UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA AGRÍCOLA

USO DE TÉCNICAS DE CLASSIFICAÇÃO AUTOMÁTICA NA ANÁLISE AMBIENTAL: UM ESTUDO DE CASO

Antonio Cesar de Barros MUNARI

Março - 2001

CAMPINAS - SP

UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA AGRÍCOLA

USO DE TÉCNICAS DE CLASSIFICAÇÃO AUTOMÁTICA NA ANÁLISE AMBIENTAL: UM ESTUDO DE CASO

Antonio Cesar de Barros MUNARI

Dissertação apresentada à Universidade Estadual de Campinas para a obtenção do Título de Mestre em Engenharia Agrícola, sob orientação do Professor Doutor Luiz Henrique Antunes Rodrigues.

Março - 2001

CAMPINAS - SP

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

M92u

Munari, Antonio Cesar de Barros
Uso de técnicas de classificação automática na análise ambiental: um estudo de caso / Antonio Cesar de Barros Munari. --Campinas, SP: [s.n.], 2001.

Orientador: Luiz Henrique Antunes Rodrigues.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Análise ambiental. 2. Árvores de decisão. 3. Indução (Lógica). 4. Geomorfologia. I. Rodrigues, Luiz Henrique Antunes. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

"Antes da Iluminação,
cortar lenha e carregar água.
Depois da Iluminação,
cortar lenha e carregar água."

Provérbio Zen

A meus pais Dorival e Maria Lúcia, meus irmãos Carmem e Vitor e à Regina, pela presença, compreensão e apoio constantes.

Agradecimentos

Ao Dr. Luiz Henrique Antunes Rodrigues, pela orientação, camaradagem e ensinamentos que foram essenciais na elaboração deste trabalho.

Ao Dr. José Paulo Marsola Garcia, pela generosidade, disponibilidade e informações fornecidas, todas elas indispensáveis a um trabalho desta natureza.

Ao Dr. Jansle Vieira da Rocha, pelo apoio e ensinamentos sempre presentes.

À Dra. Cláudia Maria Bauzer Medeiros, pela disponibilidade e orientação em momentos críticos do projeto.

À todos que direta ou indiretamente colaboraram para a elaboração deste trabalho, principalmente à Dra Maria Carolina Monard e ao Gustavo Batista do ICMS-Usp em São Carlos e ao Rogério, Fábio e Daniela, do setor de informática da Faculdade Prudente de Moraes, em Itu.

Resumo

Este trabalho tem por objetivo analisar as possibilidades e condições necessárias à utilização de uma ferramenta para a indução automática de regras de classificação em uma pesquisa de reconhecimento de unidades ambientais. A adoção desse tipo de ferramenta permite automatizar uma parte significativa da etapa referente ao processo investigatório desses estudos, de maneira que o especialista possa concentrar-se com maior ênfase nos aspectos naturalmente mais dependentes da intervenção humana, como aqueles ligados à criatividade ou julgamento, por exemplo. A análise consistiu da condução de um estudo de caso onde aplicou-se uma ferramenta indutora de regras de classificação sobre os dados de uma pesquisa acadêmica voltada ao reconhecimento de unidades ambientais que foi concluída originalmente sem o emprego de tais técnicas. Diversas adaptações necessárias foram implementadas e aspectos problemáticos foram identificados, discutidos e, quando possível, solucionados. Os resultados obtidos são em geral convergentes com os do trabalho de referência e também indicam que o emprego de ferramentas indutoras de regras pode ser bastante útil como um elemento de apoio à decisão para o especialista na condução de suas análises, apesar de requerer cuidados especiais na definição e coleta dos dados a serem utilizados, em função das características da ferramenta escolhida.

Use of Automatic Classification Techniques in Environmental Analysis: A Case Study

Abstract

This work aims to analyze the necessary possibilities and conditions concerning the use of a tool that induces to automatic classification rules in an environmental unit recognition research. The adoption of this kind of tool allows the automation of a very important part of the phase referred to the investigatory process of these studies, in such a way that the expert can concentrate with more emphasis on aspects naturally more dependent of human intervention, such as those related to creativity or judgement. This analysis consisted in the conduction of the study of a case where an inductive rule classification tool was applied to the data of an academic research devoted to the recognition on environmental units that was concluded originally without the use of such techniques. A lot of necessary adaptation was implemented and problematic aspects were identified, discussed and, when possible, solved.

Final results converged to the reference work and also indicate that the use of inductive rule tool can be very useful as a supportive element for expert's decisions on conducting their analysis, although some special care is required to define and collect data to be used, due to the features of the chosen tool.

Sumário

1. Introdução 1
2. Revisão bibliográfica 3
   2.1 Descoberta de conhecimento em bancos de dados 3
   2.2 Classificadores baseados em árvore de decisão 7
      2.2.1 Aspectos gerais 7
      2.2.2 Análise de classificadores 13
3. Material e metodologia 25
   3.1 O trabalho de referência 25
      3.1.1 O local 26
      3.1.2 Metodologia utilizada 26
      3.1.3 Resultados obtidos 31
   3.2 A ferramenta See5 34
      3.2.1 Recursos para a geração de classificadores 35
      3.2.2 Estrutura dos arquivos utilizados 41
      3.2.3 Leitura dos resultados 42
      3.2.4 Análise da ferramenta 44
   3.3 Aspectos quantitativos 45
      3.3.1 Amostragem 46
      3.3.2 Estatística e mineração de dados 50
      3.3.3 Considerações sobre o problema 52
   3.4 O banco de dados 54
      3.4.1 Estratégia 55
      3.4.2 Os dados originais 55
      3.4.3 Adaptações realizadas 59
   3.5 Os experimentos 61
4. Resultados 65
   4.1 Experimento 1 67
      4.1.1 Versão 1a (10-20 cm) 67
      4.1.2 Versão 1a (10-40 cm) 73
      4.1.3 Versão 1b (10-20 cm) 81
      4.1.4 Versão 1b (10-40 cm) 88
   4.2 Experimento 2 94
      4.2.1 Versão 2a (10-20 cm) 94
      4.2.2 Versão 2a (10-40 cm) 99
      4.2.3 Versão 2b (10-20 cm) 105
      4.2.4 Versão 2b (10-40 cm) 111
      4.2.5 Versão 2c (10-20 cm) 118
      4.2.6 Versão 2c (10-40 cm) 125
5. Conclusões 130
6. Referências bibliográficas 132

   Anexo 1: Resumo estatístico dos experimentos 137

   Anexo 2: Escalas ordenadas para atributos discretos 141

Lista de figuras

Pág.

1. Estrutura básica de uma árvore de decisão 9
2. Árvore de um dos classificadores possíveis para o caso das espécies vegetais 10
3. Representação de um segundo classificador possível para o caso das espécies vegetais 18
4. A árvore da figura 2 após uma poda 20
5. Localização da área de estudo 27
6. Localização dos pontos de coleta de material 29
7. Delimitação das SGHs 32
8. Opções para a geração de classificadores no See5 37
9. Arquivos utilizados pelo See5 para a geração de classificadores 41
10. Exemplo de relatório de saída do See5 43
11. Representação gráfica da árvore indicada no relatório de saída do See5 . 44
12. Comportamento geral do erro em função da amostra 50
13. Passos principais do processo de extração de conhecimento em bancos de dados 54
14. Estratégia de produção do arquivo de dados para o See5 55

Lista de tabelas

Pág.

1. Um levantamento fictício sobre as características físicas de 10 regiões .. 8
2. Superfícies Geneticamente Homogêneas da Planície de Picinguaba 31
3. Exemplo de normalização dos valores de um atributo 58
4. Escala ordinal para os valores do atributo Grau de Seleção das amostras 61
5. As diferentes versões dos experimentos e os dados utilizados 63

Lista de gráficos

Pág.

1. Número de espécies das famílias mais representativas da flora vascular da restinga do Núcleo Picinguaba 30
2. Dados granulométricos do ponto de coleta TL-III 33
3. Dados granulométricos do ponto de coleta CF-IV 33
4. Dados granulométricos do ponto de coleta TQ-III 33

Capítulo 1: Introdução

A trajetória da raça humana tem sido marcada por incontáveis demonstrações de sua notável capacidade de inovação e adaptação ao meio-ambiente, num constante exercício de superação de adversidades. O aprimoramento e a reafirmação contínua dessas habilidades propiciou um grande crescimento populacional, fazendo com que nossa espécie rapidamente ocupasse todo o planeta, até chegarmos ao atual estado de coisas, marcado pela intensa exploração dos espaços. As conseqüências negativas desse processo são bastante conhecidas de todos nós: degradação e devastação ambiental; extinção de espécies; fome e miséria; guerras; etc e representam em seu conjunto talvez o maior obstáculo com que a Humanidade se deparou em toda a sua existência. A tomada de consciência quanto a essa situação tem dado origem a iniciativas diversas no sentido de solucionar ou amenizar tais problemas, sendo uma delas o levantamento sistemático dos recursos ambientais de cada região, hoje entendido como um elemento estratégico para a sociedade moderna, uma vez que permite estabelecer controles com relação aos processos de degradação ambiental (GUERRA, 1966) e também inventariar um patrimônio de valor inestimável para o futuro da pesquisa científica.

Um elemento de fundamental importância nesse contexto é a compreensão do papel desempenhado por cada componente de um ecossistema na manutenção do equilíbrio ambiental, temática essa que tem inspirado um vasto número de trabalhos científicos nas mais diversas áreas. Estudos que procuram vincular os elementos físicos de um local com as formas de vida ali encontradas são muito importantes e geralmente envolvem uma grande complexidade, com uma característica fortemente interdisciplinar (GUERRA, 1966). Um caso representativo desse tipo de estudo é o mapeamento geomorfológico de detalhe, realizado em conjunto com levantamentos florísticos e de ecologia vegetal, onde procura-se estabelecer uma classificação dos ecossistemas baseada na fisionomia da vegetação e em sua localização geográfica, assim como nas feições geomorfológicas onde ocorre (GARCIA, 1995).

Inúmeros projetos de estudo científico do meio-ambiente são desenvolvidos atualmente em todo o mundo, com custos e estruturas bastante variadas conforme seus propósitos finais e o tamanho da região considerada. As metodologias utilizadas geralmente envolvem etapas de mapeamento e interpretação de mapas e fotos já existentes sobre a região; de coleta de material e observações em campo; ensaios de laboratório com as amostras recolhidas e uma posterior interpretação dos resultados obtidos (PIRES NETO, 1991; GUERRA, 1998), sendo que as três primeiras tarefas são relativamente sistematizadas e objetivas, delineadas através de métodos formais bem definidos, enquanto que a última, por sua própria natureza, consiste de um trabalho de investigação freqüentemente demorado e ainda bastante dependente da sensibilidade do especialista. Estudos como o anteriormente citado, sobre a correlação entre a geomorfologia de um local e a sua vegetação se encaixam nessa situação, e envolvem um volume considerável de dados a serem trabalhados, com uma etapa final que é tradicionalmente executada de forma totalmente não automatizada, requerendo muito tempo do especialista para a elaboração e validação de suas conclusões e dificultando, ou até mesmo impossibilitando em alguns casos, a execução de análises exaustivas do material

coletado, que poderiam levar a uma melhor compreensão das complexas inter-relações entre os diversos elementos de um ecossistema.

Portanto, contribuições no sentido de se coletar rápida e automaticamente evidências sobre os dados disponíveis e suas interligações são bastante desejáveis e fatores aceleradores do processo final de análise, além de poderem melhorar a precisão e mesmo a qualidade das conclusões obtidas em alguns casos. Uma série de estudos nas áreas da Estatística e da Computação têm sido conduzidos, já a vários anos, com o objetivo de se tratar adequadamente problemas dessa natureza (JOHNSON, 1998; HAN, 1995; MICHALSKI *et al.,* 1998), ligados em última análise à manipulação e gerência do conhecimento e à proposição de técnicas eficientes para a sua sistematização. É o caso, por exemplo, de áreas como a lógica matemática, a análise estatística multivariada, a teoria da informação, a inteligência artificial, os sistemas especialistas e a tecnologia de bancos de dados que, cada uma à sua maneira, têm oferecido significativos avanços no sentido de se prover meios para um suporte automatizado à análise de dados e ao processo decisório, seja nas aplicações comerciais, seja nas pesquisas científicas.

Um indicador da força que os estudos e técnicas para o tratamento do conhecimento têm adquirido ao longo do tempo é o significativo número de trabalhos publicados anualmente e o sucesso que as principais conferências nessa área têm conseguido. Termos como "Sistemas baseados em conhecimento", "Descoberta de conhecimento em bancos de dados", "Aprendizagem de máquina", etc tem proliferado e atingido, em alguns casos, até mesmo a grande imprensa. Uma forte base teórica já se encontra desenvolvida neste momento, e inúmeras formas de aplicação desses conceitos têm sido propostas, sendo que dentre as mais importantes estão a representação de conhecimento a partir de regras e a indução dessas regras a partir de algoritmos computacionais (RICH, 1993; MICHALSKI *et al.,* 1998; SILBERSCHATZ, 1999).

Este trabalho tem por objetivo analisar as possibilidades e condições necessárias à utilização de uma ferramenta para a indução automática de regras de classificação no processo de reconhecimento de unidades ambientais. A adoção desse tipo de ferramenta automatizaria uma parte significativa da etapa referente ao processo investigatório desses estudos, permitindo ao especialista concentrar-se nos aspectos naturalmente mais dependentes da intervenção humana, como aqueles ligados à criatividade ou julgamento, por exemplo. Em resumo, o que se pretende é verificar a viabilidade da utilização de classificadores de dados baseados em regras e árvores de decisão para acelerar o processo de análise, interpretação e validação dos resultados em um caso específico de estudo do meio ambiente.

Capítulo 2: Revisão bibliográfica

A pesquisa referente à descoberta e manipulação automáticas do conhecimento tem sido reconhecidamente bastante fértil ao longo dos anos, existindo atualmente um imenso volume de trabalhos à disposição da sociedade em geral e da comunidade científica em particular. O objetivo deste capítulo é apresentar os elementos necessários para uma compreensão da área que seja suficiente para o acompanhamento do restante deste trabalho, concentrando-se nas vertentes de maior interesse para este estudo de caso específico, procurando com isso evitar uma incursão excessivamente detalhada na vastidão de conceitos e abordagens atualmente utilizados. Assim, esta revisão constará de uma apresentação bastante geral da área da descoberta de conhecimento em bancos de dados seguida por um maior detalhamento sobre os classificadores baseados em árvores de decisão.

2.1 Descoberta de conhecimento em bancos de dados

A expressão "descoberta de conhecimento em bancos de dados" (ou KDD, da abreviação do termo em inglês - *Knowledge Discovery in Databases*) é freqüentemente utilizada para representar o processo de extração / inferência de conhecimento significativo, na forma de regras, restrições e regularidades, a partir de dados contidos em bancos de dados (HAN, 1995; FAYYAD, 1996a). Outros termos de emprego muito comum para esse fim são extração de conhecimento, descoberta de informação, colheita de informação, arqueologia de dados, dragagem de dados (*data dredging)* e processamento de padrões de dados. A expressão mineração de dados (*data mining*), também tem sido bastante utilizada com os mesmos propósitos apesar de, em princípio, representar apenas uma das etapas centrais do processo (FAYYAD, 1996a). Um outro termo também diretamente associado à descoberta de conhecimento é "*machine learning*" (ou aprendizagem de máquina ou ainda aprendizagem automática, numa tradução livre), que representa uma abordagem proposta já a cerca de 40 anos e que pesquisa formas de aquisição de conhecimento geralmente baseadas em indução a partir de exemplos (MICHALSKI *et al.,* 1998).

Esforços para a descoberta de conhecimento em bancos de dados justificam-se nos dias de hoje devido ao imenso volume de dados mantidos pelos sistemas convencionais que, analisados em uma perspectiva mais ampla, tipicamente histórica ou espacial, podem revelar informações importantes sobre os padrões de relacionamento entre os dados. Trata-se, assim, de uma análise em busca de um novo tipo de conteúdo presente nos bancos de dados, que não pode ser visualizado de maneira imediata pela simples leitura de seus dados, e que pode expressar significados mais profundos sobre os mesmos. De uma certa forma, é a transição da idéia de "dados" para a de "informação", "conhecimento", "visão geral", obtida através da exploração das regularidades eventualmente presentes na base de dados. O estudo de métodos e princípios para KDD tem se constituído em uma das frentes de pesquisa mais importantes da ciência da computação na atualidade e envolve principalmente as áreas de Inteligência Artificial, Estatística e Bancos de Dados.

O processamento dos dados com vistas à aquisição automática de conhecimento envolve a aplicação de algoritmos que, de uma maneira geral, procuram detectar a ocorrência de

padrões diversos entre os dados, certificados através de algum tipo de suporte estatístico. Por isso, uma questão muito importante que se coloca é a da qualidade dos dados disponíveis para o processamento, que afeta diretamente o grau de confiabilidade nos resultados obtidos. Assim, nessa perspectiva, dados relativamente estáveis, completos e sem redundâncias são mais adequados ao processo KDD do que aqueles altamente dinâmicos, incompletos, imprecisos, redundantes, poluídos e esparsos. Como freqüentemente as bases de dados disponíveis apresentam algumas dessas inconsistências quanto ao conteúdo e, eventualmente, problemas com a estrutura física das tabelas (ou seja, no nível do esquema do banco de dados original), uma etapa típica do KDD é a preparação desse conteúdo para o processamento, tanto no aspecto da 'limpeza' dos dados, filtrando aqueles que sejam mais relevantes para os propósitos em vista e removendo as inconsistências detectadas como, quando necessário, no aspecto físico da base de dados, adaptando os formatos e estruturas originais para esquemas mais adequados às técnicas e equipamentos que se pretende utilizar (FAYYAD, 1996a). Quando o esforço para a melhoria da qualidade dos dados não consegue atingir resultados considerados satisfatórios, torna-se necessário conduzir o processo de descoberta de informação com o uso de técnicas e algoritmos mais complexos, que tentam tratar as diversas imperfeições possíveis (MICHALSKI *et al.,* 1998).

Uma noção muito importante, principalmente na área de aprendizagem automática, é a de conceito: uma abstração estabelecida para um conjunto de objetos que compartilham algumas características que os diferenciam de outros conceitos (MICHALSKI *et al.,* 1998). Ao tratar da descoberta de conhecimento estamos interessados em identificar novos conceitos, caracterizá-los e detectar padrões na sua ocorrência, podendo envolver ou não esforços para a sua generalização. O conhecimento *não baseado em generalização*, também chamado de *conhecimento em nível primitivo,* é caracterizado pela ausência de um esforço de tradução (abstração) dos conceitos originalmente representados na base de dados para conceitos de mais alto nível. São exemplos de conhecimento em nível primitivo as regras de dependência funcional, de dependência multi-valorada e as regras de dedução. O conhecimento *baseado em generalização*, por sua vez, assume a existência de uma hierarquia de conceitos, que pode ser definida explicitamente por especialistas ou então ser gerada automaticamente por processos de análise de dados. Dispondo dessa estrutura, torna-se possível trafegar verticalmente nos diversos níveis de abstração, abordando informações em graus de generalidade variados, mas preservando uma consistência com os conceitos de mais baixo nível originais.

Uma regra é uma especificação que descreve um comportamento padrão em um dado domínio, e permite representar vários tipos de conhecimento (SILBERSCHATZ, 1999). Não obstante os diversos formalismos passíveis de serem utilizados para expressar regras, estas, de uma maneira geral, apresentam uma estrutura típica e bastante simples:

$$\forall\ X\ antecedente \Rightarrow conseqüente$$

em que X é uma lista de uma ou mais variáveis. Conforme o caso, esses elementos podem receber outras denominações, como por exemplo LHS (*Left Hand Size* - Lado esquerdo) ou corpo da regra para indicar o antecedente e RHS (*Right Hand Size* - Lado direito) para o conseqüente.

A avaliação da pertinência de uma regra geralmente está associada a duas medidas extremamente importantes, que revelam a natureza estatística da abordagem: o *suporte* e a *confiança*, calculadas com base no confronto entre os resultados obtidos pela aplicação da regra face ao conjunto de dados em que ela se originou.

Assim, seja a regra a avaliar representada na forma $F(o) \Rightarrow G(o)$, onde O indica o conjunto de dados utilizado, $F(o)$ é o antecedente da regra e $G(o)$ o conseqüente. Podemos definir as duas medidas da seguinte forma:

a) Fator de confiança: também chamado de força da regra, indica a fração dos objetos em O que satisfazem F e também satisfazem G (HAN, 1995), ou seja, mede a freqüência com que, dada uma ocorrência do antecedente, o conseqüente também ocorre (SILBERSCHATZ, 1999). Assim, se temos uma regra que afirma que dor de cabeça e irritação na mucosa indicam uma doença D, e verificamos que em 90% dos pacientes onde os sintomas eram dor de cabeça e irritação da mucosa o diagnóstico se confirmou como sendo realmente da doença D, temos um fator de confiança de 90%, o que faz dessa regra um conhecimento bastante interessante, dado o seu alto grau de certeza na interpretação dos sintomas. Portanto, quanto maior o fator de confiança, mais forte a regra, sendo que os valores que indicam quais percentuais são mais adequados para que uma regra seja considerada forte varia conforme o caso.

b) Restrição de suporte: ou significância estatística do padrão, representa a fração dos objetos em O que satisfazem F e G (HAN, 1995). Em outras palavras, é a parcela do conjunto de dados que satisfaz tanto o antecedente como o conseqüente da regra (SILBERSCHATZ, 1999). Por exemplo, detectamos uma regra que afirma que clientes do sexo feminino (o antecedente) compram carros vermelhos (o conseqüente). Se verificamos que, por exemplo, 0.02% de todas as transações envolvem a compra de carros vermelhos por mulheres, temos uma regra que, apesar de eventualmente poder ser verdadeira, apresenta uma incidência muito pequena sobre os fatos retratados na base, o que coloca em questão a sua pertinência. Quanto maior o suporte de uma regra, maior a sua representatividade em expressar fatos sobre a base de dados.

Existem diversos tipos de conhecimento que podem ser "garimpados" através de processos KDD, cada um deles dependendo das características e do teor do banco de dados disponível para uso: regras de associação de dados; regras de classificação; detecção de padrões seqüenciais; etc. Cada uma dessas variedades pode ser expressa formalmente através de regras, e encontram-se um pouco mais detalhadas a seguir. A compreensão do significado de cada uma das principais formas de conhecimento é importante na medida em que, conforme o objetivo da pesquisa (ou seja, da natureza do conhecimento a ser buscado), serão as técnicas mais adequadas para utilização no processo.

Regras de associação: identifica conjuntos de objetos que aparecem juntos (HAN, 1995). Um exemplo clássico é aquele em que buscamos identificar, numa base de dados comercial, quais os produtos que comumente aparecem na mesma transação de venda, indicando um padrão de consumo que pode ser interessante para o negócio.

Regras de classificação: buscam distribuir os elementos de um conjunto de dados em categorias previamente definidas, com base no valor de alguns dentre seus atributos

(JOHNSON, 1998; CHEN, 1996). Uma regra de classificação (ou de *alocação*, segundo alguns autores) é produzida com o objetivo de facilitar a atribuição de novos casos às classes (categorias) consideradas. Um caso típico é aquele em que uma empresa procura distribuir seus clientes em categorias com base em algum critério relevante para suas atividades.

Regras (ou análise) de agrupamento (Clustering analysis): consiste em segmentar um conjunto de dados (sem um atributo de classe pré-definido), compondo grupos identificáveis e homogêneos, com base em algumas características relevantes. Esses agrupamentos (ou classes, ou *clusters*) representam conjuntos de entidades com alta coesão conceitual (MICHALSKI *et al.,* 1998) e são produzidos de acordo com o princípio básico do agrupamento, que é "maximizar a similaridade intra-classe e minimizar a similaridade inter-classe", sendo essa similaridade ou proximidade medida através do cálculo das distâncias entre os valores observados, que depois são confrontados com uma escala quantitativa que determina, em última análise, o grau de pertinência dos objetos aos agrupamentos (JOHNSON, 1998). Como se pode observar, existe uma certa semelhança entre as tarefas de agrupar e de classificar dados, já que ambas objetivam a formação de grupos. A diferença essencial entre elas reside no critério usado no processamento: na classificação, ele é conhecido inicialmente, enquanto que no agrupamento isso não acontece (FAYYAD, 1996b), sendo por esse motivo também chamado de classificação não-supervisionada (CHEN, 1996). A construção de agrupamentos pode ser útil, por exemplo, para identificar conjuntos afins de observações feitas em campo, constituindo categorias bem definidas certificadas por similaridades métricas ou probabilísticas, de maneira a facilitar uma posterior análise dos dados.

Regras (ou análise) de seqüências: procura reconhecer modelos de padrão seqüencial, ou seja, dados com algum tipo de dependência temporal, de maneira a identificar desvios e tendências através do tempo (FAYYAD, 1996b). Constitui-se, assim, num tipo especial de regra de associação, que busca correlacionar dados que aparecem em transações separadas, em oposição aos dados que aparecem na mesma transação, no caso da associação (HAN, 1995). Uma aplicação típica de análise seqüencial é a previsão do comportamento de indicadores econômicos como, por exemplo, identificar uma regra que afirme que "quando os juros sobem, o índice da bolsa de valores cai dentro de 2 dias" (SILBERSCHATZ, 1999). Conforme o enfoque dado na análise, este tipo de conhecimento pode receber outras denominações como, por exemplo, análise de evolução e análise de desvio.

Regras de caracterização, generalização e sumarização: caracterizam cada grupo no conjunto de observações de acordo com um nível de abstração, resumindo ou detalhando informações (HAN, 1995).

Um aspecto importante na descoberta de conhecimento é a forma de condução do processo, que pode ser dirigida por comandos, onde o usuário especifica (com o auxílio de uma linguagem especial geralmente derivada da SQL, uma tradicional linguagem de consulta a bancos de dados, ou interfaces gráficas) exatamente o que ele pretende pesquisar sobre os dados, ou autônoma, sem a interferência humana, que pode gerar conhecimentos extremamente interessantes e originais mas que tipicamente requer pesquisas exaustivas sobre a base de dados e freqüentemente conduz à descoberta de conhecimento irrelevante. Muitas vezes é conveniente utilizar num estudo as duas abordagens de condução do processo, considerando as características e os propósitos de cada etapa individualmente.

O conhecimento em nível primitivo, por não envolver a conversão de conceitos para um nível mais genérico, é de detecção mais direta. Exemplos típicos de conhecimento neste nível são regras de classificação, associação, dedução, dependência funcional e dependência multi-valorada. As abordagens mais comumente empregadas na sua detecção envolvem o uso de algoritmos de árvore de decisão (como ID3/C4.5/See5 entre outros). Uma interessante introdução sobre esses algoritmos é encontrada em (MICHALSKI *et al.,* 1998; QUINLAN, 1993; HART, 1986; RUSSEL, 1995), comentários sobre algumas de suas características principais aparecem também em (DYMOND, 1994; HUANG, 1996), enquanto que uma discussão sobre as classes de problemas que eles podem resolver e exemplos de sua utilização pode ser encontrada em (MICHALSKI *et al.,* 1998; QUINLAN, 1993; FAYYAD, 1996a).

2.2 Classificadores baseados em árvore de decisão

2.2.1 Aspectos gerais

Um classificador é uma função que, dada uma coleção de dados *D*, composto por *n* objetos, cada um deles descrito por um mesmo conjunto de atributos *A*, produz para cada elemento de *D* um valor de classe *C*, com base nos valores internos de seus atributos. Os valores das classes são denominados "rótulos" e são definidos antes da função ser executada. A classe em si, por sua vez, é também um atributo do objeto classificado, denominado atributo categórico ou atributo dependente, enquanto que os demais atributos, aqueles que são utilizados para determinar a classe de um objeto em particular, são chamados de atributos de predição ou não-categóricos (GANTI, 1999; CHEN, 1996). Os classificadores geralmente são construídos por programas chamados indutores, que implementam algoritmos computacionais especiais, que operam sobre uma massa de dados inicial considerada representativa do domínio do problema e na qual tanto o valor dos atributos comuns quanto da classe de cada objeto são conhecidos. Um programa indutor de classificadores procurará, com base nas ocorrências desse conjunto de dados inicial, chamado de conjunto de treinamento (*training set*), estabelecer qual a ligação entre os valores dos atributos não-categóricos e as classes encontradas na massa de dados. A expectativa é de que essas relações encontradas, chamadas de regras de classificação e que representam em última instância o classificador em si, possam ser empregadas posteriormente para determinar o valor da classe para objetos onde essa informação é desconhecida, num tipo de atividade chamada predição (da classe) (WEISS, 1998). Para que o grau de acerto de um classificador assim produzido possa ser avaliado antes de sua efetiva utilização prática geralmente procura-se aplicá-lo sobre um segundo conjunto de dados onde o valor da classe é igualmente conhecido, chamado conjunto de teste (*test set*), posteriormente comparando-se o grau de concordância entre a classe prevista pelo classificador para cada objeto e a classe realmente observada.

Exemplo: em uma pesquisa fictícia foram coletados e organizados dados sobre algumas características físicas de dez regiões diferentes, e busca-se agora determinar elementos sobre a distribuição de uma dada espécie vegetal. As características levantadas foram o Grau de Umidade do terreno, sua Profundidade, Teor de Argila, a Altitude Média e o Tipo de Cobertura Vegetal encontrada no local. Além disso, observou-se qual a variedade da

espécie vegetal estudada que apresentava-se como dominante naquela região, com os resultados das supostas observações organizados na tabela apresentada a seguir.

Região	*Umidade*	*Prof.*	*Argila*	*Altitude*	*Cobertura*	*Dominante*
1	MÉDIA	120	ALTO	300	SECUNDÁRIA	B
2	ALTA	120	BAIXO	20	PRIMÁRIA	C
3	MÉDIA	110	BAIXO	200	ANTRÓPICA	B
4	MÉDIA	50	ALTO	200	ANTRÓPICA	A
5	BAIXA	110	MÉDIO	12	PRIMÁRIA	B
6	ALTA	10	ALTO	15	SECUNDÁRIA	B
7	ALTA	200	MÉDIO	220	PRIMÁRIA	A
8	ALTA	40	BAIXO	200	SECUNDÁRIA	A
9	MÉDIA	110	ALTO	30	PRIMÁRIA	B
10	MÉDIA	20	BAIXO	25	SECUNDÁRIA	C

Tabela 1: Um levantamento fictício sobre as características físicas de 10 regiões.

A coleção de dados *D* é representada pela tabela, onde cada uma de suas 10 linhas corresponde a um objeto distinto a ser classificado. Todos os objetos possuem o mesmo conjunto *A* de atributos {Região, Umidade, Prof., Argila, Altitude, Cobertura, Dominante}, cada um dos quais com seus valores específicos. O atributo 'Dominante', neste caso, constitui a classe ou atributo categórico do objeto e os seus valores específicos ou rótulos possíveis são 'A', 'B' e 'C'. Pretende-se descobrir um critério para determinar a classe de cada região em função dos seus demais atributos, que formam o conjunto de atributos de predição ou atributos não-categóricos. Uma vez gerado o classificador para este problema, com base nos dados deste conjunto de treinamento inicial, cujo tamanho reduzido serve apenas para fins de exemplo, seria possível então determinar a variedade dominante do vegetal estudado em qualquer outra área onde se conhecesse apenas os atributos não categóricos utilizados pelo classificador.

Existem várias técnicas para a indução de classificadores como, por exemplo, as redes neurais, os algoritmos genéticos, métodos estatísticos diversos, algoritmos baseados em árvore de decisão, etc (GANTI, 1999) cada uma possuindo suas próprias vantagens e desvantagens. Algumas técnicas levam à produção de funções de classificação eficazes que, entretanto, não podem ser representadas logicamente no vocabulário e nível de abstração do usuário comum, sendo indicadas para tarefas de classificação automática, onde o que interessa é apenas o resultado da classificação, e não exatamente o critério de classificação adotado. Esses classificadores "caixa-preta" tipicamente são gerados por abordagens conexionistas (redes neurais) e muito utilizados em diversas áreas, adequando-se a tarefas onde a análise dos dados inicia-se efetivamente a partir do resultado da classificação. Outros métodos produzem classificadores mais compreensíveis para os seres humanos, como os baseados em árvore de decisão, permitindo estruturar hierarquicamente, nos termos do domínio do problema, os critérios de decisão adotados pela função. Tais classificadores são adequados para tarefas onde o entendimento do critério de seleção utilizado é parte importante da análise, como por exemplo em estudos sobre a correlação entre certas propriedades de um objeto (representadas pelos atributos selecionados) e sua classificação final, permitindo automatizar e acelerar uma parte significativa do processo analítico, reduzir erros e viabilizar análises mais rigorosas e exaustivas de um problema.

Esse potencial faz das ferramentas indutoras de classificadores baseados em árvore de decisão um recurso bastante atrativo para usuários tanto do mundo acadêmico como corporativo, e por esse motivo, sua análise constitui o foco deste trabalho.

É possível, caso necessário, distinguir entre o classificador em si e a sua representação: o primeiro é uma função computacional típica, com entrada e saída de dados e o segundo, uma descrição mais ou menos inteligível da lógica interna desse processamento. Algumas técnicas de geração de classificadores propiciam representações mais compreensíveis destes do que outras, sem que esse aspecto esteja ligado necessariamente à eficácia do classificador (LU, 1996; WEISS, 1998), sendo que as duas formas de representação de classificadores mais comumente empregadas são as árvores de decisão e as regras de decisão (WEISS, 1998; BERSON, 1997). Uma árvore de decisão é uma estrutura gráfica hierárquica composta por nós, ramos e folhas, como a mostrada na figura 1.

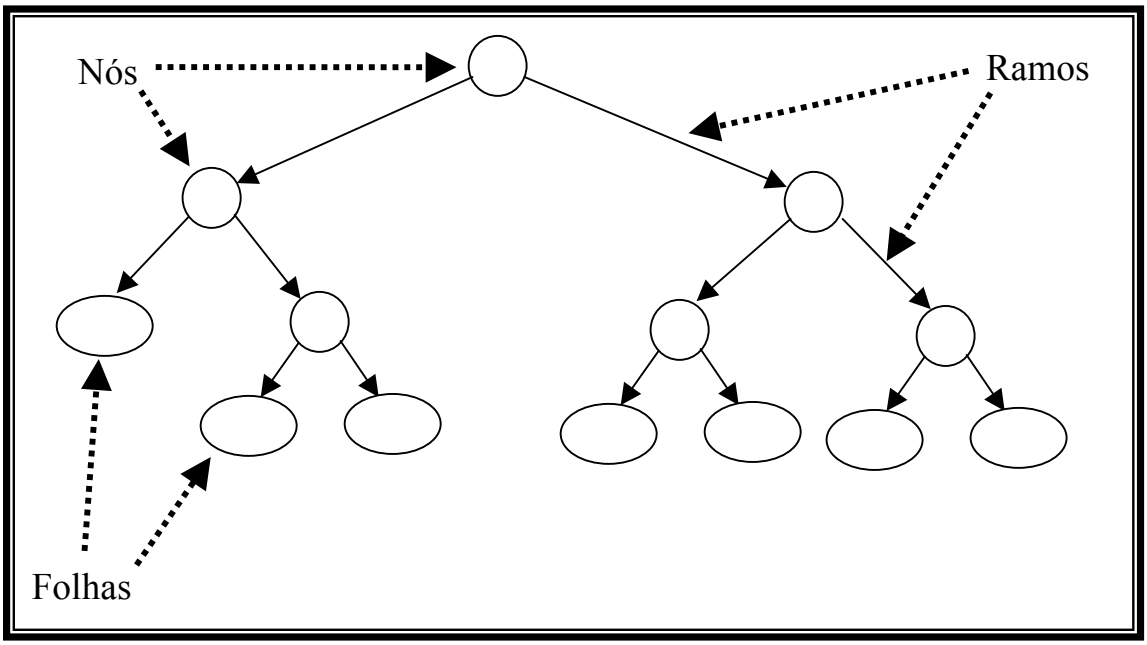


Figura 1: Estrutura básica de uma árvore de decisão.

Cada nó refere-se a um ou mais atributos de predição e cada ramo ali originado representa um possível valor ou faixa de valores para aqueles atributos. Esse encadeamento segue até que se atinja uma folha, um tipo de nó especial que não possui desdobramentos. Em uma árvore de decisão uma folha corresponde a um valor de atributo categórico ou seja, uma classe para o objeto analisado, conforme indica a figura 2, que representa um critério de classificação para o exemplo da distribuição das espécies vegetais citado anteriormente.

Uma árvore de decisão como a apresentada na figura 2, que possui apenas um atributo testado em cada nó, é chamada de árvore univariada (*univariate tree*) e constitui o caso de implementação mais simples e comum na literatura atualmente disponível. Já existem, entretanto, algoritmos que conseguem induzir árvores de decisão nas quais um nó pode representar uma expressão lógica que envolve a combinação de dois ou mais atributos. Tais árvores são chamadas de árvores multivariadas (*multivariate trees*) e permitem tratar condições complexas gerando classificadores mais acurados e compactos para algumas

situações onde há forte inferência entre alguns atributos, e têm sido um importante objeto de estudo nos últimos anos (BRESLOW, 1996; CHEN, 1996).

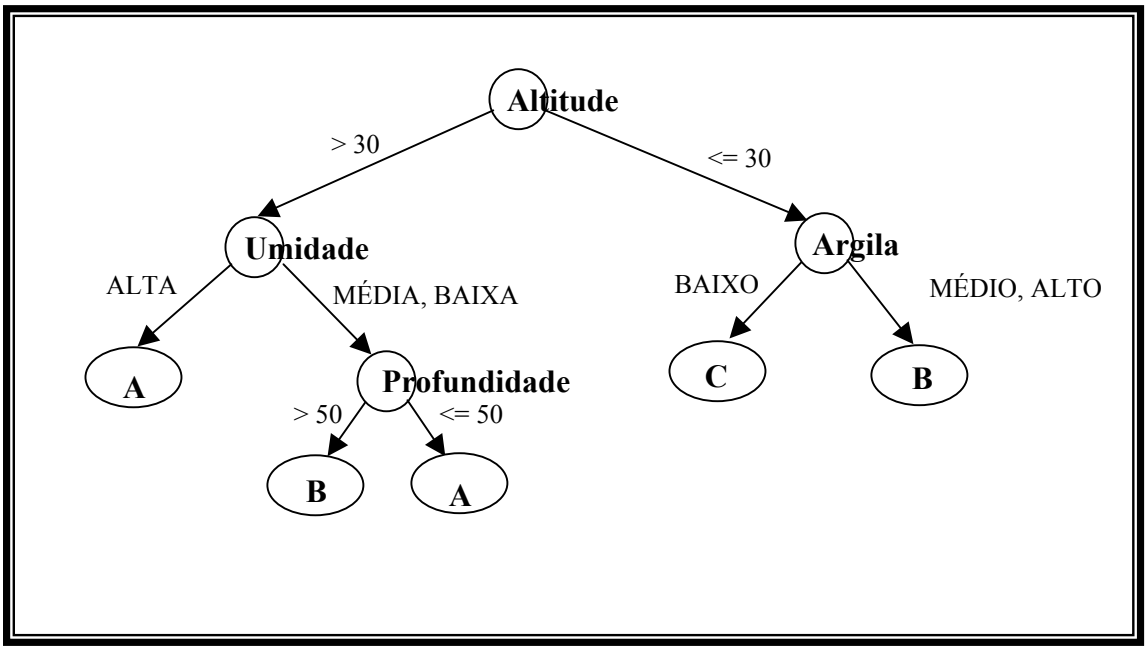


Figura 2: Árvore de um dos classificadores possíveis para o caso das espécies vegetais.

Uma regra de decisão por sua vez procura descrever os critérios adotados pelo classificador na forma de um conjunto de especificações SE <condição> ENTÃO <ação>, de compreensão geralmente fácil para os seres humanos. Para o classificador ilustrado na figura 2, poderia ser montado o seguinte conjunto de regras de decisão:

Regra 1:
SE Umidade = ALTA E Altitude > 30
ENTÃO Dominante = A

Regra 2:
SE Profundidade <= 50 E Altitude > 30
ENTÃO Dominante = A

Regra 3:
SE Argila EM {ALTO, MÉDIO} E Altitude <= 30
ENTÃO Dominante = B

Regra 4:
SE Umidade = MÉDIA E Profundidade > 50
ENTÃO Dominante = B

Regra 5:
SE Argila = BAIXO E Altitude <= 30
ENTÃO Dominante = C

Observa-se que há uma grande concordância entre as duas representações, com diferenças parciais apenas nas regras 2 e 4, a primeira porque não considera o atributo Umidade e a última porque além de desconsiderar a possibilidade do valor “Baixa” para o atributo

Umidade, também não utiliza o atributo Altitude, já que a combinação dos valores de Profundidade e Umidade é suficiente para classificar as ocorrências do conjunto de treinamento. Uma outra diferença, mais genérica, é que as regras não refletem explicitamente uma hierarquia, como ocorre nas árvores de decisão, o que pode limitar ou não o seu poder expressivo dependendo dos propósitos do usuário final.

O estudo de algoritmos para a indução de classificadores baseados em árvore de decisão data de meados da década de 1960, quando estatísticos buscavam meios para automatizar o processo de determinar quais campos em suas bases dados seriam mais úteis para o entendimento de um problema específico, de onde se originou seu emprego relativamente comum para a formulação de hipóteses em pesquisas (BERSON, 1997).

Desde então têm sido desenvolvidas várias abordagens e algoritmos para a indução de regras de classificação a partir de conjuntos de dados de treinamento, fazendo com que essa área seja uma das mais firmemente estabelecidas no espectro da prospecção de conhecimento em bases de dados. A forma mais tradicional para a indução de regras de classificação baseia-se em uma estratégia de "divisão-e-conquista" conhecida por TDIDT (*Top-Down Induction of Decision Trees*) ou ID3 (MICHALSKI *et al.,* 1998), que pode ser resumida num algoritmo genérico como:

> Sendo S o conjunto de treinamento, faça:
>
> 1. Encontre o "melhor" atributo (de predição) *at*;
> 2. Divida o conjunto S nos subconjuntos S_1, S_2, ..., de maneira que todos os exemplos no subconjunto S_i possuam $at = v_i$. Cada subconjunto será um nó da árvore de decisão;
> 3. Para cada S_i, verificar se todos os elementos de S_i possuem o mesmo valor para o atributo categórico (classe). Caso isso seja verdade, criar um nó folha, com rótulo igual ao valor de classe encontrado; senão re-executar o procedimento, a partir do passo 1, considerando que $S = S_i$.

Esse algoritmo genérico terminará quando todos os subconjuntos estiverem associados a uma classe ou quando não houverem mais atributos de predição disponíveis para os subconjuntos ainda não-rotulados (MICHALSKI *et al.,* 1998).

Partindo dessa abordagem genérica, é possível derivar várias implementações distintas, que variam quanto a aspectos de custo computacional e, principalmente, no que se refere ao critério para determinar qual seria o "melhor" atributo de predição (atributo de *split*) num dado contexto, como será discutido posteriormente. Os principais algoritmos de indução de regras de classificação atualmente disponíveis são representados no quadro a seguir.

Algoritmo	***Características***
ID3 / C4.5 / C5.0	Desenvolvidos a partir do final da década de 1970 por J. Ross Quinlan, utilizam como critério para a seleção do melhor atributo de predição medidas derivadas do campo da Teoria da Informação. O algoritmo ID3 inicial foi posteriormente aperfeiçoado, dando origem ao C4 e C4.5, que estenderam suas funcionalidades ao permitir o tratamento de conjuntos de treinamento com valores de atributo desconhecidos; valores de atributos contínuos; implementação de estratégias de poda de árvore e recursos para a extração de regras SE-ENTÃO a partir da árvore inicialmente induzida. É provavelmente a família de algoritmos de indução de regras mais conhecida na área do aprendizado computacional. Mais recentemente uma nova geração desse algoritmo foi desenvolvida, oferecendo diversos aprimoramentos e passando a ser oferecida como um produto comercial denominado C5.0 ou See5, conforme o sistema operacional a que se destina.
CART	*Classification and Regression Trees* (CART), foi desenvolvido e apresentado em 1984 pelos pesquisadores Leo Breiman, Jerome Friedman, Richard Olshen e Charles Stone das Universidades de Stanford e da California (em Berkeley) (BERSON, 1997). Representa uma abordagem híbrida, pois utiliza conceitos tanto de Inteligência Artificial como de Estatística. O critério de escolha do "melhor" atributo pode ser a entropia (uma medida da Teoria da Informação); índices Gini (baseados na teoria das probabilidades) ou o chamado *twoing criterion*, similar aos índices Gini, mas que tendem a produzir divisões mais balanceadas na árvore resultante (BERSON, 1997; CHEN, 1996). CART atualmente é implementado em algumas ferramentas de mineração de dados comerciais.
CHAID	*Chi-Square Automatic Interaction Detector* (CHAID), desenvolvido por G.V. Kass que o apresentou originalmente no artigo *"An Exploratory Technique for Investigating Large Quantities of Categorical Data"* de 1980, baseia-se na utilização de testes de significância Chi-Quadrado (χ^2) sobre tabelas de contingência para determinar o "melhor" atributo para a subdivisão da árvore. É portanto um método fortemente fundamentado no campo da Estatística. Costuma apresentar bons resultados quanto à qualidade dos classificadores produzidos e, a exemplo do CART, é também implementado atualmente em alguns produtos comerciais (BERSON, 1997).
SLIQ	*Supervised Learning In Quest* (SLIQ) foi desenvolvido por Manish Mehta, Rakesh Agrawal e Jorma Rissanen, dentro do projeto QUEST da IBM e apresentado em 1996. O algoritmo é projetado para tratar grandes volumes de dados em um conjunto de treinamento e também utiliza o grau de entropia como critério para a seleção de atributos. Entretanto, possui uma abordagem própria para o desenvolvimento da árvore de decisão, fazendo inicialmente sua expansão na horizontal, para depois aprofundar os ramos até as folhas, numa estratégia conhecida como *breadth-first*, oposta à utilizada pelo ID3 ou CART por exemplo que, para cada caso, fazem o desenvolvimento do ramo

	correspondente até a folha, para só depois processar um novo ramo (*depth first*). Um novo algoritmo de classificação, chamado ***S**calable **P**a**R**allelizable **IN**duction of decision **T**rees* (SPRINT) foi desenvolvido pela mesma equipe para substituir o SLIQ, eliminando alguns problemas ligados à grande necessidade de memória em certos casos (MEHTA, 1996; CHEN, 1996; SHAFER, 1996).

Uma interessante comparação entre trinta e três diferentes algoritmos de indução de classificadores pode ser encontrada em (LIM, 2000), que fornece ainda uma breve descrição das características básicas de cada um.

2.2.2 Análise de classificadores

A avaliação da qualidade de um classificador é feita considerando-se prioritariamente sua acurácia, ou seja, a precisão com que executa seu trabalho de predição e, em segundo lugar, o seu grau de complexidade (CHEN, 1996; BRESLOW, 1996). Outros fatores podem eventualmente também ser considerados, como determinadas características ligadas ao seu custo computacional (HAN, 1995); número de exemplos necessários ou escalabilidade (MICHALSKI *et al.,* 1998) mas esses aspectos muitas vezes são de menor importância prática que os dois primeiros, que serão abordados com maiores detalhes nas próximas seções.

Acurácia de um classificador

Um classificador tem sua acurácia avaliada em termos do grau de acerto de suas previsões tanto sobre o conjunto de treinamento como, principalmente, sobre dados novos, medida numa fase imediatamente posterior ao treinamento. Essa medição baseada em novos dados é realizada aplicando-se o classificador sobre um conjunto de testes (*test set*) que possui valores para os atributos de predição e também valores de classe para cada caso, ou seja, é estruturalmente idêntico ao conjunto de treinamento, mas seus dados não são os mesmos. Em seguida é contabilizado o total de erros e de acertos obtidos pelo classificador gerado, produzindo-se uma medida chamada taxa de erro (*error rate*), que representa a razão entre os erros observados e o total de casos do conjunto de teste:

$$\text{taxa de erro} = \text{erros} / \text{número de casos} \quad (1)$$

Existem algumas variações possíveis no cálculo das taxas de erro observado, sendo que as principais delas referem-se à ponderação de custos e riscos de alguns tipos de erro e são implementadas atribuindo-se pesos aos casos mal-classificados, conforme a natureza do erro (WEISS, 1998; MICHALSKI *et al.,* 1998). A acurácia de um classificador corresponde ao complemento da taxa de erro, ou seja, à porcentagem de acertos verificados (MICHALSKI *et al.,* 1998):

$$\text{acurácia} = 1 - \text{taxa de erro} \quad (2)$$

O grau de acerto de um classificador é dependente de uma série de fatores, dentre os quais merecem destaque:

a) a qualidade do conjunto de treinamento, que deve possuir um volume suficiente de dados confiáveis, extraídos da população de forma aleatória (WEISS, 1998);

b) o chamado "critério de parada", que define quando o algoritmo indutor de regras deve encerrar o desenvolvimento de um classificador, dando-o por concluído;

c) as características da ferramenta utilizada para induzir o classificador a partir dos dados, como por exemplo o fundamento matemático utilizado para calcular os atributos de *split*, o tipo de abordagem para o desenvolvimento da árvore (*depth first* ou *breadth first*), etc;

d) a sistemática de avaliação dos resultados, que deve utilizar conjuntos de teste com dados representativos; tamanhos adequados e constituição independente do conjunto de treinamento (WEISS, 1998);

e) a natureza do domínio do problema.

A construção de árvores de decisão possui algumas características próprias que precisam ser levadas em conta quando se busca um melhor entendimento de seu comportamento em termos da precisão nas predições realizadas. Um primeiro aspecto é a natureza por assim dizer fragmentadora do processo: o conjunto de treinamento vai sendo subdividido à medida que os critérios vão sendo definidos durante a construção do classificador, fazendo com que a análise vá se realizando com base em conjuntos progressivamente menores quanto ao número de casos. Isso tem um efeito importante na definição dos nós mais profundos da árvore, uma vez que o suporte estatístico para a tomada de decisão vai sendo também progressivamente diminuído, colocando em dúvida a representatividade do conjunto de elementos correspondentes a essas folhas (MICHALSKI *et al.,* 1998). Por esse motivo muitos mecanismos indutores de regras de classificação oferecem recursos para o descarte (a "poda") dos ramos teoricamente menos representativos da árvore de decisão, ou ainda, a produção de montagens de diversos classificadores individuais que trabalham em conjunto, estratégias essas que serão apresentadas com maiores detalhes mais à frente neste capítulo.

Um segundo elemento importante a considerar diz respeito à forma como os atributos são selecionados para a subdivisão dos dados a serem classificados. Não existe até o momento uma solução computacionalmente viável para que se obtenha sempre a melhor árvore de decisão possível, uma vez que esse é um problema NP-complexo (HYAFIL, 1976 *apud* BRESLOW, 1996). Isso significa que o custo de se proceder a uma busca exaustiva da melhor solução é proibitivo, crescendo a taxas exponenciais à medida que o tamanho do conjunto de treinamento aumenta. Problemas com essa característica de complexidade requerem soluções baseadas em algum tipo de conhecimento prévio sobre as propriedades dos dados, de maneira a direcionar os esforços na procura de uma boa solução (mas não necessariamente a melhor), constituindo aquilo que é chamado de busca heurística. Uma ilustração comparativa entre esses dois tipos de abordagem para a solução de problemas seria imaginar o conjunto de todas as soluções possíveis (chamado espaço de estados ou ainda, no caso das árvores de decisão, floresta de decisão (MURPHY, 1994)) e um procedimento para percorrer todo esse conjunto comparando cada elemento (no caso, cada uma das possíveis soluções ou árvores) até que todos os elementos tenham sido avaliados. Nesse caso temos uma busca exaustiva que garante que sempre será obtida a melhor solução. Uma busca heurística, por sua vez, faria uma procura mais tendenciosa na floresta, visitando apenas as soluções com maior potencial de serem boas, com base em algumas premissas previamente conhecidas. Com isso, a rapidez do processo aumenta, mas é

possível que a melhor solução entre todas não tenha sido encontrada, já que ela pode ter ficado fora do trajeto percorrido. A tarefa de indução de um classificador é, portanto, levada a efeito por meio de algoritmos computacionais que implementam heurísticas para a seleção dos atributos que melhor permitem prever a classe de uma dada ocorrência no conjunto de treinamento. É isso que significa escolher o "melhor" atributo de predição num dado contexto, conforme mencionado no passo 1 da formulação genérica do algoritmo TDIDT apresentado anteriormente. A questão agora é descobrir alguma característica sobre os dados que permita definir um critério para a identificação desse melhor atributo (também chamado de atributo de *split*) em cada nível da árvore, e para isso existem várias abordagens, algumas das quais serão apresentadas a seguir, considerando por questão de simplicidade apenas a indução *top-down* de árvores univariadas.

Abordagem baseada na Teoria da Informação

O processo de construção da árvore de decisão constitui-se de uma sucessão de escolhas de critérios de divisão dos dados disponíveis, conforme mostrado anteriormente no algoritmo genérico TDIDT. Uma boa subdivisão é aquela que produz para os dados disponíveis os grupos mais homogêneos com relação ao atributo categórico, enquanto que as más subdivisões caracterizam-se por formar grupos com pouca identidade com relação à classe. A idéia por trás desse tipo de avaliação é que espera-se que a classificação evidencie as linhas gerais que fazem um elemento pertencer a uma determinada classe e isso é, pelo menos em princípio, facilitado quando conseguimos produzir agrupamentos mais organizados. Assim a questão sobre qual o melhor atributo para uma subdivisão (*split*) dos dados disponíveis pode ser redefinida em termos de se descobrir qual o atributo que permite dividir esses dados em grupos mais homogêneos com base em seu valor observado. Ou, em outras palavras, "qual o atributo mais informativo sobre a lógica dos dados num determinado contexto?".

Nesses termos, reduzimos o problema a uma questão sobre a medição da quantidade de informação presente em uma situação, e com isso passamos ao domínio da Teoria da Informação, que teve algum desenvolvimento no século passado com as pesquisas de Samuel Morse em torno do telégrafo e foi mais firmemente estabelecida a partir do trabalho de Claude Shannon na década de 1940. Temos então toda uma base teórica já consolidada que pode ser empregada na montagem do classificador, cujas linhas gerais serão mostradas a seguir.

Em primeiro lugar, o grau de informação de uma mensagem, como entendido pela Teoria da Informação, é dependente da probabilidade de sua ocorrência, no sentido que mensagens menos originais e mais freqüentes são, a rigor, menos informativas que outras mais raras e inesperadas. É possível até quantificar esse grau de informação: por exemplo, se temos n mensagens possíveis num dado contexto, todas elas eqüiprováveis ou seja, com probabilidade de ocorrência de $1/n$, a informação coberta por cada uma é dada por:

$$- \text{Log}_2 (1/n) = \text{Log}_2 (n) \qquad (3)$$

Retomando o problema da escolha de um atributo de *split* para um nó da árvore de decisão, estamos interessados em descobrir aquele que forme os subconjuntos mais homogêneos no que diz respeito ao atributo categórico. Em outras palavras, aquele que forma os grupos menos "confusos" com relação à classe. A Teoria da Informação utiliza um conceito

originado da Termodinâmica chamado Entropia, para representar o grau de confusão presente nos dados disponíveis. A entropia, que é inversamente proporcional ao grau de informação presente no contexto analisado, é expressa por meio de um valor situado entre 0 e 1, obtido pela fórmula

$$\text{Entropia}(S) = - \sum_{i=1}^{n} p_i \log_2(p_i) \quad (4)$$

onde S é a distribuição de probabilidade das n mensagens possíveis e p_i é a probabilidade de ocorrência da i-ésima mensagem. Supondo que se esteja construindo um classificador para um problema com duas classes possíveis, chamadas A e B, um atributo não categórico x vai permitir dividir os dados em tantos subconjuntos S quantos forem os seus possíveis valores. A entropia de cada um desses subconjuntos S_k seria então calculada, com base na fórmula anterior, por:

$$\text{Entropia}(S_k) = - p_A \log_2(p_A) - p_B \log_2(p_B) \quad (5)$$

O quadro a seguir mostra os valores de entropia obtidos para algumas probabilidades possíveis de A e B, e pode-se observar que quanto mais uniforme a distribuição de probabilidade, maior o grau de entropia. Se, no subconjunto analisado, todos os elementos pertencem a uma mesma classe, a entropia (ou "confusão") é mínima ou seja, zero.

p(A)	***p(B)***	***Entropia***
0,50	0,50	1,00
0,67	0,33	0,92
1,00	0,00	0,00

Agora vamos admitir, como exemplo, que o atributo x possa assumir apenas valores inteiros entre 1 e 3, o que leva à formação de três subconjuntos de S, cada um com seu próprio grau de entropia. É possível então avaliar a entropia em S quando considerado o atributo x através da média ponderada dos graus de entropia dos subconjuntos gerados (S_1, S_2 e S_3, no exemplo), o que é expresso na fórmula

$$\text{Entropia}(x, S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropia}(S_i) \quad (6)$$

Finalmente, após esses cálculos, podemos determinar em quanto a utilização do atributo x reduz a entropia original de S, o que significa determinar qual o ganho (*gain*) de informação devido a x na predição da classe quando partimos dos dados disponíveis no conjunto S. Essa medida, chamada de Ganho de Informação ou *Information Gain* é dada por:

$$\text{Ganho de Informação}(x, S) = \text{Entropia}(S) - \text{Entropia}(x, S) \quad (7)$$

A estratégia para determinar o melhor atributo de *split* resume-se então a calcular o ganho de informação para cada atributo não categórico disponível e escolher aquele que apresentar o maior valor, descartando-o em seguida do processo de escolha para os próximos níveis da árvore de decisão. Algumas variações nesse processo podem ser

encontradas, de maneira a equilibrar a importância de atributos com grande número de valores discretos possíveis, como o uso de *Gain Ratio* proposto por Quinlan (QUINLAN, 1993), ou permitir o tratamento de atributos não categóricos que possuem como domínio uma faixa de valores contínuos, por exemplo.

Abordagem baseada em métodos estatísticos

A seleção do atributo de *split* pode ser realizada por métodos que não empregam conceitos específicos da Teoria da Informação, mas fazem uma abordagem probabilística convencional do problema, produzindo diversos tipos de funções para a avaliação dos atributos disponíveis. Uma função de avaliação bastante comum pertencente a esta categoria é chamada de Índice Gini de Diversidade (*Gini Index*), que produz um valor para cada subconjunto com base na fórmula

$$\text{Gini}(S) = 1 - \sum_{i=1}^{n} p_i^2 \qquad (8)$$

onde p_i é a freqüência relativa da classe i em S.

Distribuições mais homogêneas ("confusas") correspondem a índices Gini maiores, a exemplo do que ocorre com a entropia. Adicionalmente, foram propostas também algumas variações para o cálculo desse índice, com o objetivo de minimizar eventuais desbalanceamentos nos subconjuntos gerados em alguns casos (BERSON, 1997; GUPTA, 1998).

Uma outra possibilidade para a identificação do melhor atributo de *split* é o uso de medidas mais tradicionais, como por exemplo o teste do Chi-Quadrado (χ^2) (CHEN, 1996).

Complexidade de um classificador

Além da acurácia do classificador, e conseqüentemente de sua árvore de decisão, existe um outro fator de qualidade igualmente importante, que é o seu grau de complexidade. Para um dado conjunto de treinamento é possível gerar mais de um classificador, eventualmente com graus de eficácia equivalentes entre si (WINSTON, 1992; WEBB, 1996). No caso específico dos classificadores baseados em árvore de decisão, o conjunto de soluções possíveis para um dado *training set* é chamado de floresta de decisão (*decision forest*) (MURPHY, 1994). Duas árvores dessa floresta terão, muito provavelmente, níveis de complexidade diferentes, conforme mostra a figura 3, que apresenta a árvore de decisão correspondente a um classificador alternativo para o caso das espécies vegetais.

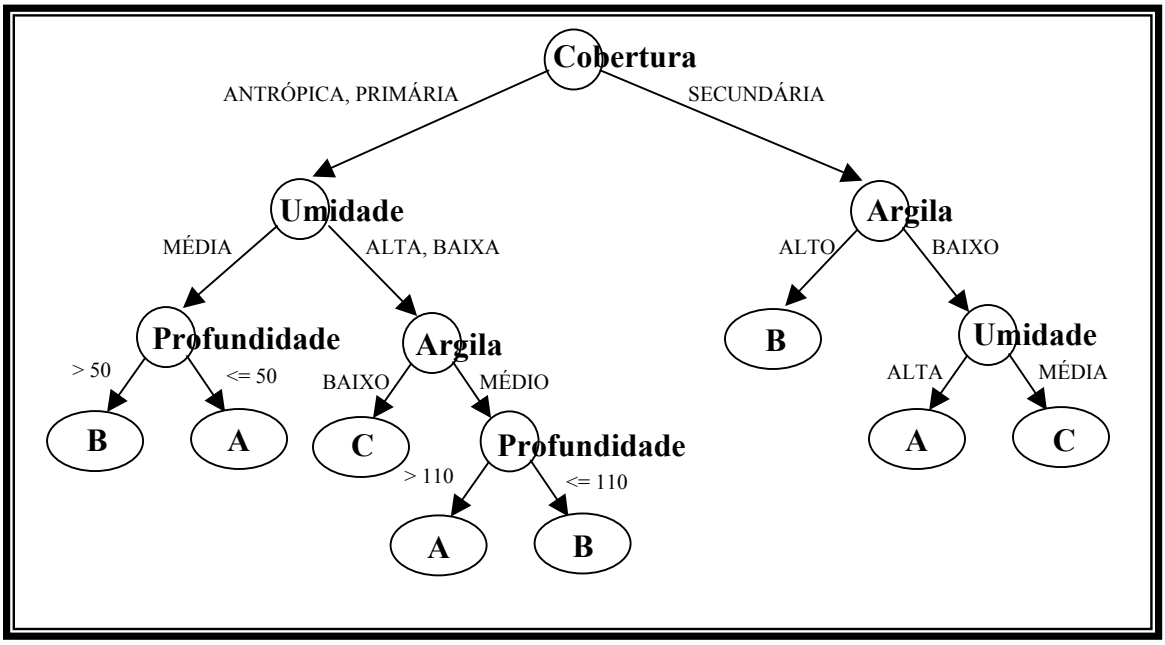


Figura 3: Representação de um segundo classificador possível para o caso das espécies vegetais.

Resta então escolher entre o classificador da figura 3 e o primeiro, da figura 2 e, nesse caso, costuma-se recorrer a um princípio geral vindo da filosofia, conhecido como 'Navalha de Occam'. Esse conceito, devido a um filósofo franciscano da Idade Média chamado Guilherme de Occam, diz que 'É inútil usar mais para fazer o que se pode fazer com menos', o que é uma tradução bastante livre do original em latim "*Entia non sunt multiplicanda praeter necessitatem*" (RUSSEL, 1995). A Navalha de Occam propõe que as coisas tendem a ser naturalmente simples e assim, de duas explicações alternativas para o mesmo fenômeno, a mais complicada tem mais chance de estar errada, pois baseia-se em um número maior de suposições (MAGEE, 1999; OLIVER, 1998). Portanto, se não houver diferença entre as duas explicações fornecidas pelas duas abordagens, a mais simples deve ser adotada. Atribui-se a Albert Einstein, inclusive, uma frase nessa mesma linha de pensamento: "Tudo deve ser tornado o mais simples possível, mas não mais simples que isso" (MAGEE, 1999). No caso específico das árvores de decisão a sua complexidade tem sido medida tradicionalmente através do seu tamanho, indicado pelo número de nós nas árvores univariadas. Nas árvores multivariadas a avaliação não é tão simples, e deve-se considerar nessa medida o grau de complexidade de cada nó, geralmente somando-se o número de atributos incluídos nos nós internos da árvore e também o seu número de folhas (BRESLOW, 1996). Desse modo, pode-se propor uma adaptação da Navalha de Occam para a avaliação de árvores de decisão: "(...) a menor árvore de identificação que é consistente com os exemplos é aquela que é a mais promissora para identificar corretamente objetos desconhecidos (...). Conseqüentemente, a questão transforma-se de 'qual é a árvore de identificação correta' para 'como se pode construir a menor árvore de identificação'" (WINSTON, 1992), o que é válido também para as árvores de decisão. Cabe aqui um esclarecimento sobre o termo "árvore de identificação": ele refere-se a uma forma alternativa para a representação de uma árvore de decisão onde o conteúdo de uma folha não é a classe e sim o conjunto de exemplos do *training set* utilizado que se encaixa

naquele caminho raiz-folha específico. Dessa forma, numa árvore de identificação a informação sobre a classe correspondente a cada folha é expressa de maneira implícita, por meio dos exemplos relacionados na folha.

Essa perspectiva 'minimalista' no tratamento da prospecção de conhecimento é largamente empregada na mineração de dados, embora existam alguns estudos que procuram fazer uma análise mais crítica de sua utilização indiscriminada (WEBB, 1996). Abordagens como MDL (*Minimum Description Length principle*, ou princípio da Descrição de Tamanho Mínimo) e MML (*Minimum Message Length*, princípio da Mensagem de Tamanho Mínimo) são consoantes com a Navalha de Occam e forneceram, juntamente com o avanço da teoria da aprendizagem computacional uma fundamentação teórica para essa perspectiva (RAMAKRISHNAN, 1999). A adoção desse critério de complexidade ligado ao tamanho da árvore é particularmente importante ao se gerar classificadores, na medida em que árvores muito grandes tendem a ser encaradas como "caixas-pretas" pelo usuário final do processo (MICHALSKI *et al.,* 1998; CHEN, 1996; QUINLAN, 1996), perdendo-se com isso talvez o maior apelo desta forma de representação, que é a sua inerente inteligibilidade.

Na prática observa-se que a produção de árvores de decisão mais enxutas é favorecida pelo uso de heurísticas que privilegiam os atributos mais significativos nos ramos mais próximos da raiz através, por exemplo, do uso da entropia ou índice Gini na seleção do atributo de *split*.

Poda de classificadores

A indução de árvores de decisão é baseada nas observações sobre o conjunto de treinamento, onde cada nó é definido em função de algum critério que procura identificar o atributo mais significativo para a predição da classe em um particular nível da árvore. Como já foi mencionado anteriormente, esse processo tem características que fazem com que ao se produzir os nós finais, diretamente ligados às folhas, o número de elementos disponíveis para avaliação seja muito pequeno, às vezes chegando à unidade. Isso coloca em questão a representatividade desses conjuntos finais, uma vez que seu tamanho reduzido muitas vezes não permite afirmar que os valores de classe ali presentes não o sejam apenas devido ao acaso. Essa questão é crítica na medida em que se espera que com base nesse treinamento seja possível posteriormente classificar casos novos, originais, e se nos baseamos em condições pouco seguras, o resultado final falhará quanto à precisão. Um outro fator que contribui para esse tipo de problema é a presença de incorreções ou "ruídos" (*noise*) nos dados ou ainda de atributos irrelevantes, cujo efeito se torna estatisticamente mais pronunciado nos nós mais distantes da raiz, fazendo com que a árvore de decisão resultante acabe modelando também essas anomalias presentes no conjunto de treinamento, um problema chamado de *overfitting*. Outras vezes a árvore produzida é demasiado complexa, impedindo uma melhor compreensão da estrutura interna do classificador. Finalmente, estudos relativamente recentes (JENSEN, 1997; OATES, 1999) indicam que o tamanho da árvore de decisão aumenta à medida que o conjunto de treinamento cresce, mesmo que esse crescimento não traga nenhuma melhoria significativa nos seus níveis de acurácia. Uma estratégia para resolver ou minimizar todos esses problemas é a poda (*pruning*) das árvores de decisão (MICHALSKI *et al.,* 1998), um processo que pode ser realizado segundo diversos critérios e em vários momentos e cujos efeitos afetam tanto a acurácia quanto a complexidade do classificador.

Durante a poda de uma árvore de decisão substitui-se por um nó folha cada subárvore cujo erro esperado é maior que o erro previsto para a respectiva folha substituta. A estimativa do erro pode ser feita por meio de uma variedade de cálculos, como por exemplo o estimador-*m*, o estimador de Laplace, etc (MICHALSKI *et al.,* 1998). Assim, suponha que analisando o classificador elaborado inicialmente e apresentado na figura 2, identificamos, apoiados em alguma evidência qualquer verificada sobre o conjunto de treinamento, que a distinção feita com base na profundidade para as amostras de maior altitude e menor umidade é questionável por apresentar uma taxa de erro esperado relativamente elevada. Nesse caso poderíamos podar a árvore original, substituindo a subárvore correspondente ao teste da profundidade por uma folha afirmando que o valor esperado é 'B', por exemplo, conforme mostra a figura 4 a seguir. Com isso, passaríamos a ter ocorrências do *training set* classificadas com erro, mas a expectativa de acertos nos conjuntos novos aumentaria.

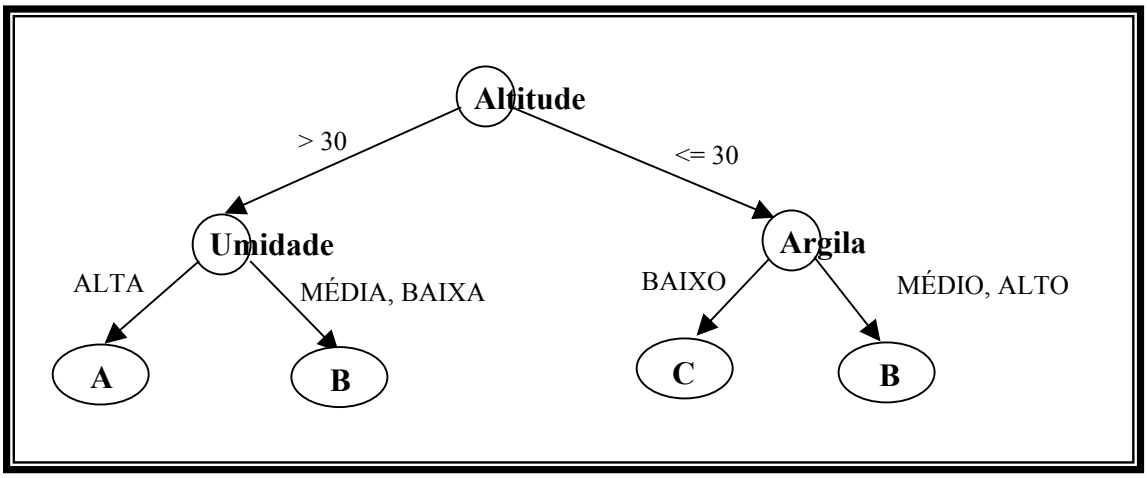


Figura 4: A árvore da figura 2 após uma poda.

Ainda que esse exemplo seja ilustrativo do conceito da poda, ele representa apenas uma das possíveis formas, chamada de poda posterior (*post pruning*), para se atingir o objetivo proposto. Existem também estratégias de poda que são embutidas no próprio processo de construção da árvore e são chamadas de poda prévia (*pre-pruning*) e uma visão geral dos principais aspectos dessas abordagens é apresentada a seguir.

Poda prévia (*pre-pruning*)

O algoritmo TDIDT mostrado anteriormente utiliza em seu passo 3 um critério de parada que só será satisfeito quando todos os nós folha corresponderem a elementos de uma mesma classe, ou seja, a árvore é desenvolvida exaustivamente e nenhum elemento do *training set* deixa de ser classificado corretamente. Esse é o chamado de critério de homogeneidade, de concepção bastante simples e que leva à construção de árvores que podem padecer dos problemas de *overfitting* já mencionados. Uma abordagem mais cautelosa seria avaliar, a cada nível da árvore, se o ganho de informação esperado com o desdobramento dos próximos níveis é maior que um determinado valor: se for, o processo avança, senão, um nó folha é criado para representar a classe mais provável para os elementos remanescentes e a árvore é dada como finalizada. A avaliação embutida nesse critério de parada pode empregar medidas de natureza diferente daquela utilizada para

selecionar o atributo de *split*, de maneira a evitar uma interrupção prematura ou tardia do processo de indução do classificador.

A poda prévia ocorre portanto durante o crescimento da árvore e evita que desdobramentos pouco significativos sejam gerados o que, de alguma forma, acelera o processo de indução.

Poda posterior (*post-pruning*)

Esta modalidade de poda é executada quando a árvore já foi totalmente desenvolvida, e comporta uma série de variações importantes. Algumas estratégias procuram dividir o conjunto de treinamento em duas partes, uma para a indução da árvore e outra para a poda, chamadas respectivamente de conjunto de crescimento (*growing set*) e conjunto de poda (*pruning set*), de maneira que as subárvores a serem descartadas não o sejam com base apenas em estimativas, mas também através de testes reais sobre os dados. Uma outra distinção pode ocorrer com relação à forma de avaliação dos erros a serem eliminados, havendo estimativas baseadas no cálculo do desvio padrão, outras através de mecanismos de validação cruzada. Finalmente, algumas estratégias de poda percorrem a árvore a partir da raiz, outras a partir de suas folhas. A seguir é apresentado um breve resumo dos principais métodos de poda posterior de árvores.

MCCP (*Minimal Cost-Complexity Pruning* – Poda do custo de complexidade mínimo) é o método de poda mais antigo, tendo sido proposto por Leo Breiman em 1984 como elemento do sistema CART (BRESLOW, 1996). A poda é baseada em um índice de custo que considera a somatória dos erros da árvore, seu número de folhas e o valor de um parâmetro definido pelo usuário para expressar um compromisso entre complexidade e custo. A árvore original induzida pela ferramenta é avaliada e podada com base no seu índice de custo, dando origem a uma nova árvore, que sofre o mesmo processo, até que reste apenas uma árvore, correspondente ao nó raiz. O conjunto de árvores gerado é então avaliado e a melhor árvore é escolhida, com base tanto em seu tamanho como em sua acurácia sobre o conjunto de treinamento.

REP (*Reduced Error Pruning* – Poda de erro reduzido) adota uma estratégia em que o conjunto de treinamento é dividido em duas partes, uma para a criação da árvore e outra específica para a poda. Uma vez induzida a árvore, a poda é realizada a partir dos seus ramos finais (*bottom-up*), descartando-se os nós cuja eliminação não diminuam a acurácia sobre o conjunto de poda.

EBP (*Error Based Pruning* – Poda baseada em erro) é um método pessimista proposto por Ross Quinlan que parte do princípio que as estimativas de erro sobre o *training set* tendem a valores menores do que deveriam ser. O processo, que não utiliza conjunto de poda, realiza uma estimativa do erro em um nó mais negativa do que o normal, considerando as amostras a que ele se refere como sendo binomialmente distribuídas e dispondo de limites de confiança bem definidos. A taxa de erro esperado corresponde ao limite superior de sua distribuição de probabilidade.

Não há consenso sobre a existência de um método de poda que seja invariavelmente superior aos demais, o que faz com que a escolha de uma determinada abordagem para a redução de uma árvore de decisão seja dependente de uma série de fatores, inclusive do volume de amostras disponíveis para o *training set*. Além dos mencionados, há ainda um grande número de outros métodos de poda importantes, como MEP (*Minimum Error*

Pruning), CVP (*Critical Value Pruning*), PEP (*Pessimistic Error Pruning*), MDL (*Minimum Description Length*), etc (BRESLOW, 1996).

A simplificação das árvores de decisão, realizada com o objetivo tanto de melhorar sua acurácia nos casos não vistos como também torná-la de mais fácil compreensão pelo usuário final, não se restringe apenas a estratégias de poda. Uma interessante categorização das diversas possibilidades de simplificação pode ser encontrada em (BRESLOW, 1996), que identifica 5 abordagens principais:

a) controle do tamanho da árvore, onde se encaixam as estratégias de poda;

b) modificação do espaço de testes;

c) modificação do espaço de busca, através da seleção de medidas e do tratamento de valores contínuos;

d) aplicação de restrições sobre o banco de dados onde residem os dados de treinamento e teste, selecionando os casos a serem processados e os atributos a serem considerados no processo;

e) utilização de outras estruturas que não árvores de decisão, como regras e gráficos de decisão.

A maioria dessas abordagens pode ser utilizada em conjunto, de maneira a somar suas vantagens individuais e melhorar os resultados finais do processo de classificação.

Estratégias baseadas em votação (ou montagens) de classificadores

Ainda que um classificador seja criado em condições ideais, com dados representativos e possua uma alta acurácia em seus resultados, uma taxa de erro estará sempre envolvida, seja devido aos dados utilizados no seu treinamento, seja devido à natureza das medidas utilizadas como critério de *split*, seja ainda devido às características próprias do algoritmo empregado. Em muitas situações, entretanto, pode-se obter um resultado final mais preciso pela utilização combinada de vários classificadores distintos, cujos resultados são reunidos e considerados para a elaboração da resposta final a ser fornecida. Isso ocorre particularmente nos casos em que há uma instabilidade natural no algoritmo de indução dos classificadores, onde relativamente pequenas mudanças nos dados de entrada causam relativamente grandes diferenças no classificador gerado (DIETTERICH, 2000). Tais abordagens tem sido bastante exploradas ultimamente e podem ser usadas, com diferentes níveis de ganho final (BAUER, 1999), por algoritmos baseados em árvores de decisão, redes neurais, etc. No momento as principais estratégias de uso combinado de classificadores são as denominadas *Boosting* e *Bagging*, descritas resumidamente a seguir.

Boosting

Esta técnica consiste na construção de um determinado número de classificadores, com base no mesmo conjunto de treinamento (*training set*), onde foi introduzida uma nova propriedade numérica indicando o peso da instância para fins de classificação. Assim, um primeiro classificador é gerado a partir dos exemplos, onde o peso inicial é igual para todas as instâncias, e seus erros são avaliados de maneira que as ocorrências que foram classificadas incorretamente recebam um valor de peso maior que as demais. É então gerado um novo classificador, que dará algum tipo de prioridade para as instâncias de

maior peso, com o processo se repetindo tantas vezes quantas forem desejadas. No final temos um conjunto de classificadores que são considerados para a produção do resultado final: uma nova ocorrência a ser classificada é submetida a todos os classificadores, que indicam o seu valor de classe para aquela instância, e o resultado final é compilado a partir das diversas respostas obtidas, ganhando aquela mais freqüente, numa espécie de votação. Em algoritmos como *AdaBoost* o voto de cada classificador possui também um peso, que é determinado com base na sua acurácia sobre o *training set* (FREUND, 1999); e em outras abordagens, como *Arc-x4*, o voto de cada classificador tem a mesma importância que o dos demais (BAUER, 1999).

Bagging

A denominação é uma composição sobre as palavras **B**ootstrap **agg**regat**ing**, e refere-se a uma abordagem proposta por Leo Breiman em 1996 para aumentar a acurácia de classificadores. A estratégia baseia-se na geração de diversos classificadores, cada um deles construído com base em seu próprio conjunto de treinamento (*training set*), extraído das amostras aleatoriamente e com reposição O classificador final é uma composição dos classificadores originais participantes, de maneira que para uma dada entrada, todos os componentes produzem seu próprio resultado, e o valor de classe a ser retornado será aquele que possuir maior votação, ou seja, o que for predito pelo maior número de participantes. Existem implementações dessa estratégia para diversas técnicas de classificação, como redes neurais, árvores de decisão, k-*nearest neighbor*, etc (BAUER, 1999). As principais vantagens desta abordagem são a possibilidade do processamento paralelo na geração dos classificadores, impossível com *Boosting* e em geral um melhor resultado em situações com muito ruído nos dados de treinamento (DIETTERICH, 2000).

Outras abordagens

Existem algumas variações sobre as técnicas básicas dos classificadores baseados em votação como *Bagging* e *Boosting*, e que são implementadas em diversos algoritmos. Basicamente elas consistem em introduzir pesos em variações de *Bagging* (*Wagging*); colocar re-amostragem em algumas variações de *Boosting*; implementar ou não estratégias de poda nos classificadores e recalcular o erro estimado do classificador (*backfitting*). Cada variação procura potencializar as características de uma dada abordagem de acordo com o tipo de dado a ser processado, não havendo uma solução que seja indiscutivelmente a melhor para todos os problemas (BAUER, 1999).

Capítulo 3: Materiais e metodologia

A avaliação das possibilidades do uso de uma ferramenta de indução automática de regras de classificação para a identificação de unidades ambientais, definida como o objetivo deste trabalho, foi conduzida na forma de um estudo de caso sobre uma pesquisa ambiental realizada anteriormente e que não fez uso desse tipo de ferramenta. A escolha do trabalho acadêmico a ser utilizado como referência neste estudo revestiu-se de grande importância, pois deveria ilustrar convenientemente o trabalho de um especialista no domínio da ecologia e da geografia física, possibilitando:

a. identificar as tarefas analíticas típicas envolvidas nesse classe de atividade, para que fosse possível investigar quais delas são adequadas à utilização do método computacional selecionado e porque;

b. validar os resultados obtidos pela automatização das análises escolhidas;

c. fornecer o conjunto de dados reais indispensável à aplicação efetiva do método computacional.

Foi escolhida a monografia "Análise Geomorfológica e Distribuição Espacial da Vegetação na Planície Litorânea de Picinguaba (Ubatuba - SP)", de autoria de José Paulo Marsola Garcia, apresentada à Universidade de São Paulo como dissertação de mestrado em Geografia Física no ano de 1995. Esse trabalho, atende aos requisitos básicos identificados para a pesquisa pois, além de ser relativamente recente, aponta para um tipo de estudo de grande representatividade quanto ao que se faz atualmente em pesquisa do meio-ambiente. Com relação à ferramenta computacional a ser utilizada, optou-se por uma versão do programa See5, um produto comercial desenvolvido a partir dos clássicos algoritmos de indução de classificadores ID3 e C4/5 de Ross Quinlan.

Este capítulo procurará apresentar e descrever os métodos e recursos que foram utilizados e também discutir alguns aspectos conceituais e de ordem prática considerados importantes e que causaram um forte impacto no desenvolvimento do trabalho e nos resultados finais atingidos, constituindo-se em elementos de alto grau de interesse para outros trabalhos similares. Inicialmente, a pesquisa ambiental adotada como base para o estudo de caso será apresentada, com seus objetivos, estratégias e resultados principais. Em seguida, a ferramenta de indução de regras utilizada será descrita quanto às suas características mais relevantes. Uma discussão sobre alguns aspectos quantitativos do problema em questão será então realizada, seguida de uma descrição do banco de dados que foi gerado para a avaliação. O capítulo termina com a apresentação dos experimentos a serem realizados com o auxílio da ferramenta sobre banco de dados.

3.1 O trabalho de referência

O estudo denominado "Análise Geomorfológica e Distribuição Espacial da Vegetação na Planície Litorânea de Picinguaba (Ubatuba - SP)" corresponde à dissertação de mestrado em Geografia Física apresentada por José Paulo Marsola Garcia à Universidade de São Paulo em 1995, e será aqui descrito em suas linhas gerais, com base tanto no material

publicado originalmente na versão final da monografia como também em comunicações pessoais diversas do autor, ocorridas em várias entrevistas.

Esse trabalho tem como objetivo principal verificar as possibilidades do mapeamento geomorfológico de detalhe como instrumento para o reconhecimento e a caracterização de relações de causalidade entre o meio físico e a vegetação. Isso imprime a essa iniciativa uma forte característica interdisciplinar, típica de estudos mais recentes sobre o meio ambiente, articulando elementos de geologia, biologia, geografia e ecologia, principalmente. Um outro aspecto de importância do trabalho de Garcia é que, até o momento de sua realização, poucas propostas de emprego da análise geomorfológica na classificação vegetal deixavam o nível teórico e ganhavam uma implementação efetiva, principalmente no nível de detalhe adotado (1:10.000). O trabalho justifica-se ainda por viabilizar um conhecimento dos limites de exploração de um ambiente complexo e sensível, fundamental para a definição de futuras propostas de manejo e preservação.

A pesquisa foi realizada em uma reserva natural do litoral paulista e consistiu em um detalhado mapeamento do local, na identificação e caracterização de suas diversas regiões naturais e no estudo das relações de distribuição de espécies de orquidáceas por essas áreas, inserindo-se num contexto mais amplo denominado "Projeto de Florística e Biossistemática do Núcleo de Desenvolvimento de Picinguaba (Ubatuba - SP)", sob responsabilidade e coordenação do Departamento de Botânica do Instituto de Biociências da UNESP, campus de Rio Claro. O mapeamento e a classificação regional desenvolvidos no trabalho têm sido utilizados desde então em diversas atividades desse projeto maior e, em suas considerações finais, Garcia conclui, entre outras coisas, que há um grande potencial da análise geomorfológica para caracterizar unidades ambientais tendo a vegetação para abalizar esses limites.

3.1.1 O local

A região de estudo é denominada Planície Litorânea de Picinguaba, localizada ao norte do município paulista de Ubatuba, e constitui-se em parte integrante do Parque Estadual da Serra do Mar, apresentando uma área de aproximadamente 8.2 quilômetros quadrados, mostrada na figura 5. O local foi escolhido por apresentar-se em condições relativamente boas de conservação e por, na época, abrigar uma equipe de pesquisadores em botânica da UNESP, que seriam indispensáveis para as tarefas de identificação da flora requeridas pelo projeto.

3.1.2 Metodologia utilizada

O levantamento do meio físico com propósitos de se correlacioná-lo com elementos bióticos, como o pretendido no trabalho de Garcia, geralmente é realizado por meio da classificação de solos. Entretanto, devido a fatores como a dificuldade em se reconhecer unidades individuais do solo na escala adotada no trabalho (1:10.000) e a questões decorrentes da falta de um consenso quanto a critérios mais firmemente estabelecidos para sua classificação, entre outros, foi adotada a análise geomorfológica como método de classificação de terreno. Essa escolha, segundo o autor, encontra eco em outros trabalhos internacionais de levantamento ambiental e define as principais caraterísticas metodológicas do estudo desenvolvido.

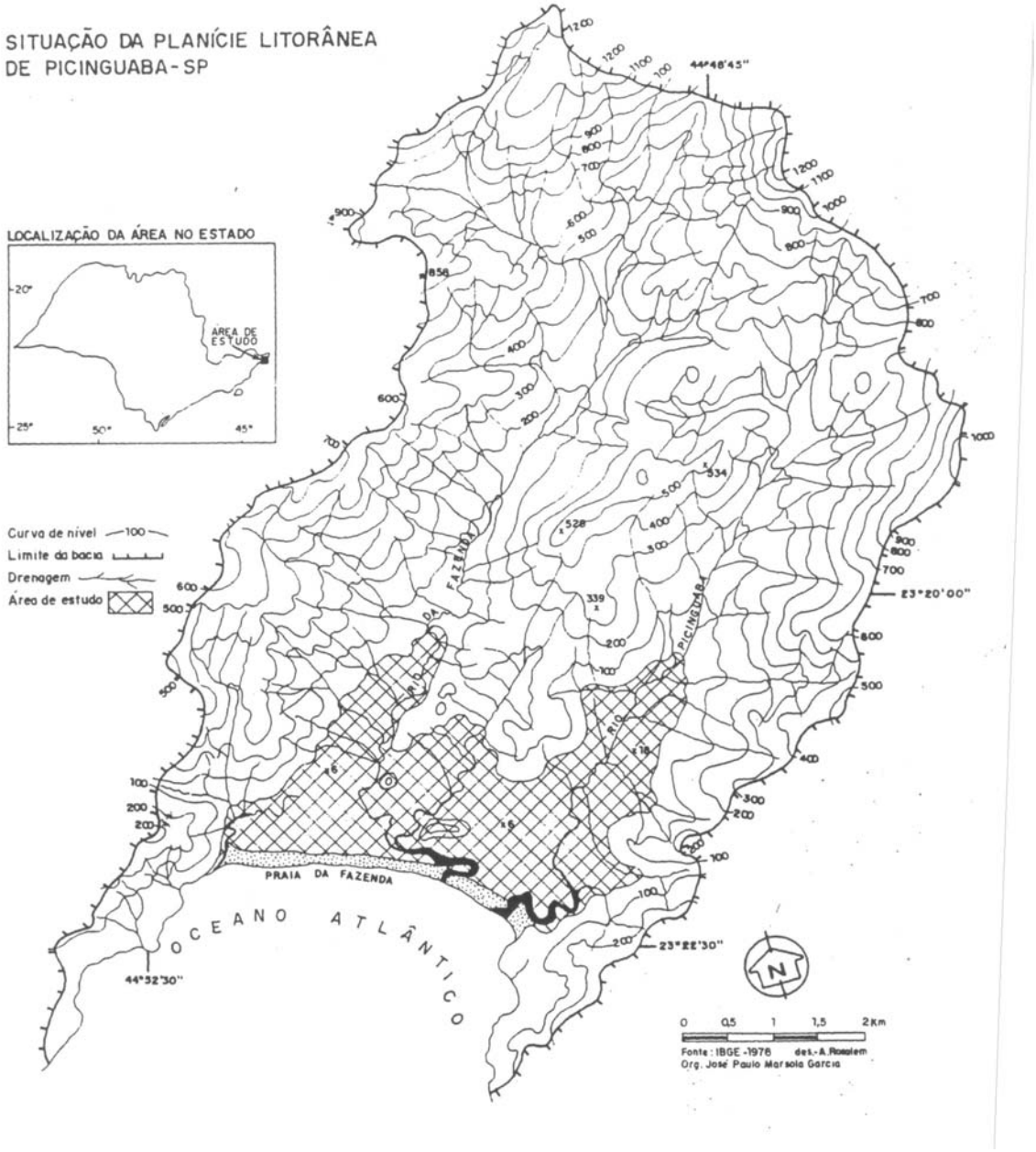


Figura 5: Localização da área de estudo. Adaptado de (GARCIA, 1995).

A análise geomorfológica em escala de detalhe procura delinear feições de relevo na dimensão de centenas de metros a poucos quilômetros, chamadas de Superfícies Geneticamente Homogêneas (SGHs), com base em três fatores: inclinação do terreno; sua origem e sua idade. Em áreas planas, como é o caso da Planície Litorânea de Picinguaba, a inclinação muito pequena e geralmente uniforme e a dificuldade no estabelecimento da idade precisa das formações, devido tanto ao seu custo como à quase ausência de carvão para a datação por carbono 14, tornam determinante o fator origem do terreno que, no caso, está ligada principalmente a processos de sedimentação. Adicionalmente, a análise da composição física do material de superfície é também uma importante aliada na diferenciação das unidades geomorfológicas de áreas muito planas.

A identificação das feições geomorfológicas foi realizada com base na análise de fotografias aéreas da região, com posterior comprovação das observações em campo. Além do delineamento das SGHs, a etapa de fotointerpretação propiciou a definição de uma estratégia de amostragem estratificada orientada, considerada mais adequada para trabalhos que envolvem aspectos espaciais, onde cada SGH corresponde a um estrato. Assim, com base no zoneamento obtido pela fotointerpretação foram selecionados 22 pontos distintos de onde foram extraídas amostras de material para análise de sua composição física e que também serviram como pontos para a conferência do processo de fotointerpretação.

As amostras foram retiradas a profundidades fixas, de 10 em 10cm até 50cm de profundidade e de 20 em 20cm desse nível até ser atingido o lençol freático no local da escavação, o que levou à extração de cerca de 160 amostras no total. O material devidamente acondicionado em sacos plásticos, etiquetado e anotado em cadernetas de campo propiciou posteriormente a análise física de suas características em três perspectivas distintas:

a) macroscópica, realizada sobre o material seco através de lupas especiais, observando aspectos de cor, textura, composição mineralógica e seu grau de seleção;

b) granulométrica, feita em laboratório de sedimentologia, determinando inicialmente as proporções de Areia Total, Silte e Argila do material amostrado. Posteriormente, a fração Areia Total foi decomposta em 5 classes granulométricas, conforme a Escala de Wentworth, sendo esse maior detalhamento das areias um fator diferenciador da abordagem adotada neste trabalho em relação aos demais e também um elemento de extrema importância nos resultados obtidos pelo autor;

c) química, realizada sobre 65 das amostras coletadas até 40cm de profundidade, procurou determinar a sua composição em termos das proporções observadas de carbono, matéria orgânica, pH em água e em cloreto de potássio, nitrogênio, fósforo, potássio, sódio, alumínio, cálcio e magnésio. A não utilização de todas as amostras neste tipo de análise deve-se a fatores de custo e também porque os nutrientes do solo concentram-se em maior quantidade nas suas camadas superiores. Os resultados obtidos foram disponibilizados para outras pesquisas de biossistemática, florística e fitossociologia conduzidas no mesmo local.

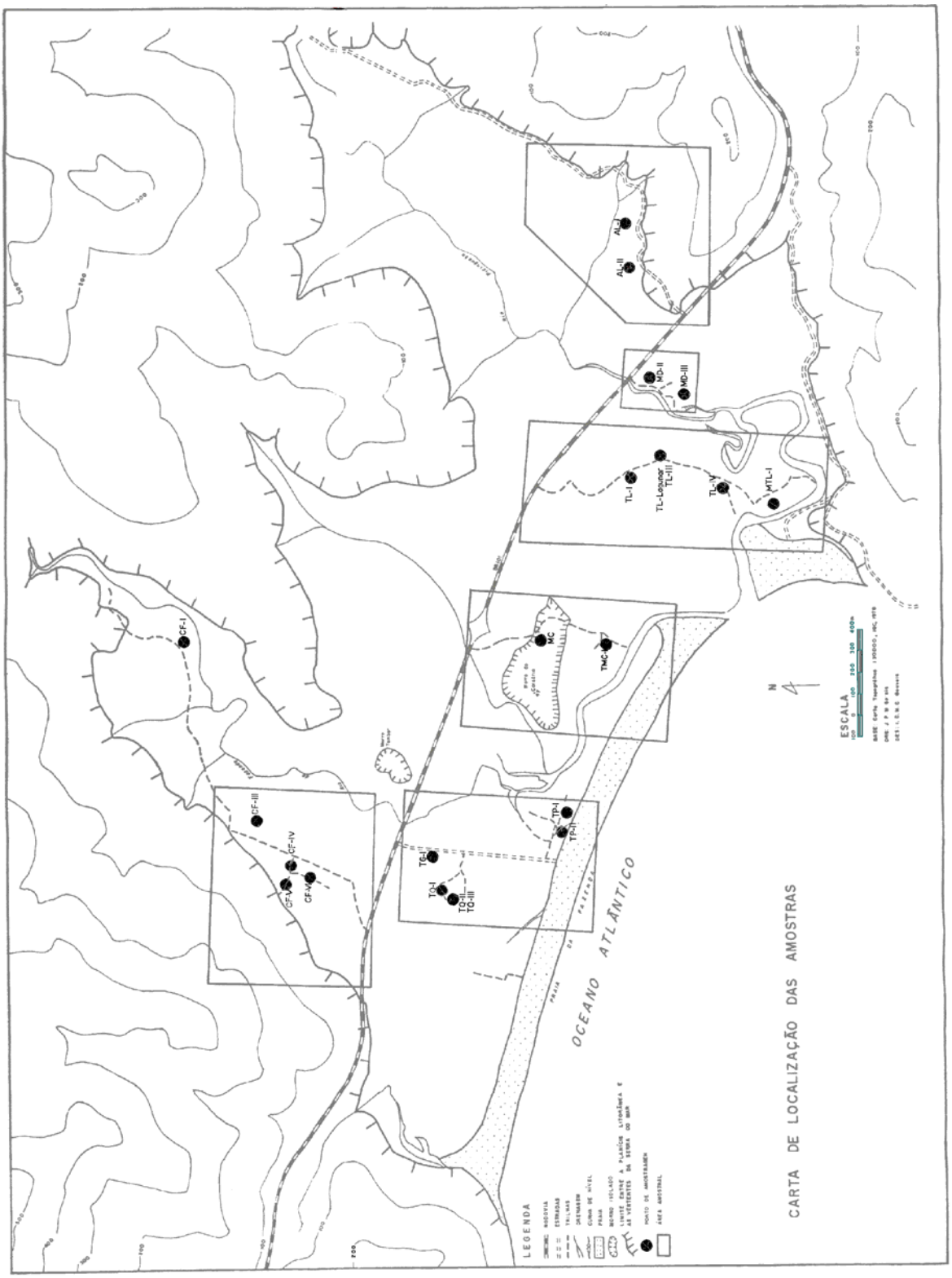


Figura 6: Localização dos pontos de coleta de material. Adaptado de (GARCIA, 1995).

Os resultados finais dessas análises foram depois comparados com a classificação dos terrenos em SGHs, esperando-se encontrar algum tipo de perfil que permitisse afirmar que as feições geomorfológicas identificadas anteriormente com base principalmente na fotointerpretação (e que procuraram considerar também inferências sobre sua dinâmica e processo de origem) possuem uma certa identidade em termos da composição física de seus solos, sendo que essa expectativa é baseada em duas suposições fundamentais. A primeira delas assume que "diferentes agentes de transporte fracionam os resíduos de maneiras diferentes, fazendo com que depósitos sedimentares (modernos), conforme os meios de transporte e deposição, passem a apresentar maior predomínio de seixos, de areias, de siltes, ou de argilas" (GARCIA, 1995, p29). A segunda suposição diz que "Tanto a maturidade textural (física) como a mineralógica (química), ocorrem durante a história de transporte de uma população arenosa e estão intimamente relacionados entre si. Portanto, em geral uma areia fisicamente matura (menor diâmetro, mais selecionada e com alto grau de arredondamento) é também quimicamente matura (minerais mais resistentes ao intemperismo) e vice-versa. Isso porque a composição mineralógica é essencialmente dependente da proveniência, enquanto a textural é o resultado de processos de deposição e transporte" (GARCIA, 1995, p30).

Finalmente, uma análise dos padrões de distribuição florística foi também realizada, com o objetivo de aferir a representatividade ecológica do mapeamento geomorfológico de detalhe. Para isso foi escolhida a família vegetal *Orchidaceae*, por ser a de maior representatividade na área de estudo, conforme mostra o gráfico a seguir, e por também contar com bibliografia afirmando seu potencial como indicador biológico das condições ambientais. Merece atenção ainda o fato de que a maioria das subfamílias e tribos de orquidáceas ocorrentes no Brasil são encontradas em Picinguaba, o que torna essa área propícia para estudos taxonômicos e ecológicos com a família por representar uma significativa parcela da diversidade morfológica e das estratégias adaptativas desse grupo tão complexo, conforme afirma Garcia em seu trabalho.

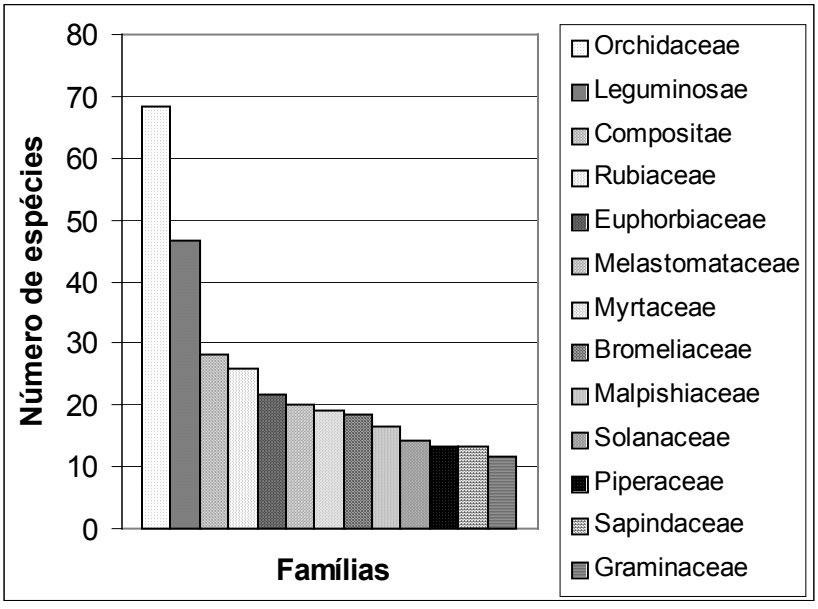


Gráfico 1: Número de espécies das famílias mais representativas da flora vascular da restinga do Núcleo Picinguaba. Adaptado de (GARCIA, 1995).

3.1.3 Resultados obtidos

Com base nas considerações iniciais e na metodologia anteriormente descrita em suas linhas gerais, a Planície Litorânea de Picinguaba foi então dividida nas seguintes Superfícies Geneticamente Homogêneas:

SGH	***Nome***	***Idade Relativa***[1]
I	Planície de Maré	8
II	Berma	9
III	Duna	7
IV	Planície Litorânea de Cordões Regressivos	1
V	Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes	3
VI	Planície de Retrabalhamento Fluvial-Marinho	6
VII	Planície Colúvio-Aluvionar	2
VIII	Planície Colúvio-Aluvionar com Micro-Canais Interligantes	4
IX	Áreas Alteradas	10
X	Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos	5

1 – áreas mais antigas aparecem indicadas com os números menores

Tabela 2: Superfícies Geneticamente Homogêneas da Planície de Picinguaba.

Como pode ser observado na Tabela 2, foi determinada também qual a ordem de antigüidade de cada SGH, com base na análise de seus sedimentos e em sua morfogênese. Levando em conta os dois pressupostos anteriormente citados, essa classificação é um elemento de grande importância para as conclusões obtidas.

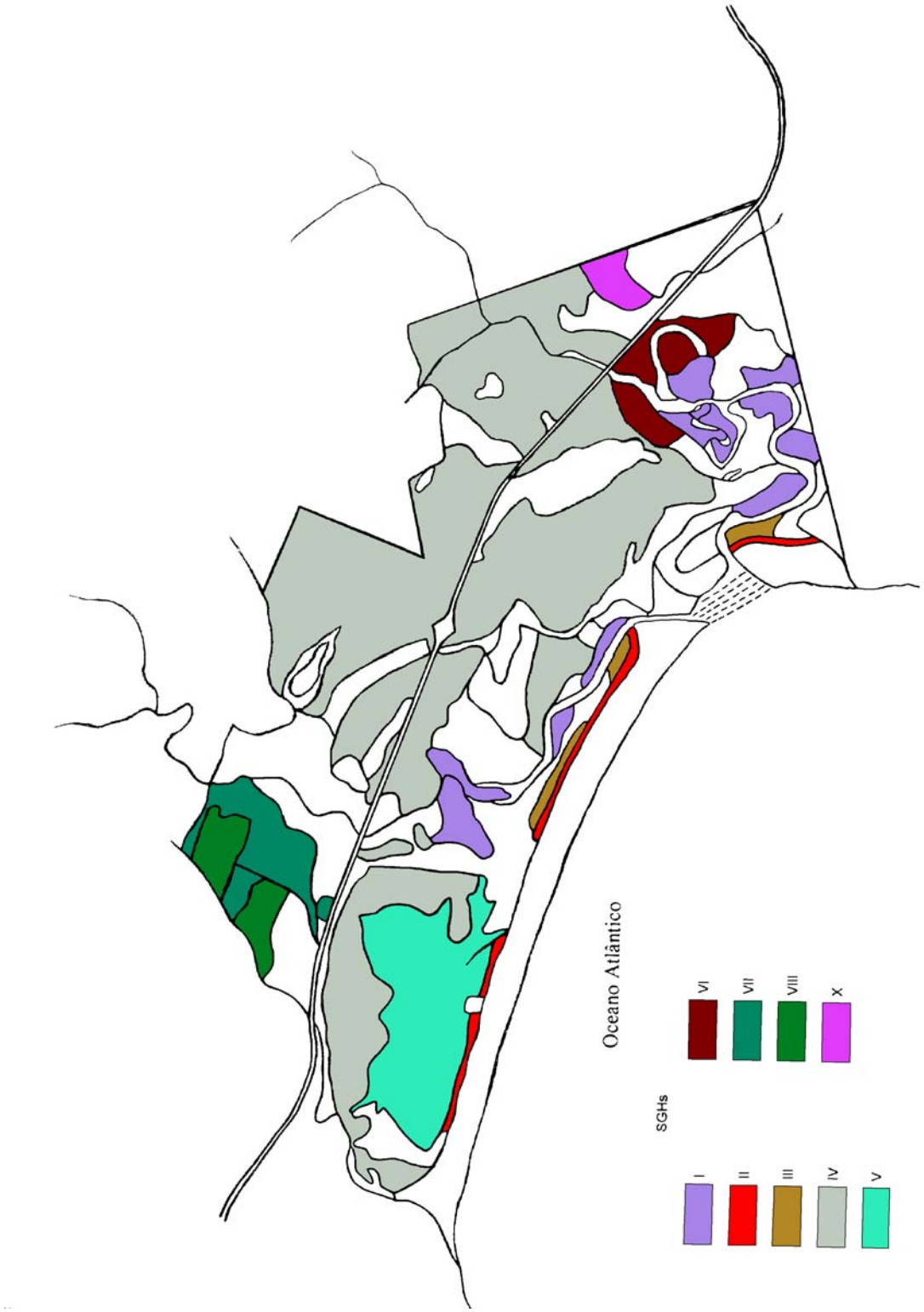
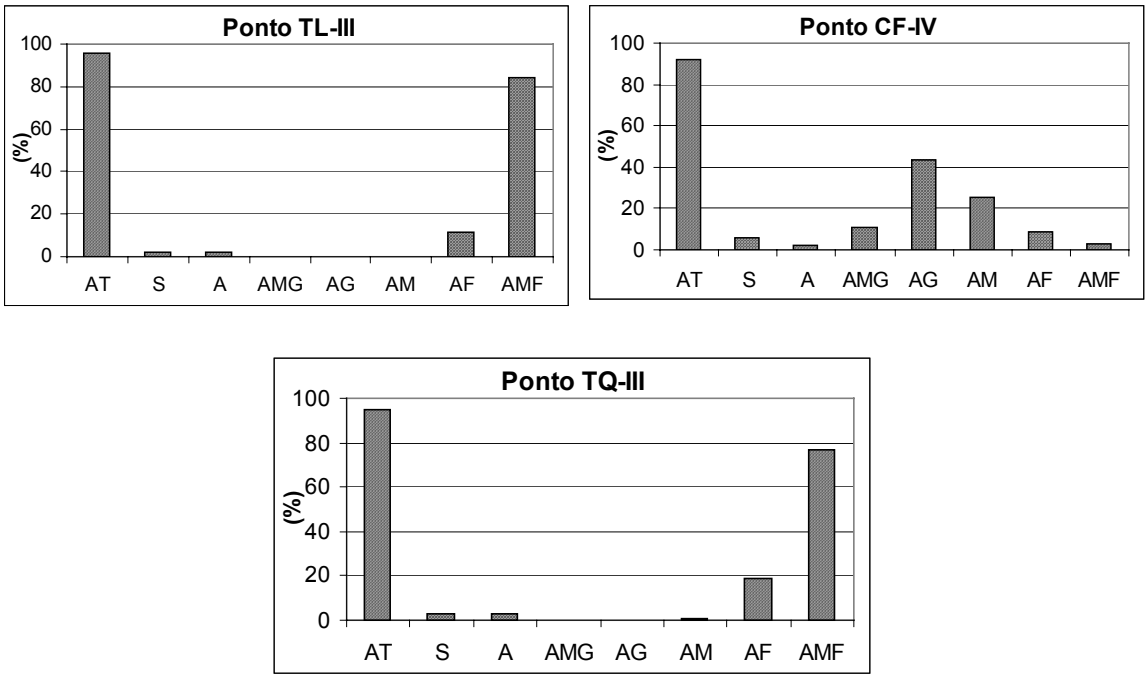


Figura 7: Delimitação das SGHs. Adaptado de (GARCIA, 1995).

A comparação entre a classificação realizada pelo mapeamento geomorfológico e os dados das análises de composição física dos terrenos mostrou, segundo o autor, fortes indícios de uma identidade própria para cada SGH com base em suas características macroscópicas e granulométricas. As evidências foram consideradas tão fortes a ponto de praticamente eliminar a necessidade de confrontos também com os dados da análise química, que não foram então realizados inclusive por questões ligadas à falta de tempo nos estágios finais do projeto. Os três gráficos apresentados a seguir foram extraídos da versão final da dissertação e constituem um exemplo desse tipo de identidade.



Gráficos 2, 3 e 4: Dados granulométricos dos pontos de coleta TL-III, CF-IV e TQ-III.

Com base apenas na subdivisão Areia Total (AT), Silte (S) e Argila (A), observa-se uma grande semelhança entre as suas estruturas, que não permite fazer com segurança uma distinção significativa entre eles. Entretanto, se for observada as distribuição das areias nos seus diversos níveis de granularidade (de AMG até AMF), percebe-se uma grande diferença entre as três estruturas, que fazem parte de SGHs distintas. O ponto TL-III, que pertence à SGH "IV" (a mais antiga) possui a estrutura granulométrica mais fina, com forte predomínio de areias finas (AF) e muito finas (AMF). No ponto CF-IV, pertencente à SGH "VII" (de idade relativa posterior à SGH "IV"), por sua vez, predominam areias médias (AM) e grossas (AG), parecendo afirmar que estruturas mais antigas são mais maduras texturalmente. Entretanto, no ponto TQ-III, que pertence à SGH "V" (a mais recente das três), o padrão granulométrico parece mais próximo daquele observado no ponto TL-III, o mais antigo. Isso tem a ver com a natureza dos processos de criação e dinâmica atuantes nos terrenos (e que em parte podem ser inferidos a partir de elementos da análise macroscópica, não apresentados no gráfico), conforme mencionado anteriormente, o que coloca então um importante componente de classificação que situa-se além da simples leitura dos resultados da granulometria, e que por suas características, exige uma análise muito mais elaborada da situação. Assim observa-se que a granulometria não é suficiente,

por si só, para atestar a identidade física de um terreno nos termos das SGHs mas, como bem enfatizado pelo autor na página 29 de seu trabalho, pode constituir-se em "(...) um INDICADOR, para auxiliar a caracterização das feições geomorfológicas em áreas de acumulação e muito planas (menor que 5% de declividade)".

O confronto entre as unidades geomorfológicas identificadas e a distribuição das orquidáceas, realizado por meio de uma tabela compilando a ocorrência de 75 das 77 espécies encontradas nas diversas SGHs, bem como pelo mapeamento detalhado dessas ocorrências, levou o autor a concluir que "a cada tipo de superfície geneticamente homogênea, encontramos uma fisionomia vegetacional diferente, com parte de seu componente florístico também variando". Um indicador do grau de precisão da classificação nesse contexto seria identificar em quantas SGHs diferentes cada espécie de orquídea é encontrada. Se uma proporção mais elevada de espécies for típica de um pequeno número de SGHs, configurando um certo tipo de endemismo no local, pode-se considerar que os ambientes estão bem caracterizados ecologicamente falando. No estudo verificou-se que um total de 45 espécies (algo em torno de 60% do total) são localizadas em apenas uma SGH, não necessariamente a mesma, obviamente, e cerca de 90% das espécies apresentaram preferência por poucas SGHs (no máximo 3), o que leva o autor da pesquisa a afirmar que "(...) esse dado já demonstra a grande capacidade da família como parâmetro para caracterizar diferentes fisionomias ou indicar variação ambiental" e também quanto ao "(...) grande potencial da análise geomorfológica em caracterizar unidades ambientais, tendo a vegetação abalizando esses limites".

3.2 A ferramenta See5

O indutor de árvores de decisão See 5 é a versão comercial dos resultados das pesquisas desenvolvidas pelo australiano John Ross Quinlan a partir da década de 1980 na área de aprendizado de máquina. Tais trabalhos propuseram em 1986 um primeiro algoritmo chamado ID3 que se tornou muito conhecido pelo emprego de conceitos derivados da Teoria da Informação no processo de indução, tendo sido extensivamente estudado pela comunidade acadêmica. O lançamento em 1993 do livro *C4.5: Programs for Machine Learning*, de autoria do mesmo pesquisador, apresentou uma versão bastante aperfeiçoada desse algoritmo original, estendendo suas funcionalidades para permitir o tratamento de conjuntos de treinamento com valores de atributo desconhecidos; valores de atributos contínuos; implementação de estratégias de poda de árvore e recursos para a extração de regras SE / ENTÃO a partir da árvore inicialmente induzida. Na seqüência dessa família de indutores surgiu o C5.0, implementado em versões para Unix e em uma implementação correspondente para ambiente MS-Windows®, chamado See5, produtos esses comercializados pela empresa RuleQuest Research® a partir de 1997. O objetivo desta seção é apresentar as características principais do produto, para o que será considerada a versão See5 Release 1.12, que inclui também uma implementação para uso via linha de comando (no *prompt* do MS-DOS), chamada See5X.

A proposta do programa é atingir uma faixa de usuários relativamente ampla, não requerendo conhecimentos avançados sobre Estatística e Aprendizado de Máquina, mas com bom domínio sobre a área de concentração do problema a que os dados se referem. Por esse motivo a interface do programa é bastante simples e, em linhas gerais, a operação da

ferramenta a partir do momento em que se dispõe dos dados já formatados pode ser dividida em duas etapas principais:

1. Geração do classificador, onde se procura ajustar as características das regras de classificação a serem induzidas por meio de diversos parâmetros que afetam tanto sua acurácia como a forma de representação das regras e da interpretação dos dados. É um trabalho de natureza incremental e analítica, na medida em que tipicamente vários ajustes vão sendo processados pelo usuário com base nos erros observados sobre os dados de treinamento e de teste até que se obtenha um classificador considerado adequado para os propósitos desejados.
2. Utilização (interativa) do classificador sobre novos dados. Nesta etapa os dados de casos reais ainda não classificados são digitados, um a um, e a classe prevista é fornecida pela ferramenta na tela, juntamente com um número que indica o grau de certeza daquela predição. Se mais de uma classe é aplicável para um determinado caso, todos os resultados possíveis são listados, seguidos de seus respectivos graus de certeza.

Considerando essa proposta da ferramenta e sua utilização neste trabalho, esta seção está organizada como se segue. Inicialmente as principais características do mecanismo de indução da ferramenta são descritas, acompanhadas por uma discussão sobre seus recursos mais importantes. Posteriormente uma visão geral da estrutura dos arquivos previstos e utilizados pelo See5 é fornecida, procurando compor um quadro sobre sua forma de operação. O texto é então finalizado com uma análise crítica das características do produto.

3.2.1 Recursos para a geração de classificadores

Como um descendente direto do algoritmo ID3, o See5 induz árvores de decisão aplicando conceitos e medidas originárias da Teoria da Informação para a seleção dos testes a serem colocados em cada ramo interno da árvore. Em princípio, o atributo cuja utilização minimiza a entropia do sistema, ou seja, produz as subdivisões do conjunto de treinamento com maior grau de homogeneidade internas é escolhido para compor o próximo nó intermediário da árvore. Para os atributos que possuem um domínio discreto (como por exemplo: sexo, estado, cor, etc) são considerados todos os possíveis valores observados no conjunto de treinamento, o que significa em princípio gerar um subconjunto para cada valor distinto encontrado e avaliar seu grau de entropia com relação à classe. Os atributos cujos conteúdos são obtidos a partir de faixas de valores contínuos (por exemplo: tamanho, peso, temperatura, etc) requerem que seja calculado, com base nos dados do conjunto de treinamento, um valor (geralmente chamado de valor de subdivisão ou *split value*) que maximiza o potencial de classificação daquele atributo, fazendo com que se formem dois grupos, um com os elementos do conjunto de treinamento cujo valor do atributo contínuo é maior que o *split value* e outro para os restantes.

O emprego dessa abordagem baseada no ganho de informação, apesar de conceitualmente adequado, tende a favorecer a escolha dos atributos que possuem maior número de valores distintos no conjunto de treinamento (QUINLAN, 1996). A solução adotada inicialmente para os algoritmos ID3 e C4.5 foi utilizar uma nova medida, chamada Taxa de Ganho (*Gain Ratio*) calculada pela fórmula

$$\text{Taxa de Ganho}(x, S) = \frac{\text{Ganho}(x, S)}{\textit{SplitInfo}(x, S)} \tag{9}$$

que procura ponderar o ganho de informação devido a um atributo x por considerar também os tamanhos dos subconjuntos por ele originados, determinado por *SplitInfo*, cujo cálculo é dado por

$$\textit{SplitInfo}(x, S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Log}_2 \left(\frac{|S_i|}{|S|} \right) \tag{10}$$

Entretanto, observou-se que mesmo o uso de Taxas de Ganho ainda não bastava para eliminar um certo favorecimento concedido aos atributos contínuos durante o processo de escolha do atributo de *split*, na medida em que o cálculo de um valor de corte (um limiar ou *threshold*) para um atributo contínuo é feito de maneira a maximizar seu efeito, em detrimento de atributos discretos para os quais esse benefício não pode ser estendido. Aperfeiçoamentos posteriores no algoritmo passaram a dar um tratamento diferenciado aos atributos contínuos, procurando evitar esse favorecimento por meio de três modificações (QUINLAN, 1996):

a) utilização do Ganho de Informação para a escolha do valor de corte, em vez da Taxa de Ganho;

b) descarte dos atributos contínuos para os quais nenhum valor de corte produz um Ganho de Informação suficiente para compensar uma penalidade, dada por $\text{Log}_2(N - 1) / |D|$ onde N é o número de valores distintos observados, imposta a essa categoria de atributos;

c) reclassificação dos candidatos a atributo de *split* penalizando aqueles que envolvem faixas contínuas de valores.

Tais aperfeiçoamentos, disponíveis a partir do Release 8 do C4.5, permitiram a indução de árvores de decisão mais compactas e com índices de acurácia ligeiramente mais elevados que as versões originais (QUINLAN, 1996).

Além de correções como essa, que afetam aspectos centrais do processo de indução, uma série de outros recursos foram incorporados ao longo do tempo e são gerenciados na versão atual do See5 através da caixa de diálogo "*Classifier Construction Options*" disponível nas versões para uso em ambiente gráfico e interativo ou através de parâmetros digitados em linha de comando (versão See5X do programa). A figura 8 apresentada a seguir mostra a aparência da caixa de diálogo padrão para a definição desses parâmetros, dos quais aqueles tidos como mais importantes serão posteriormente descritos em suas linhas gerais. O produto oferece ainda um tutorial (RULEQUEST RESEARCH) sobre o processo de indução e os principais recursos que podem ser utilizados, escrito numa linguagem clara e com farto uso de exemplos, em um enfoque voltado para um público leigo na área de Aprendizado de Máquina.

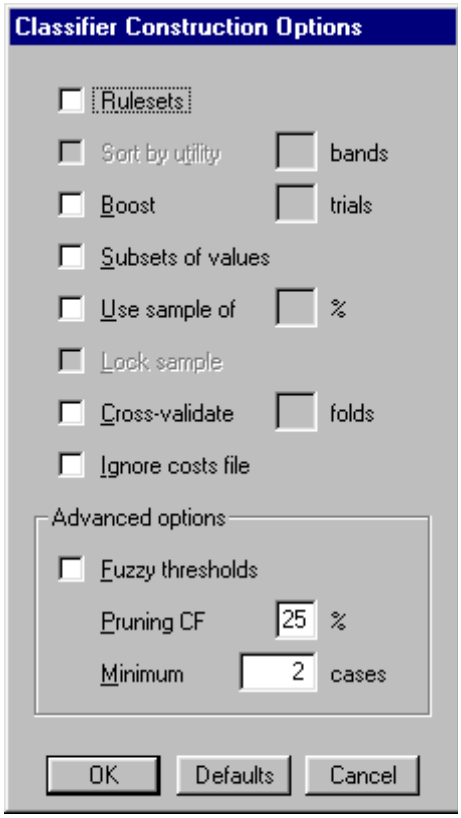


Figura 8: Opções para a geração de classificadores no See5.

Arquivos de custos de erro

Um recurso de grande utilidade quando se monta classificadores é a possibilidade de se estabelecer custos relativos para os erros de classificação observados no processo de indução. Tais erros ocorrem conforme os parâmetros que foram utilizados para a indução e os dados disponíveis para treinamento e teste, sendo que a medição da acurácia é realizada pela contagem simples de cada erro verificado, independentemente do tipo de erro ocorrido. Assim, se duas classes "A" e "B" são possíveis, classificar erroneamente um caso "B" como sendo "A" tem o mesmo peso que classificar erroneamente um caso "A" como "B". Por diversos motivos é possível, entretanto, que desejemos minimizar a possibilidade da ocorrência de um dos tipos, e isso pode ser obtido especificando-se um peso maior para esse tipo de erro, através de um arquivo de custos. Um arquivo de custos é um arquivo texto não formatado, cujo nome é igual ao do conjunto de testes com a extensão .costs, e que contém uma linha para cada custo de erro que se queira estipular, no seguinte formato:

```
classe prevista, classe verdadeira: custo
```

Portanto, se queremos dizer ao mecanismo indutor que classificar um caso da classe "B" como "A" tem o dobro do custo de se classificar um caso da classe "A" como "B", deve-se colocar nesse arquivo a seguinte especificação:

```
A, B: 2
```

Regras de produção

Selecionando-se a opção *Rulesets*, é possível obter um conjunto de regras SE-ENTÃO a partir de um *training set*, com o resultado final sendo apresentado em um formato similar ao do exemplo contido no quadro a seguir.

```
Rule 3: (3/1, lift 1.2)
          Umidade = MÉDIA
          Altitude > 30
          ->   class B   [0.600]
```

Conforme pode ser observado nesse exemplo, uma regra possui um número de identificação arbitrário; uma indicação sobre o número de casos do *training set* que serviram de base para sua definição; um conjunto de predicados que correspondem aos testes do lado esquerdo da regra, sendo que em uma regra cada predicado é ligado ao seguinte por meio do conectivo lógico "E" (ou seja, correspondente ao operador AND ou ^, conforme a notação adotada); um resultado para a classe, colocado após o sinal de implicação "−>", que corresponde ao lado direito da regra e, finalmente, um valor no intervalo [0,1] que indica o seu fator de confiança.

O fator de confiança de uma regra é calculado através da fórmula

$$Confiança(x) = \frac{N * (|S_i| + 1)}{N * (|S_x| + 2)} \qquad (11)$$

onde N é o número total de casos (ou seja, o tamanho do conjunto de treinamento ou teste); S_i é o conjunto de casos classificados corretamente pela regra x; e S_x é o conjunto de casos correspondentes à regra x (QUINLAN, 1993). Para a regra 3 fornecida como exemplo, produzida com base nos dados constantes na tabela 1 apresentada no capítulo anterior, temos um total de 10 casos a classificar (o valor de N); 2 casos classificados corretamente pela regra (o valor de $|S_i|$) e um total de 3 casos que foram por ela classificados como pertencentes à classe "B" (o valor de $|S_x|$), o que produz o fator de confiança de 0.600 indicado no quadro.

As versões mais recentes do produto passaram também a apresentar uma nova informação, chamada *lift*, que procura ponderar a confiança da regra em função do número de casos observados para a classe em questão e que representa uma espécie de medidor para o suporte da regra. Seu cálculo é feito dividindo-se a confiança da regra pela probabilidade anterior da classe, ou seja, para o exemplo dado na regra 3 anteriormente mostrada, divide-se 0.600 (a confiança) por 0.5 (a probabilidade anterior, já que haviam 5 casos com classe = "A" num universo de 10 casos), o que vai dar o *lift* de 1.2 apresentado anteriormente. Portanto, uma boa regra deve apresentar um alto fator de confiança (quanto mais próximo de 1, melhor) e um valor de *lift* relativamente baixo (quanto menor, melhor), sendo que o conceito de "alto" e "baixo" aqui é proporcional à probabilidade anterior da classe (boas regras para uma classe majoritária no conjunto de treinamento devem ter *lift* menor que boas regras para uma classe minoritária).

Com isso, podemos interpretar a descrição sobre regra apresentada no exemplo da seguinte forma:

Nome da regra:	Rule 3 (ou apenas 3)
Nº de casos em que se baseia:	3
Nº de casos classificados incorretamente pela regra:	1
Regra:	SE Umidade = 'MÉDIA' E Altitude > 30 ENTÃO Classe = 'B'
Fator de confiança:	60%
Relação entre o nº de casos da classe x confiança da regra:	1.2

Por motivos que incluem a utilização ou não de recursos de poda posterior e a adoção de critérios de parada diferenciados, entre outros, é possível que ao final do processo de indução tenhamos casos que não se enquadrem em nenhuma das regras produzidas pela ferramenta. A tais casos é atribuído um valor de classe padrão (*default class*), definido com base na classe mais freqüente no conjunto de treinamento.

Poda posterior

O descarte de subárvores com alta taxa de erro previsto é feita com base no parâmetro indicado em *Pruning CF*, que sugere inicialmente um valor padrão de 25%. Quanto menor o valor definido para esse parâmetro, maior a severidade da poda, que afetará tanto as árvores construídas como o respectivo conjunto de regras (isso caso a opção *Ruleset* tenha sido ativada). Casos extremos, em que esse fator de confiança da poda seja definido com valores muito baixos, fazem com que um classificador majoritário venha a ser gerado, ou seja, a classe a ser prevista será sempre aquela mais freqüente no conjunto de treinamento.

Boosting

A opção *Boosting* habilita a construção de um conjunto de diversos classificadores para o mesmo conjunto de treinamento, que serão posteriormente utilizados no processo de predição dos novos casos. A quantidade de classificadores a serem gerados é fornecida pelo usuário, sendo que a ferramenta sugere inicialmente o número 10. No arquivo de saída contendo o resultado são exibidos todos os classificadores que foram gerados, na forma de árvores de decisão, juntamente com a taxa de erro de cada um e também do conjunto de classificadores como um todo. Ao se processar um novo caso, todos os classificadores são consultados e cada um deles fornece a sua própria predição da classe correspondente, que é então computada em uma espécie de *ranking* das classes possíveis para o caso. Ao final, a classe melhor posicionada nesse *ranking* (a que teve um maior número de "votos") é indicada como o resultado final.

Agrupamento de valores

Um recurso bastante interessante que a ferramenta oferece é a possibilidade de se agrupar automaticamente os valores de atributos discretos para a elaboração dos testes, o que tenderia à geração de árvores de decisão mais compactas. Essa redução é obtida porque um atributo declarado como discreto que possui, por exemplo, 4 valores distintos poderá talvez levar à divisão dos dados em apenas dois subconjuntos em vez de quatro, simplificando assim a estrutura geral da árvore pela diminuição de seu número de folhas. Esse recurso está disponível desde o C4.5 e é habilitado no See5 por meio da opção *Subsets of values* sendo que um exemplo de sua utilização aparece mais à frente neste capítulo, com a árvore de decisão sendo exibida na figura 11.

Número mínimo de casos

O parâmetro *Minimum cases* expressa uma medida da liberdade que o mecanismo indutor terá para agrupar os casos do conjunto de treinamento ao construir o classificador. Se definirmos um valor para *Minimum cases* igual a 5, por exemplo, estamos ordenando que a árvore a ser construída deverá apresentar no mínimo dois ramos contendo 5 casos. A utilização deste recurso tende a produzir árvores de estrutura mais simples, mas com maior taxas de erro sobre o conjunto de treinamento, e representa uma forma de poda prévia. O valor padrão sugerido pela ferramenta para esse parâmetro é 2 e convém modificar esse recurso com cautela, uma vez que a presença de atributos com valores desconhecidos no conjunto de treinamento e a especificação de custos diferenciados para cada tipo de erro possível acabam afetando o resultado final obtido.

Suavização de valores de corte

Cada atributo contínuo é tratado pelo algoritmo indutor calculando-se um valor de corte (*split value*) que será utilizado para subdividir os dados do conjunto de treinamento em dois grupos durante o processo de escolha do atributo mais adequado para a composição de um determinado nó da árvore. Esse valor é, em princípio, definitivo: ocorrências menores ou iguais a ele são colocadas em um grupo, as restantes em outro grupo. Muitas vezes esse tratamento pode ser inadequado, na medida que casos com valores muito próximos entre si para o atributo em questão podem eventualmente ser colocados em grupos diferentes. Para minimizar essa possibilidade o See5 oferece um recurso para o abrandamento desse critério em alguns casos, que é habilitado através de sua opção *Fuzzy thresholds*, que afeta apenas as árvores de decisão, e não os conjuntos de regras produzidos por meio da opção *Rulesets*.

A estratégia de abrandamento consiste em definir para cada valor de corte um limite superior e um limite inferior que fazem com que, durante a classificação, os casos possam receber 3 tipos de tratamento com base naquele atributo:

a) os que possuem valores igual ou abaixo ao limite inferior são alocados ao ramo “esquerdo” da árvore;

b) os que possuem valores igual ou acima ao limite superior são alocados ao ramo “direito” da árvore;

c) os que possuem valor no intervalo entre os limites calculados levam o classificador a fazer uma combinação probabilística dos resultados decorrentes da alocação daquele caso a cada um dos dois ramos possíveis.

Os limites superior e inferior do atributo de corte são calculados automaticamente pelo algoritmo indutor de regras com base em critérios internos próprios que procuram considerar o seu grau de sensibilidade quanto ao resultado final da predição.

3.2.2 Estrutura dos arquivos utilizados

Durante o processo de construção de um classificador, o See5 utiliza uma série de arquivos, tanto para a leitura de parâmetros e dados dos conjuntos de treinamento e de teste como também para a gravação dos resultados do processamento da aplicação. Uma visão geral desses arquivos é mostrada na figura 9, onde os arquivos obrigatórios são representados com linha contínua, os arquivos opcionais com linha tracejada e o nome da aplicação é "xxx".

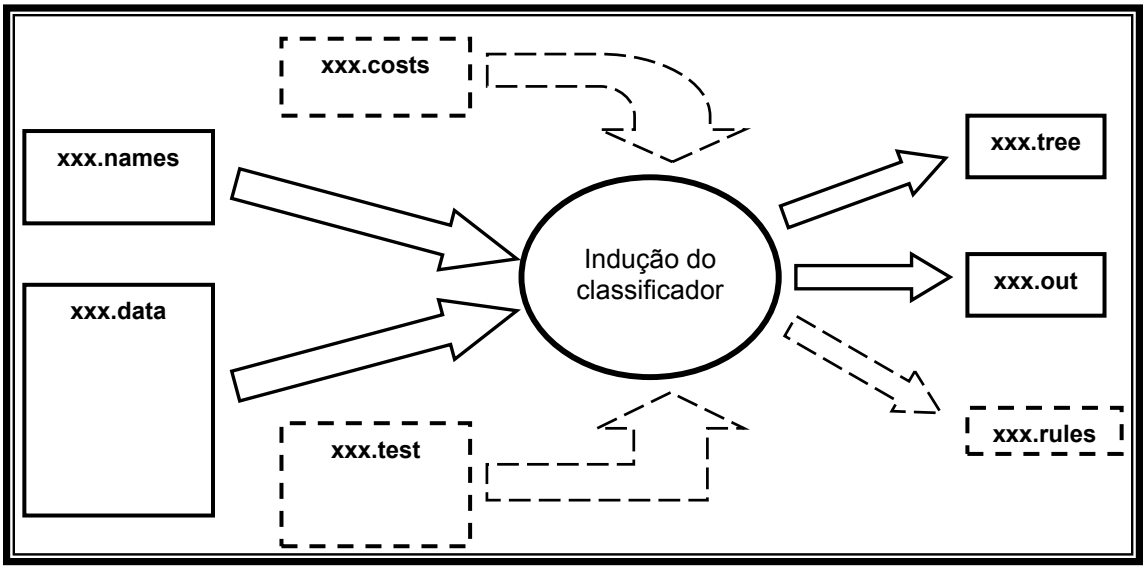


Figura 9: Arquivos utilizados pelo See5 para a geração de classificadores.

O quadro a seguir indica o formato, tipo de conteúdo e finalidade de cada um desses arquivos.

Extensão	***Descrição***
.names	Arquivo de nomes, do tipo texto não formatado, contém os nomes dos atributos, os valores válidos para cada um (ou seja, seus domínios) e indica também qual deles é o atributo categórico (de classe).
.data	É um arquivo texto não formatado que representa o conjunto de treinamento. Cada caso é um parágrafo desse arquivo, com seus atributos separados por vírgulas.
.tree	É o arquivo de saída que contém a árvore de decisão em formato binário e que permite sua utilização por outros

	programas desenvolvidos para fins específicos pelo usuário.
.out	Este arquivo texto contém o resultado final exibido na tela quando o classificador é gerado, na forma de um relatório.
.test	De estrutura similar à do arquivo .data, contém um conjunto de casos utilizado para a avaliação da acurácia do classificador que está sendo induzido.
.costs	Também do tipo texto não formatado, é um arquivo opcional que indica os custos associados com cada tipo de erro que se quer minimizar.
.rules	É o arquivo de saída que contém o conjunto de regras de decisão em formato binário e que permite sua utilização por outros programas desenvolvidos para fins específicos pelo usuário.

Informações detalhadas sobre as convenções internas para a construção de cada um desses arquivos podem ser obtidas a partir do tutorial *on-line* disponível com o produto e também analisando-se os arquivos de exemplo fornecidos com a ferramenta.

3.2.3 Leitura dos resultados

Uma vez gerado o classificador, é exibido pelo See5 um relatório que apresenta um resumo dos parâmetros utilizados e das opções habilitadas para o processo, seguido de uma descrição da árvore de classificação, do conjunto de regras gerado (caso a opção *Rulesets* tenha sido selecionada) e também uma pequena compilação dos erros observados sobre o conjunto de treinamento e, se existir, o conjunto de teste. O quadro a seguir mostra um exemplo desse tipo de relatório, que fica gravado em disco em um arquivo com extensão .out.

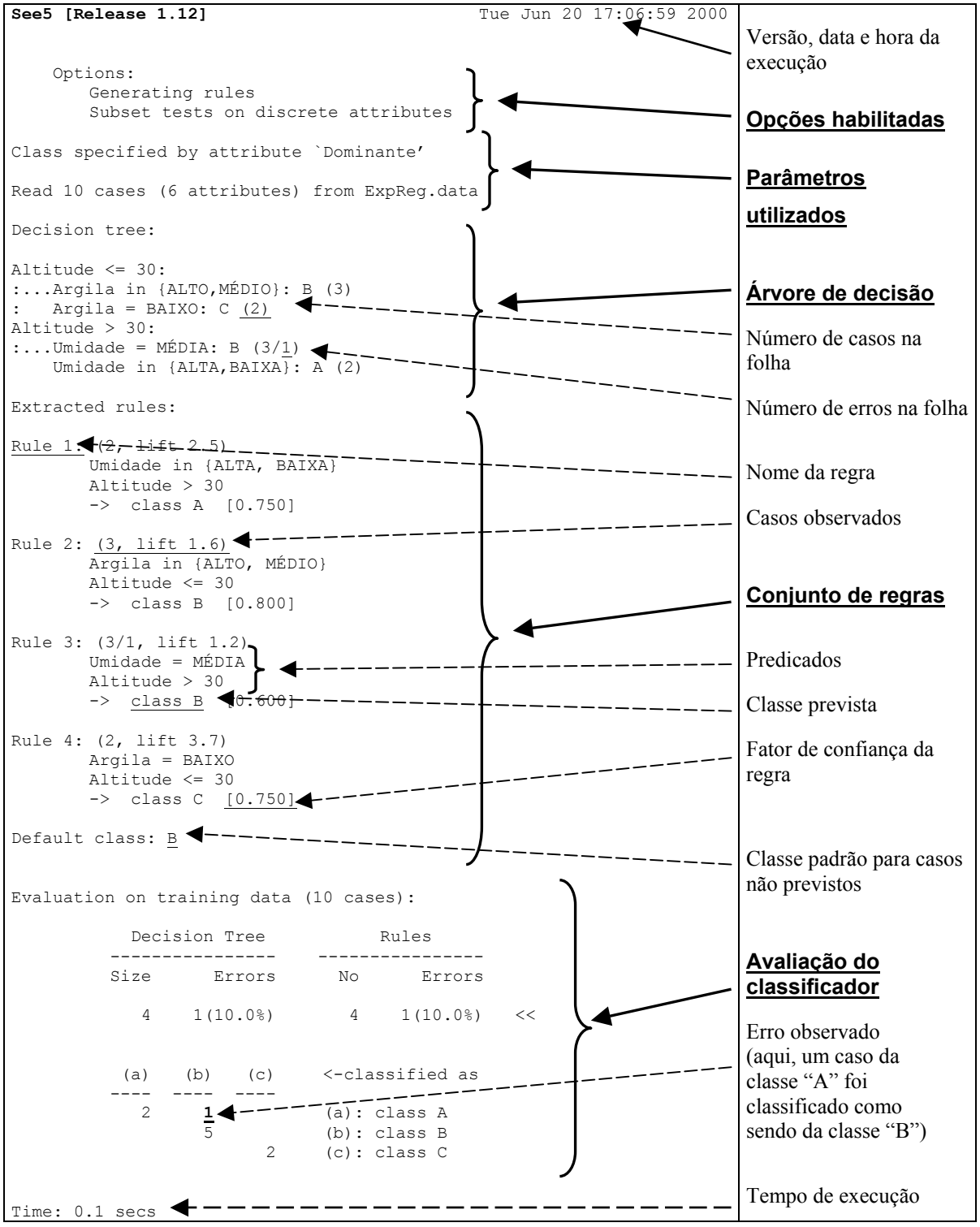


Figura 10: Exemplo de relatório de saída do See5.

Um relatório como esse indica uma árvore de decisão que pode ser melhor visualizada conforme mostrado na figura 11.

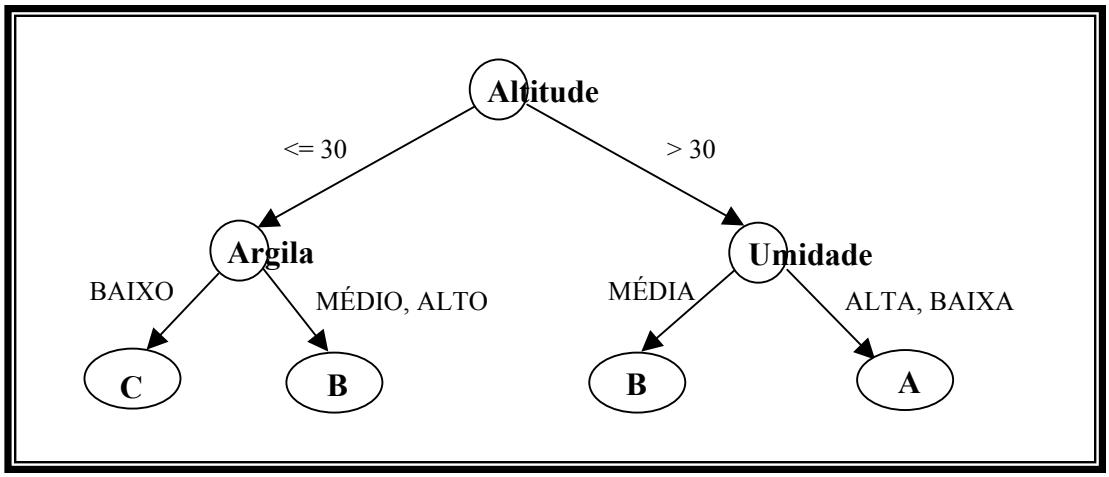


Figura 11: Representação gráfica da árvore indicada no relatório de saída do See5 da Figura 10.

Nesse exemplo, para cada folha da árvore corresponde uma regra SE / ENTÃO, que é lhe exatamente equivalente logicamente. Entretanto, muitas vezes uma regra pode ser mais compacta (com um número menor de comparações) do que o respectivo ramo na árvore, uma vez que nela são mantidas apenas as condições imprescindíveis para a determinação da classe naquele contexto.

3.2.4 Análise da ferramenta

De uma maneira geral, o produto se mostra adequado às finalidades a que se propõe, pois oferece um conjunto de recursos bastante interessante acionado através de uma interface com o usuário simples e prática. A geração de regras de decisão; a capacidade de agrupamento de valores discretos e a possibilidade de uso de arquivos de custo de erros, entre outros, tornam muito flexível o projeto dos classificadores a serem induzidos, mantendo ainda um aparente grau de simplicidade para o usuário leigo em Aprendizado de Máquina. Entretanto, alguns outros recursos bastante interessantes como a suavização de *split values* (uma espécie de "*fuzzyficação"* dos valores de corte dos atributos contínuos) e a implementação de classificadores baseados em votação (*boosting*), por exemplo, parecem oferecer um nível de dificuldade maior para a compreensão do usuário final, não obstante sua inegável utilidade prática.

Com relação à interface do aplicativo, cuja versão analisada é voltada para ambiente gráfico padrão MS-Windows, cabe observar que a estrutura do relatório de saída descrevendo o classificador gerado é suficiente para um bom entendimento de sua estrutura, apesar de a forma escolhida para a representação da árvore de decisão carecer de uma maior clareza, que talvez pudesse ser obtida por meio de um tratamento gráfico mais elaborado.

Um outro ponto fraco na versão analisada do produto é a necessidade de se usar o classificador de maneira interativa, com a digitação dos novos dados caso a caso. Essa abordagem é adequada apenas para uma classe de uso, bastante típica por sinal, onde o volume de casos novos a serem classificados é pequeno e o uso da ferramenta é eminentemente analítico, como no exemplo de um médico estudando o estado de um novo

paciente que lhe foi apresentado. Situações diversas, onde um grande volume de novos dados devem ser classificados a cada vez, requerendo a geração automática de resultados para algum tipo de arquivo em disco não são contempladas pelo produto em sua versão de prateleira e requerem a construção de aplicações específicas que façam uso dos classificadores contidos nos arquivos binários (de extensões .tree e .rules) gerados previamente pela ferramenta. A interface para o tratamento desses arquivos é disponibilizada para *download* gratuito na internet mas a presença de um módulo adicional na ferramenta que realizasse a classificação de novos casos em lote, lendo-os do disco e realizando a saída para algum tipo de arquivo de formato mais geral como texto, planilha ou dbf, por exemplo, seria também interessante.

Neste trabalho, onde pretende-se que as regras de classificação induzidas automaticamente sejam abordadas sob uma perspectiva descritiva, ou seja, como um elemento de auxílio ao especialista para o entendimento dos critérios e relações entre os diversos casos de uma classe, os aspectos ligados à facilidade de entendimento da estrutura do classificador gerado ganham muita importância. Isso faz com que alguns recursos oferecidos pelo See5 tenham um alto grau de interesse, ao mesmo tempo em que outros acabem por ser descartados dos experimentos.

Dentre os recursos de maior importância estão a capacidade de geração de regras (opção *Rulesets*), que muitas vezes são mais legíveis e compactas do que os ramos da árvore correspondente e a possibilidade de agrupamento dos valores discretos (opção *Subsets of values*), que representa um elemento de generalização de conceitos e simplificação do classificador. Dado o problema do pequeno volume de dados disponível para os experimentos, que será discutido na próxima seção deste capítulo, serão utilizados apenas os valores padrão para os parâmetros de ajuste das podas posterior (opção *Pruning CF*) e prévia (opção *Minimum cases*).

Não serão utilizados os recursos de *boosting* devido às suas características estruturais: seu objetivo é melhorar a capacidade preditiva, mas não necessariamente a descritiva, já que o significado e o papel no resultado final dos diversos classificadores gerados tende a ser de muito difícil análise e entendimento, ficando o conjunto classificador assim obtido muito próximo de uma "caixa preta". Finalmente, os recursos relativos à amostragem (opções *Use Sample* e *Lock Sample*) e validação cruzada (opção *Crossvalidate ... folds*) não terão aplicação prática devido ao tamanho muito reduzido do conjunto de amostras disponível para os experimentos.

3.3 Aspectos quantitativos

A essência do trabalho de um cientista ou especialista reside em compreender algum aspecto da realidade e propor aplicações úteis desse conhecimento. Se admitimos a generalidade dessa colocação podemos então considerar que, em última instância, a intenção do estudo científico é modelar o aspecto da realidade no qual temos interesse e/ou aplicar o modelo produzido visando atingir algum propósito específico. Dessa forma, a modelagem assume um papel de importância central no processo de investigação científica e mesmo não-científica, na medida em que compreender também é construir modelos próprios, sendo realizada através de uma grande variedade de técnicas e com um grande

número de ferramentas, selecionadas em função das características do problema que se procura resolver. Métodos quantitativos, apoiados na Matemática e Estatística, possuem largo emprego na construção desses modelos em virtualmente qualquer área de pesquisa, tanto por resultarem de um corpo de conhecimento já bastante consolidado ao longo do tempo como também pela facilidade de implementação dos modelos resultantes através de computadores, o que facilita muito a sua aplicação prática. Mais recentemente, desenvolvimentos em áreas como a computação e as ciências da informação têm oferecido importantes contribuições para a construção e utilização de modelos, muitas vezes tomando emprestados conceitos matemáticos e estatísticos mais tradicionais e fazendo uma interpretação mais original dos mesmos ou combinando-os com novos conceitos e tecnologias de suas respectivas áreas. Algumas vezes essas novas abordagens possibilitam soluções muito boas que seriam impensáveis da perspectiva matemática e estatística mais convencional, o que tem gerado um estado de coisas em que torna-se freqüentemente confuso delimitar o tipo de contribuição devida a cada área na composição de um determinado método ou abordagem. Assim, são comuns questionamentos do tipo "Qual a diferença entre *Data Mining* e Estatística?" ou "Mineração de dados e Descoberta de Conhecimento em Bancos de dados são a mesma coisa?" ou ainda afirmações como "Aprendizado de Máquina é apenas Estatística", por exemplo, muitas vezes descambando a discussão para abordagens perigosamente subjetivas. A questão dos limites entre as diversas áreas de conhecimento é reconhecidamente complexa e ao ser muitas vezes super-dimensionada, propicia debates estéreis e classificações de pouca utilidade prática.

Esta seção tem por objetivo analisar os aspectos quantitativos relacionados à viabilidade da aplicação da ferramenta See5 sobre a base de dados originária do estudo "Análise Geomorfológica e Distribuição Espacial da Vegetação na Planície Litorânea de Picinguaba (Ubatuba - SP)" (GARCIA, 1995), e terá de encarar em alguns momentos, questões referentes à delimitação das diversas abordagens possíveis para o problema. Inicialmente será feita uma breve apresentação sobre alguns elementos de amostragem; depois uma discussão sobre as principais diferenças geralmente aceitas entre o enfoque estatístico dito mais convencional e aquele preconizado por métodos de mineração de dados e, finalmente, uma caracterização do problema específico da aplicação da ferramenta de mineração de dados See5 sobre a base de dados do projeto.

3.3.1 Amostragem

Estudos científicos dos fenômenos observados na natureza são baseados na coleta de informações sobre os diversos aspectos relevantes para a compreensão do problema e na aplicação de métodos e técnicas diversas para a descrição, organização e análise do material coletado. Tipicamente o fenômeno, que pode ser por exemplo um fato ou objeto, é descrito em termos de suas características tidas como mais relevantes pelo pesquisador (seus atributos ou variáveis), expressas por algum tipo de medida julgada conveniente para o caso, permitindo a construção de modelos sobre os quais as diversas hipóteses aventadas no trabalho são então verificadas.

Uma técnica de particular importância, com grande desenvolvimento na área da Estatística, é a da amostragem, onde se procura retratar o todo de uma classe de fenômenos reais (a população) a partir de apenas algumas de suas manifestações ou instâncias (as amostras). Freqüentemente a amostragem é imprescindível em um estudo, por vários motivos:

a) o processo de estudo pode destruir o próprio material estudado durante os experimentos e isso não pode ser permitido, a exemplo do que ocorre em aplicações de controle de qualidade na indústria;

b) é impossível medir e descrever todas as manifestações do fenômeno, por um motivo qualquer, como ocorre com o solo de uma região;

c) o volume de informação resultante da descrição populacional é grande demais para ser compilado e analisado, como no caso de alguns bancos de dados comerciais com volumes de dados que excedem a capacidade de processamento de seus sistemas computacionais;

d) as técnicas empregadas requerem a definição de dois ou mais conjuntos de dados distintos mas hipoteticamente equivalentes, como ocorre, por exemplo, ao se utilizar determinadas ferramentas indutoras de regras de decisão, que necessitam de um conjunto de dados para induzir as regras, outro para podá-las e outro ainda para avaliar a acurácia do classificador, etc.

A questão chave quando se utiliza a amostragem é garantir que as manifestações (amostras) selecionadas da população sejam representativas da mesma, ou seja, permitam formar a mesma concepção que seria obtida se tratássemos diretamente da população. A abordagem estatística que é empregada para essa questão está intimamente ligada ao fato de que tratamos com modelos, onde os fenômenos são representados (modelados) por meio de simplificações que procuram reter apenas sua essência, segundo a perspectiva do pesquisador. Assim, se selecionamos como realmente importante apenas um atributo/variável do fenômeno, e sabendo que há uma variabilidade natural dessa informação nas diversas manifestações possíveis do fenômeno estudado no mundo real, basta que selecionemos para estudo algumas amostras que representem proporcionalmente essa variabilidade. Mais tecnicamente falando, basta que a amostra preserve as linhas gerais da distribuição de probabilidades dos valores possíveis no mundo real para aquela variável selecionada ou, em outras palavras ainda, que a amostra possua, aproximadamente, a mesma distribuição de probabilidades da população em que se originou.

Ao se conduzir um processo de amostragem, uma primeira questão que se coloca está ligada então à forma como as amostras deverão ser selecionadas para garantir sua representatividade, na medida em que sua distribuição pode ser dependente de uma série de fatores tais como espaciais, temporais, etc e conforme os critérios adotados no processo de seleção, algumas distorções (chamadas *bias* ou viés de seleção) podem ser geradas, comprometendo assim os resultados produzidos posteriormente com base no conjunto de amostras. Tal problema geralmente é solucionado retirando-se as amostras aleatoriamente da população original, técnica essa que faz com que, após um certo número de seleções, as distribuições amostral e populacional tornem-se naturalmente bastante próximas. Esse comportamento, chamado de convergência, significa que as variabilidades amostral e populacional tendem a tornar-se semelhantes a medida que o tamanho da amostra cresce (PYLE, 1999), sendo que as variabilidades geralmente são expressas por medidas como a variância e o desvio padrão. Naturalmente, o processo de convergência também ocorre em amostragens não aleatórias, mas com ritmos e características menos previsíveis. Em todo caso o grau de convergência sempre pode ser medido incrementalmente, conforme novas instâncias vão sendo selecionadas e adicionadas à amostra: após cada acréscimo calcula-se

a nova variabilidade e mede-se a sua diferença em relação à anterior. Essa diferença tende a diminuir progressivamente, e quanto menor ela for, mais próxima deve ser a variabilidade amostral da variabilidade populacional, o que pode ser interpretado como um bom indício de representatividade da amostra.

É importante considerar que nos trabalhos científicos em geral, e em aplicações destinadas à descoberta de conhecimento em particular, cada objeto é descrito por um conjunto de seus atributos considerados mais importantes para o problema em questão. Isso significa que a amostra é formada por um certo número de objetos cada um deles representado por uma determinada quantidade de variáveis, ou seja, o espaço amostral é multidimensional (ou multivariado). Com isso, a complexidade da amostra torna-se muito grande, porque além das considerações necessárias à compreensão e correta representação de cada variável, surge a questão das ligações entre as diversas variáveis e todas as possibilidades decorrentes desses relacionamentos. Cada nova variável que é acrescentada à estrutura da amostra gera um aumento exponencial no tamanho do espaço amostral (ou espaço de estados) a ser considerado, o que significa que o tamanho da amostra é proporcional também ao número de variáveis utilizadas. Essa característica apresenta um forte impacto no que diz respeito à representatividade da amostragem, já que torna-se necessário ajustar não apenas a variabilidade individual de cada variável como também as suas variabilidades conjuntas, já que é perfeitamente possível que uma amostra em que todas as variáveis possuem distribuição própria condizente com a população não o seja quando se comparam as respectivas distribuições conjuntas.

Um segundo aspecto importante, este decorrente do processo de convergência, refere-se à identificação de qual o número necessário de amostras para representar satisfatoriamente as características de distribuição da população original que, aliás, é desconhecida na maioria dos casos. Qual o volume de amostras necessárias para garantir que ela possa ser considerada representativa da população que a originou? Essa questão é crítica para praticamente qualquer trabalho científico que utilize algum tipo de amostragem e, felizmente, encontra-se bem amadurecida no âmbito da ciência estatística através da definição do conceito de confiança (ou nível de confiança). Adotar um nível de confiança geralmente bastante razoável de 95% significa garantir com 95% de certeza que a variabilidade de uma variável foi capturada na amostragem (PYLE, 1999). Como regra geral, quanto maior o volume de amostras aleatoriamente coletadas, maior a variabilidade capturada, sendo que um nível de 100% de confiança obviamente só será possível com uma amostra igual à população o que é, neste contexto, um contra-senso. Então o número de amostras necessário é uma função do nível de confiança que se pretende adotar no trabalho, sendo determinado na prática através de várias maneiras diferentes, uma vez que não existe uma fórmula universal para a determinação do número suficiente de amostras para todos os possíveis tipos de estudos. Algumas abordagens para esse cálculo são mostradas a seguir, a título de ilustração.

Tamanho da amostra em função da variabilidade capturada

Determinar o que é uma amostragem suficiente com base na confiança de que a variabilidade amostral está suficientemente próxima da variabilidade populacional é uma abordagem realista mas, definida apenas nesses termos, é ainda um objetivo bastante vago. A questão que se coloca a bem da clareza é definir, com base no nível de confiança que se

pretende adotar, um bom critério para o julgamento do significado prático do termo "suficiente", de maneira a determinar uma condição objetiva que permita considerar que o processo de coleta de amostras pode ser dado por encerrado. A idéia central é ir aumentando progressivamente o tamanho da amostragem e, a cada acréscimo, medir a variabilidade obtida e compará-la com a variabilidade anterior. Quando a diferença entre ambas atingir um patamar mínimo, definido com base no fator de confiança adotado, considera-se que houve efetivamente a convergência e para-se de coletar novas amostras. Eventualmente, com algumas poucas amostras pode ocorrer que essa diferença se mantenha pequena, sem que isso signifique, entretanto, que a amostragem seja suficiente. Portanto, é necessário que essa diferença se mantenha pequena por um número mínimo de comparações, valor esse que pode ser calculado com base no próprio nível de confiança adotado (PYLE, 1999).

PAC *Learning*

Existe uma abordagem no campo da inteligência artificial que parte da premissa de que, após termos avaliado um certo número de amostras representativas do conceito a ser aprendido, qualquer hipótese previamente formulada que seja seriamente errada acabará sendo percebida e descartada. Com isso ao final do processo de avaliação sobrarão somente as hipóteses teoricamente mais confiáveis ou, em inglês, *Probably Approximately Correct* (expressão que dá origem à sigla "PAC"). Essa abordagem é chamada de PAC-*Learning*, e com base em suas formulações, que são melhor descritas em (RUSSEL, 1995), podemos encontrar uma maneira de se estimar com maior precisão o tamanho mínimo da amostra requerida para um experimento dentro dessa perspectiva. Esse cálculo é feito computando o tamanho do conjunto de sentenças lógicas possíveis de serem formadas envolvendo os atributos utilizados, chamado espaço de hipóteses e que é dependente da linguagem de representação utilizada para compor as sentenças e também de eventuais restrições assumidas com relação à complexidade máxima admitida para uma sentença. Também são utilizados parâmetros para a acurácia mínima requerida e para o fator de confiança adotado.

O trabalho com uma linguagem de representação que permita sentenças de complexidade razoável e o número de atributos tipicamente considerados em aplicações de mineração de dados, que em geral não são menores que 10, às vezes chegando mesmo à casa da centena, levam a valores extremamente grandes para o tamanho da amostragem requerida, sendo possível provar que o número de exemplos necessários para que um algoritmo de indução consiga "aprender" dessa forma (*PAC-learning*) um conceito é polinomial quanto ao número de atributos considerados e exponencial quanto ao número de atributos efetivamente utilizados para compor a regra (RUSSEL, 1995).

Curva de aprendizado

Uma outra forma de se determinar o volume de amostras necessárias para um experimento, mais específica para o aprendizado de máquina, consiste em montar o conjunto de amostras incrementalmente, monitorando o grau de erro observado nas predições realizadas sobre os conjuntos de treinamento e teste (WEISS, 1999). À medida que o volume de amostras utilizado aumenta, a taxa de erro deve diminuir progressivamente, até se estabilizar num determinado nível, conforme mostra a figura 12 a seguir. Esse tipo de gráfico é chamado de curva de aprendizado e é muito útil para indicar o ponto em que novos acréscimos no

conjunto de amostras deixam de ter impacto significativo na qualidade do resultado final, sendo este o método de maior valor prático para aplicações de mineração de dados, já que de alguma forma considera também a complexidade dos conceitos e dados trabalhados e as particularidades do mecanismo de indução utilizado.

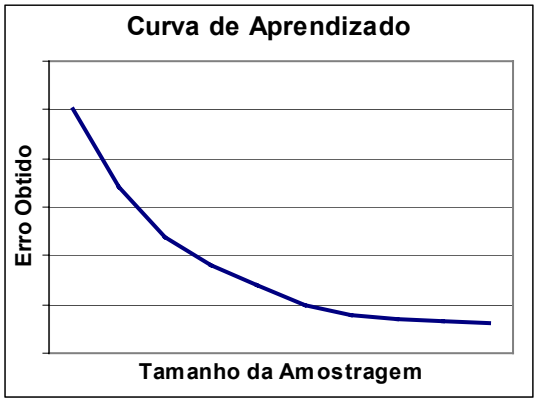


Figura 12: Comportamento geral do erro em função da amostra.

Finalmente, é importante ter sempre em mente que um outro fator de extrema importância é a qualidade dos dados utilizados no processo: independentemente da quantidade dos dados disponíveis, a qualidade do resultado final é também totalmente dependente do grau de adequação dos dados originais utilizados. No caso da elaboração de classificadores, por exemplo, a qualidade preditiva dos atributos selecionados para compor o conjunto de treinamento é fundamental para a indução de boas regras de classificação, sendo a indicação de quais os atributos mais adequados para um experimento de estrita competência do especialista no domínio do problema estudado.

3.3.2 Estatística e mineração de dados

A modelagem de situações e conceitos de interesse para um especialista geralmente envolve, em algum momento, o uso de estratégias para a descoberta e/ou a comprovação da existência de algum tipo de estrutura lógica subjacente a um conjunto de dados proveniente das observações e têm sido instrumentalizadas por métodos quantitativos desenvolvidos no campo da Matemática, da Estatística e, mais recentemente, também de outras áreas. Essas técnicas mais recentes, como as dedicadas ao aprendizado de máquina e a mineração de dados, por exemplo, empregam uma diversidade de conceitos, originários de diversas áreas de conhecimento e, em geral, procuram endereçar classes de problemas mais específicos, sendo freqüente, por esse motivo, um certo desconforto ao se definir exatamente os contornos dessas abordagens e, principalmente, em distinguir suas características em relação ao chamado conhecimento estatístico convencional. Algumas das possíveis distinções podem ser de valor prático duvidoso, como determinadas referências a questões históricas e de nomenclatura, enquanto que outras podem permitir importantes considerações para a seleção da melhor técnica de modelagem. Neste trabalho, a utilização de uma ferramenta típica de mineração de dados como o See5 acabou por colocar a necessidade de reconhecer algumas dessas diferenças em relação à abordagem estatística mais tradicional, sendo esse o objetivo desta seção.

Muitos autores já discorreram a respeito da distinção e das ligações entre mineração de dados e estatística (HAND, 1998; MANILLA, 1996; ELDER, 1996; FAYYAD, 1996a; WEISS, 1998; entre outros) e, em geral, os principais aspectos considerados estão ligados ao enfoque do processo e às características dos dados envolvidos nas análises. A maior motivação para o surgimento de técnicas alternativas como as adotadas na mineração de dados foi o avanço da tecnologia da informática, que há décadas tem permitido uma crescente capacidade de processamento de informação a baixo custo e também viabilizou o armazenamento, ao longo do tempo, de um grande volume de dados, geralmente históricos, cujo estudo pode produzir um conhecimento valioso. É dentro desse contexto que as distinções entre os métodos geralmente são situadas.

A primeira distinção entre as técnicas de mineração de dados e a estatística convencional diz respeito ao enfoque de cada abordagem, e comporta duas perspectivas principais. O estudo estatístico normal tradicionalmente diz respeito a análises primárias, ou seja, que são realizadas com dados coletados especificamente para aquele estudo em particular, enquanto que a mineração de dados é tipicamente uma análise secundária (HAND, 1998), no sentido que os dados processados foram coletados originariamente com outros propósitos (como, por exemplo, registrar as transações de vendas de uma empresa ou monitorar o comportamento de um dispositivo qualquer). Outra diferença de enfoque diz respeito ao sentido da análise, segundo a qual o estudo estatístico é caracterizado como sendo dedutivo, já que parte de hipóteses que são então confrontadas com os dados, e a mineração de dados como indutiva, porque procura formular ("aprender") conceitos mais gerais a partir das evidências fornecidas por um conjunto de dados particular (MANNILA, 1996). A propósito desta forma de distinção, é curioso notar inclusive que o termo "*data mining*" (mineração de dados) teve origem na comunidade estatística, possuindo inicialmente uma conotação pejorativa visto que era usado para designar tentativas de análise de dados sem que houvesse uma hipótese claramente definida (MANNILA, 1996), cujo valor duvidoso deriva do fato de que se alguém analisar um conjunto de dados com grau suficiente de detalhe, vai sempre acabar encontrando algum tipo de padrão ou regularidade aparentemente significativo, mesmo que os dados tenham sido gerados aleatoriamente (FAYYAD, 1996a). Esse tipo de abordagem exploratória só passou a ser reconhecido como merecedor de maior crédito a partir do trabalho de John Tukey (TUKEY, 1977), que definiu seus propósitos básicos, indicações e limitações, bem como algumas técnicas.

O segundo tipo de distinção entre o trabalho estatístico convencional e a mineração de dados baseia-se nas características dos dados utilizados no processamento, das quais a mais evidente é o seu volume. Enquanto que os métodos estatísticos tradicionais são utilizados em análises meticulosas, de alta precisão, sobre quantidades de dados geralmente mais pequenas, como ocorre por exemplo no banco de dados disponível para este trabalho, onde o número de observações utilizadas geralmente não excede em muito o milhar, a mineração de dados tipicamente envolve imensos volumes de dados, freqüentemente na casa do milhão de observações. A questão da dimensionalidade também se coloca: o número de atributos / variáveis considerados também é geralmente bem maior nos problemas de mineração de dados, não sendo de todo incomuns casos com centenas de atributos, sendo este muitas vezes o maior fator de complexidade para o estudo, pois aumenta exponencialmente o tamanho do espaço de hipóteses (MANNILA, 1996). É recomendável portanto procurar diminuir o número de atributos considerados, sempre que possível, o que é obtido tanto pelo descarte puro e simples daqueles considerados menos importantes (a

qualidade preditiva ou descritiva dos atributos selecionados é sempre mais importante que a sua quantidade e variedade) e também pelo uso de técnicas de redução de dimensionalidade, que procuram substituir um subconjunto dos atributos originais por um único novo atributo que é gerado a partir deles (WEISS, 1998). Finalmente, uma outra decorrência interessante da manipulação desses grandes volumes de dados é que o próprio conceito de significância torna-se, em alguma medida, relativizado, já que mesmo padrões ocorrentes em um número proporcionalmente pequeno de casos apresentam uma significância estatística muito forte, inclusive alguns de veracidade duvidosa, o que acaba por colocar a necessidade da adoção de um conceito mais amplo de significância, como a chamada *significância substantiva* ("*substantive significance*") mencionada em (HAND, 1998) em oposição à clássica significância estatística, por requerer uma maior participação do especialista para determinar quais resultados e conclusões podem ser aproveitados ou não.

Uma outra característica importante dos dados que permite algum tipo de distinção entre a abordagem estatística clássica e a mineração de dados é a presença, nesta última, de grande quantidade de dados não-numéricos, o que requer a utilização de técnicas especiais, tanto para o tratamento direto desses dados como também para sua conversão para escalas numéricas ordenadas, sempre que possível. Finalmente, outros aspectos da mineração de dados que podem distingui-la da estatística quanto aos dados são a ocorrência de um maior volume de dados com valores inválidos, decorrentes de erros diversos de observação e de registro; a não-estacionaridade da população amostrada, que é relativamente freqüente nessa classe de problemas e também a dificuldade em se fazer uma amostragem mais aleatória das ocorrências a considerar em algumas situações, que pode gerar problemas de seleção tendenciosa (*bias*) e tornar a adoção de técnicas estatísticas já consagradas uma tarefa mais complicada ou, mesmo, inviável.

3.3.3 Considerações sobre o problema

O conjunto de dados utilizado no trabalho original consiste de cerca de 160 amostras coletadas em campo sob um esquema de amostragem estratificada orientada, mais adequado para dados onde o aspecto espacial é relevante, como no caso em questão. Cada ponto amostral possui cerca de 30 atributos, que são descritos em detalhe no capítulo referente ao projeto do banco de dados deste trabalho. Essas amostras referem-se a 27 pontos de coleta, situados em 9 regiões diferentes, denominadas de Superfícies Geneticamente Homogêneas (SGH). É portanto uma amostra de dimensionalidade relativamente alta e com um número de ocorrências extremamente baixo, desproporcional em relação ao número de atributos considerados. Existe ainda uma característica adicional complicadora, que é o fato de as amostras terem sido retiradas de profundidades diferentes do terreno, o que gera a presença de relacionamentos verticais em cada ponto de coleta, que são de difícil modelagem. O número de pontos amostrais para cada SGH é também variado, sendo comum SGHs com apenas um ou dois pontos distintos e, por motivos de ordem operacional e financeira, não houve qualquer possibilidade de se obter novas amostras em campo. Pretende-se com esses dados tentar identificar um conjunto de regras de classificação que permita associar cada ponto amostral com sua respectiva SGH, um típico trabalho de mineração de dados.

Dessa descrição bastante sucinta do repertório de dados disponível para o estudo, e também das considerações iniciais presentes neste capítulo, sobre amostragem e sobre a natureza do trabalho estatístico, que foi caracterizado por um menor volume de dados, analisados com muita precisão por técnicas mais consolidadas e largamente influenciadas por testes de significância, podemos chegar a algumas conclusões importantes. O problema em questão se aproxima da análise estatística clássica no que se refere ao pequeno número de ocorrências disponíveis para estudo e pelo fato de os dados terem sido coletados especificamente para esse tipo de classificação, caracterizando uma análise primária. Por outro lado, a abordagem indutiva do processo de produção de regras de classificação; a relativamente alta dimensionalidade, além é claro da própria ferramenta a ser usada no estudo, trazem o problema para a arena da mineração de dados. A falta de um maior volume de dados é crítica, na medida em que inviabiliza a sua divisão em conjuntos de treinamento e teste, indispensáveis para uma avaliação mais realista da taxa de erro apresentada pelo classificador obtido, e também impede a adoção de uma amostragem incremental que permita, com base no erro obtido, identificar o tamanho mais adequado para o conjunto de treinamento. Além disso, algoritmos de indução de regras de classificação são particularmente sensíveis a volumes de dados muito pequenos, pois sua estratégia de particionamento dos dados de treinamento vai gerando subconjuntos cada vez menores e mais homogêneos. Os subconjuntos resultantes serão proporcionalmente menores quanto maior for o número de classes e quanto mais equilibrada for a distribuição das ocorrências disponíveis pelas classes, sendo isso exatamente o que ocorre neste problema, caracterizado por uma classe dominante (a SGH IV) com cerca de 40% das amostras e uma muito baixa prevalência em cada uma das SGHs restantes.

Assim, com base em todas essas considerações, é necessário reconhecer que os resultados obtidos a partir dos experimentos que serão realizados neste trabalho deverão ser encarados em princípio como apenas especulativos, dado o pequeno suporte estatístico para conclusões apoiadas tão somente nos dados disponíveis para os experimentos, e poderão eventualmente ser interpretados como indícios a serem levados em conta em estudos posteriores sobre o local, apoiados por um maior volume de amostras. Essa restrição com certeza limita o alcance das conclusões a serem obtidas sobre os relacionamentos de maior interesse que podem ser identificados no domínio do problema, mas não inviabiliza o presente trabalho já que ainda assim é possível discutir e estabelecer um procedimento para aplicação da ferramenta proposta em trabalhos científicos como o utilizado neste estudo de caso.

3.4 O banco de dados

Este trabalho apresenta-se como uma aplicação de uma ferramenta de mineração de dados na área de pesquisa do meio-ambiente e, por esse motivo, possui um ciclo de vida típico de qualquer processo informatizado de extração de conhecimento, independentemente de possuir fins comerciais ou acadêmicos, ou ainda de qual o domínio do problema em questão. Tal processo é constituído de uma seqüência de passos que, de uma maneira geral, podem ser resumidos conforme a figura 13, adaptada de (FAYYAD, 1996b).

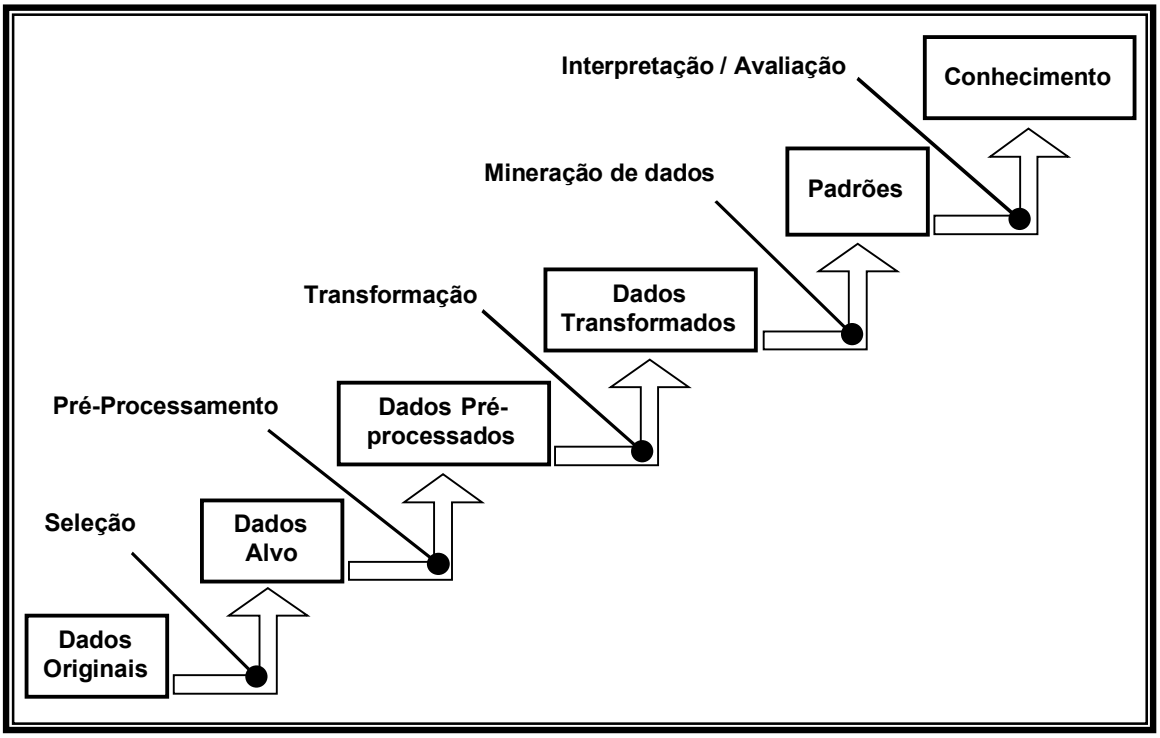


Figura 13: Passos principais do processo de extração de conhecimento em bancos de dados.

Como pode ser visto na figura, existe uma seqüência básica das etapas a serem cumpridas, sendo possível e mesmo comum retornar às fases anteriores para alterações no processo a partir das observações e avaliações finais realizadas pelo usuário. As etapas iniciais procuram reunir e organizar os dados necessários da maneira mais adequada para o processamento pelas ferramentas de mineração de dados. Tipicamente esse trabalho preliminar corresponde à identificação das diversas fontes de dados, à compilação dessas informações em um único conjunto de tabelas integrado e à depuração do conteúdo desse banco de dados, eliminando eventuais redundâncias e imprecisões. Finalmente, os dados depurados são convertidos para o formato específico requerido pelo software de mineração. A análise dos resultados é tipicamente uma tarefa do especialista, que pode fazer uso de gráficos e outros recursos visuais para facilitar seu trabalho.

O objetivo desta seção é descrever a estratégia e o esforço despendido na preparação dos dados necessários ao processamento pela ferramenta de classificação escolhida. Em linhas gerais, esses dados foram obtidos de (GARCIA, 1995), sendo que um material adicional

importante mas não publicado junto com o texto original foi conseguido através de comunicação pessoal com o autor da pesquisa.

3.4.1 Estratégia

O programa See5 utiliza informações contidas em arquivos texto sem formatação (ASCII) criados com a extensão .names para as definições gerais (metadados) sobre os casos a serem processados e .data para os valores a serem computados. Visando tornar o processo mais ágil, flexível e documentado foi gerado, através do sistema gerenciador de bancos de dados MS-Access, um banco de dados contendo todas as informações necessárias aos experimentos, no formato de tabelas normalizadas até a Terceira Forma Normal do modelo relacional (ELMASRI, 2000). Com base nessas tabelas, várias consultas foram montadas, de maneira a combinar e filtrar os dados utilizados em cada experimento, conforme mostra a figura 14.

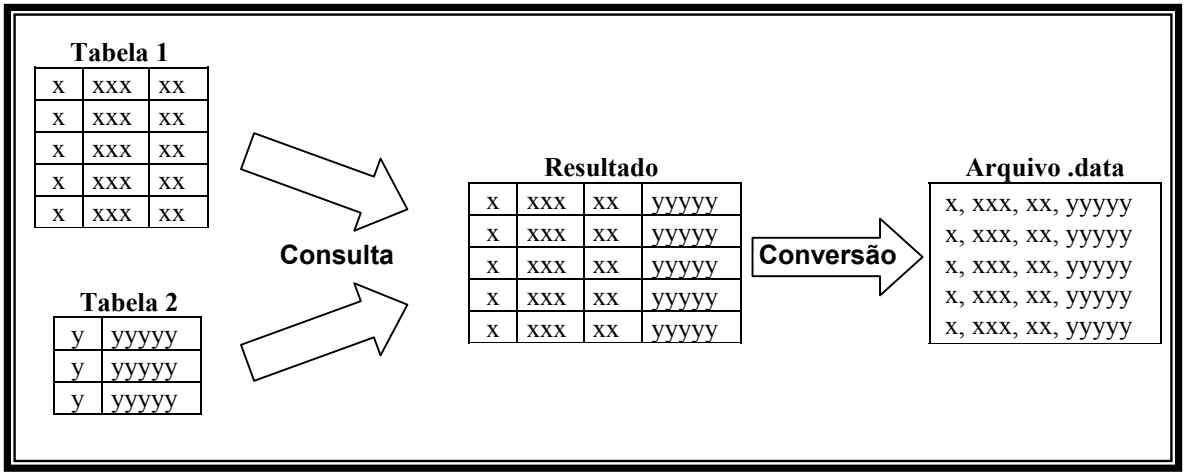


Figura 14: Estratégia de produção do arquivo de dados para o See5.

Um formato com especificações para a conversão e exportação automática de dados do MS-Access para arquivos ASCII delimitados no padrão utilizado pelo See5 foi definido para cada versão de experimento, tornando com isso trivial a geração dos dados brutos para a indução dos classificadores a partir das consultas montadas no interior do banco de dados. A construção dos arquivos .names de cada experimento foi feita manualmente, através de um processador de textos comum.

3.4.2 Os dados originais

Em seu trabalho, Garcia utiliza uma série de pontos de observação definidos na Planície Litorânea de Picinguaba com base em fotointerpretação, cada um deles gerando amostras de terreno em vários níveis de profundidade e descritos segundo 3 perspectivas:

a) uma análise macroscópica que retrata o aspecto geral da amostra com base em observação visual através de lupas especiais;

b) sua granulometria, expressa por meio de uma série de variáveis que representam a proporção entre as quantidades observadas em laboratório para areia, silte e argila na amostra;

c) sua composição química, que indica a quantidade verificada na amostra de alguns elementos considerados de maior importância para o domínio do problema.

Cada uma dessas perspectivas descritivas do material possui um significado próprio e complementar às demais, sendo que os dados macroscópicos e granulométricos desempenham um papel chave nas análises desenvolvidas no trabalho original e a composição química, apesar de reconhecida como promissora, praticamente não foi utilizada.

Os quadros a seguir apresentam as variáveis ou atributos utilizados e suas características principais, tendo sido montados com base em (GARCIA, 1995) e em comunicações pessoais com o autor do trabalho. As siglas colocadas entre parênteses indicam abreviaturas presentes no trabalho original e servem principalmente como referência.

Análise macroscópica

Atributo	***Descrição***
Cor	Cor predominante do material componente da amostra, expressa pelo seu nome na escala de Munsell.
Matéria orgânica	Indica a presença ou ausência de vestígios de matéria orgânica na amostra. No banco de dados esse atributo apresenta-se esparsamente povoado, com grande proporção de casos onde seu valor é desconhecido.
Textura	Representa textualmente características gerais de apresentação da amostra, como seu grau de agregação e aspecto granulométrico geral, por exemplo.
Composição mineral	Indica, textualmente, a presença dos principais minerais observados, como quartzo, micas e feldspato, eventualmente fazendo referência à tendência de comportamento dessa característica entre os diversos níveis de profundidade na amostra.
Grau de Seleção	Expressa o grau de facilidade no isolamento visual dos diversos componentes macroscópicos verificados na amostra.
Lençol freático	Indica a presença ou ausência do lençol freático na profundidade em que foi obtida a amostra.

Granulometria

Variável	***Descrição***
Areia total (AT)	É o percentual de areia verificado na amostra, com base na escala de Wentworth. O processo utilizado para a obtenção dos resultados foi uma adaptação da Análise Granulométrica por Dispersão Total de Argila pelo Método de Pipetagem.
Argila	Percentual de argila, obtido segundo o mesmo processo que AT.
Silte (S)	Percentual de silte, estimado pela diferença 100 – AT – Argila.
Areia muito grossa (AMG)	Percentual, observado na fração Areia Total, correspondente a areias com grãos entre 1 e 2 mm.
Areia grossa (AG)	Percentual, observado na fração Areia Total, correspondente a areias com grãos entre 0,5 e 1 mm.
Areia média (AM)	Percentual, observado na fração Areia Total, correspondente a areias com grãos entre 0,25 e 0,5 mm.
Areia fina (AF)	Percentual, observado na fração Areia Total, correspondente a areias com grãos entre 0,125 e 0,25 mm.
Areia muito fina (AMF)	Percentual, observado na fração Areia Total, correspondente a areias com grãos entre 0,064 e 0,125 mm.

Composição Química

Variável	***Descrição***
Nitrogênio total (N.T.)	Expresso em porcentagem.
Fósforo (P)	Expresso em partes por milhão.
Potássio (K)	Expresso em partes por milhão.
Cálcio (Ca)	Expresso em mEq (mili-equivalente).
Magnésio (Mg)	Expresso em mEq (mili-equivalente).
Alumínio (Al)	Expresso em mEq (mili-equivalente).
Sódio (Na)	Expresso em partes por milhão.
Carbono (C.O.)	Expresso em porcentagem.
Matéria orgânica (M.O.)	Expresso em porcentagem.
Índice pH em água (pH H2O)	
Índice pH em Cloreto de Potássio (pH KCl)	

Na montagem do banco de dados foram utilizados inicialmente os dados publicados com o trabalho original, digitados diretamente nas tabelas do banco de dados. Uma minuciosa verificação do conteúdo das tabelas foi realizada, tanto por meio de inspeção simples como também pela criação de consultas de totalização para aferir a coerência de valores numéricos como aqueles referentes à granulometria, que são interdependentes, sendo as discrepâncias anotadas para investigação. Posteriormente foram obtidas, através de entrevistas com o autor da pesquisa, cópias de planilhas, mapas e cadernetas de campo para fins de correção de possíveis erros de digitação e de impressão na versão final publicada. Alguns dados presentes nesse novo material que não constavam da publicação oficial foram também acrescentados, após autorização expressa do autor. Também algumas notações

textuais foram normalizadas nesse momento, pois abordagens automatizadas como a proposta requerem que nenhum tipo de ambigüidade ou redundância não intencional esteja presente nos dados. Um exemplo dessa prática pode ser observado no quadro a seguir, onde são apresentados os valores original e corrigido para a variável "Seleção", resultando na redução de seu domínio de 8 para 6 valores distintos, e que por isso representa muito mais do que apenas uma mudança de forma, podendo repercutir diretamente na estrutura e no desempenho das regras a serem induzidas.

Valor original	***Valor corrigido***
Mal selecionado	Mal selecionada
Bem selecionado	Bem selecionada
Muito mal selecionado	Muito mal selecionada
Muito bem selecionado	Muito bem selecionada
Boa seleção	Bem selecionada
Seleção média	Medianamente selecionada
Média seleção	Medianamente selecionada
Mal selecionado à muito mal selecionado com aumento em prof	Mal selecionada a muito mal selecionada com aumento em prof

Tabela 3: Exemplo de normalização dos valores de um atributo.

Finalmente, algumas considerações adicionais se fazem necessárias quanto ao volume e características dos dados disponíveis. As amostras foram obtidas em campo cavando-se um corte vertical no terreno escolhido e separando-se material de 10 em 10cm até a profundidade de 50cm e depois, de 20 em 20cm, até que o lençol freático fosse atingido. Isso faz com que, em princípio, tenhamos para um ponto hipotético onde o lençol freático se encontrasse a um metro de profundidade, amostras para os níveis 10cm, 20cm, 30cm, 40cm, 50cm, 70cm, 90cm e 110cm. Dadas as características do local, constituído por uma planície litorânea pequena mas bastante diversificada, o número de amostras por ponto varia substancialmente, em razão da própria variação da profundidade em que o lençol freático é encontrado nos diversos pontos. Em geral o lençol se encontra a pouca profundidade, sendo que em 22 dos 27 pontos as amostras foram coletadas até a profundidade máxima de 110cm e muitas delas não vão além de 50 ou 70cm. Assim, uma primeira característica importante dos dados é a sua diversidade quanto ao número de amostras por ponto, sendo que existem casos extremos de pontos com apenas uma ou duas amostras.

Um segundo aspecto a ser considerado deve-se à natureza bastante instável de alguns terrenos, que dificultou bastante o trabalho de escavação e fez com que um dos pontos tivesse suas amostras recolhidas às profundidades de15cm, 30cm e 45cm, diferindo de todo o restante quanto a esse espaçamento.

Motivos diversos ligados ao processamento do material de campo fizeram com que a informação referente à variável "Matéria Orgânica" da análise macroscópica fosse omitida para mais de 50% dos casos e cerca de 10% das amostras não possuam valores conhecidos para as variáveis "Seleção", "Composição mineral" e "Textura", também da análise macroscópica e aproximadamente 15% das amostras não possuem dados de granulometria conhecidos. No trabalho original tais índices não foram de importância significativa, dadas

as características da análise e interpretação final realizadas, mas representam uma informação importante quando se pretende proceder a uma abordagem baseada na aplicação de ferramentas de mineração automática de dados.

A realização de ensaios em laboratório para a identificação da composição química do material presente nas amostras requer uma série de produtos e equipamentos específicos que a tornam uma etapa bastante dispendiosa dentro de qualquer projeto na área de meio-ambiente. Esse importante fator financeiro, aliado ao menor peso relativo da análise química para as propostas do trabalho original fez com que apenas pouco mais de um terço das amostras disponham de valores conhecidos para as variáveis referentes à composição química.

Assim, pelo que foi exposto, temos caracterizado um quadro com um volume significativo de informações incompletas, principalmente se levarmos em conta que nem todos os problemas apontados ocorrem nas mesmas amostras. A rigor, apenas 7% das amostras possuem todas as informações possíveis conhecidas e cerca de 30% dispõem pelo menos do conjunto completo de atributos mais importantes ou seja, granulometria mais composição química.

3.4.3 Adaptações realizadas

Várias adaptações foram necessárias para que os dados originais pudessem ser processados com sucesso pela ferramenta, tanto devido às características limitadoras impostas pelo formato de dados dos arquivos de entrada adotado pelo programa como, principalmente, por razões técnicas ligadas ao significado atribuído pelo See5 aos dados durante a geração dos classificadores. Determinadas variáveis constantes no banco de dados não possuem utilização prevista nos experimentos inicialmente delineados mas mesmo assim sofreram adaptações nos casos em que algum tipo de problema mais sério ligado à uma eventual utilização pelo See5 tenha sido detectado, prática essa adotada tanto para explorar as possibilidades de ajustes do banco de dados como também para viabilizar testes adicionais que viessem a se fazer necessários.

No primeiro caso, foram eliminadas as possibilidades de alguns valores textuais mais extensos, geralmente descritivos e que possuem vírgulas em seu interior, já que a ferramenta atribui um significado próprio para o símbolo “,” quando encontrado em um arquivo .data. Nesses casos, optou-se por ignorar o atributo quando possível, ou então codificá-lo numericamente, através de tabelas auxiliares onde a cada valor distinto de texto foi atribuído um número de identificação.

Problemas relacionados com o significado dos dados, que constituem o segundo caso motivador para as adaptações, ocorreram principalmente ao se definir os arquivos .names para os ensaios. O mecanismo interno do See5 que gera os classificadores precisa saber como interpretar um determinado valor para um atributo e, para isso, consulta as definições do arquivo .names correspondente. Ali, na declaração de cada atributo, é especificado também o seu conjunto de valores válidos, chamado de domínio do atributo, no jargão de banco de dados, que pode ser (RULEQUEST RESEARCH):

a) <continuous> (contínuo), para faixas de valores numéricos;

b) <lista de rótulos separados por vírgula> (discreto), para valores textuais, que podem ser tratados também como uma lista ordenada se o parâmetro [ordered] for especificado para o atributo;

c) <discrete N> (discreto), para valores inteiros;

d) <date> (data), para datas no formato ano, mês e dia, com valores válidos a partir de 1601 dC;

e) <ignore> (ignorar), para atributos que não devem ser considerados na classificação;

f) <label> (rótulo), não é utilizado na classificação, mas apenas para referir casos individuais, sendo útil ao se fazer a validação cruzada dos dados.

Freqüentemente as codificações adotadas assumem valores numéricos e, em alguns casos, podem levar a listas mais ou menos extensas dos códigos possíveis, de maneira que um atributo com esse domínio, conforme seja declarado como sendo *continuous*, *discrete* ou como uma lista de valores no arquivo .names, apresentará impactos diferentes e freqüentemente significativos no classificador resultante do processamento. Isso ocorre porque para valores declarados como discretos ou configurados como listas de valores em sua definição padrão (sem fazer uso da cláusula opcional [ordered]), o See5 considera que não existe nenhuma relação de ordem entre os dados, e por isso os mesmos são tratados apenas como nomes individuais ou, no máximo, como passíveis de algum agrupamento pelo programa, desde que definidos expressamente pelo usuário através da opção *Discrete Value Subsets*. Por outro lado, um atributo declarado como contínuo é segmentado automaticamente pela ferramenta, que então pode fazer uso de comparadores como "maior que", "menor que", "maior ou igual a" que permitem um certo nível de generalização no resultado. Conforme o caso, também uma utilização de limiares nebulosos (que implementa conceitos de lógica *fuzzy* através da opção *Fuzzy Thresholds*) pode ser adotada, recurso esse só possível para valores contínuos e que pode fornecer critérios de classificação mais flexíveis, ainda que ligeiramente mais complexos.

Os casos que mereceram adaptações no projeto foram relativos aos atributos "Composição Mineral", "Texturas", "Seleções" e "Cores". O primeiro deles, "Composição Mineral", apresenta 31 distintos valores textuais descritivos, onde aparece a vírgula (que impede a exportação desse dado diretamente para o arquivo .data devido à formatação padrão requerida pela ferramenta) e onde a codificação numérica simples como um tipo discreto seria inconveniente, já que sua alta diversidade tenderia a torná-lo pouco representativo como critério de classificação. Isso levou ao estabelecimento de uma escala ordenada de valores para o atributo, que pôde então ser definido como do tipo "contínuo" nos experimentos. O atributo "Cores" também possui 31 valores distintos no banco de dados, e mostrou-se passível de conversão em uma escala ordenada, ainda que não venha a ser utilizado nos experimentos devido ao pequeno valor prático que representa para o domínio do problema em questão. Com relação aos atributos "Textura" e "Seleção", com 27 e 7 valores distintos respectivamente, foi também possível estabelecer uma escala ordenada de valores, com um claro sentido de orientação (de um menor valor para um maior), conforme mostra o exemplo contido na tabela a seguir. Uma descrição do conteúdo das tabelas referentes a "Seleção", "Textura" e "Composição Mineral" pode ser encontrada em anexo, no final deste trabalho.

Cód.	*Seleção*	*Ordem*
1	Muito mal selecionada	0
6	Mal selecionado a muito mal selecionado com aumento em prof	1
2	Mal selecionada	2
7	Média seleção piorando com o aumento da profundidade	3
3	Medianamente selecionada	4
4	Bem selecionada	5
5	Muito bem selecionada	6

Tabela 4: Escala ordinal para os valores do atributo Grau de Seleção das amostras.

Esse tipo de escala é adequado para medições qualitativas como a que ocorre no caso (PEREIRA, 1999), e permite que se faça uso das relações de ordem, independentemente de seus valores absolutos, o que viabiliza a interpretação, para fins da geração de classificadores, desses atributos como contínuos. Assim, ao exportar os dados para o arquivo .data, é enviado o código ordinal do atributo (correspondente ao seu conteúdo na coluna "Ordem" do exemplo) e não o valor textual em si, permitindo então à ferramenta gerar classificadores que ao final afirmem, por exemplo, que amostras com grau de seleção <= "Medianamente selecionada" (o que engloba 5 das 7 categorias de Seleção) tipicamente pertencem a uma dada região, numa tradução mais ou menos livre do resultado, o que seria impraticável se o atributo fosse definido como discreto, mesmo que ordenado.

3.5 Os experimentos

A avaliação das possibilidades do emprego de técnicas de classificação automática no estudo ambiental baseou-se na análise tanto do seu processo de utilização como dos resultados fornecidos pelo método computacional adotado, aplicado em um caso em que se pretende a indução de critérios de classificação de uma região em unidades ambientais homogêneas. Para isso foram realizados dois experimentos aplicando o *software* selecionado sobre os dados extraídos do trabalho original. Como esses dados foram coletados em diversos níveis de profundidade, e o número total de amostras é muito reduzido, optou-se por utilizar apenas o material proveniente dos níveis entre 10cm e 40cm, procurando com isso minimizar a interferência de eventuais relacionamentos verticais presentes dentro de cada ponto de coleta. Os experimentos são subdivididos em versões, cada uma delas correspondendo ao conjunto de atributos que foi considerado para a elaboração do classificador, de maneira a permitir entender o efeito de cada grupo de atributos no resultado final. A tabela 5 mostrada mais adiante apresenta um resumo sobre a estrutura de versões dos dois experimentos realizados.

Inicialmente foi gerado um classificador com base apenas nos dados sobre a granulometria e análise macroscópica, para uma comparação com a classificação realizada pelos meios interpretativos convencionais, como ocorre no trabalho de referência. Em seguida, um novo experimento procurou considerar também os resultados da análise química das amostras, disponíveis no trabalho original mas que não foram ali utilizados. Cada experimento encontra-se descrito em detalhes a seguir, sendo que o resultado de cada um é

posteriormente comparado com aqueles encontrados no trabalho de Garcia, de maneira a verificar o grau de correspondência entre as classificações. É importante ressaltar que os resultados deste estudo não têm, em momento algum, a finalidade de contestar as conclusões do trabalho de referência, uma vez que seu foco está na proposição de métodos válidos para o auxílio à investigação, e não no mérito da análise final. Por fim, devem ser mais uma vez ressaltadas as limitações quantitativas do material disponível para os experimentos, que não permitem a adoção de estratégias mais adequadas para a geração de classificadores, como a avaliação dos erros através de conjuntos de teste independentes dos conjuntos de treinamento ou ainda o uso de abordagens como N validações cruzadas, por exemplo.

Experimento 1: classificação pelos dados da análise granulométrica e macroscópica

O objetivo deste procedimento é colher elementos que permitam, a exemplo do que ocorre no trabalho de referência, verificar se há algum tipo de contradição entre as características observadas nos aspectos granulométricos e macroscópicos de cada amostra em relação ao perfil da respectiva SGH.

Essa verificação é feita induzindo-se um classificador a partir de um conjunto de amostras (o *training set*) onde consta as proporções relativas entre os componentes Areia (fracionado em 5 categorias), Argila e Silte bem como o código da SGH correspondente a cada amostra. O resultado é então representado tanto através de árvore de decisão como de regras de produção expressas na forma antecedente / conseqüente, de maneira a ilustrar o critério em termos compreensíveis para um especialista no domínio. Uma segunda variação mais completa deste experimento considera também os dados mais importantes da análise macroscópica, no caso, os atributos Grau de Seleção, Textura e Composição Mineral do material coletado.

Em uma situação ideal, cada classificador obtido seria então aplicado sobre um novo conjunto de amostras (o chamado conjunto de teste), descritas nos mesmos termos do *training set* inicial, para uma melhor estimativa do erro de classificação observado. Entretanto, como não há um volume de dados suficiente para esse procedimento, a taxa de erro a ser considerada é aquela obtida sobre o próprio conjunto de treinamento, que é menos confiável por tender a ser mais reduzida que a taxa de erro real. O resultado da classificação automática é em seguida comparado com aquele presente no trabalho de referência, para uma análise sobre a adequação e a qualidade do critério que foi induzido.

Experimento 2: classificação pelos dados das análises granulométrica, macroscópica e química

O objetivo deste experimento é verificar a influência da composição química na caracterização das amostras de cada SGH. A estratégia, nesse caso, consiste em gerar um *training set* semelhante ao do primeiro experimento, mas com os casos descritos em termos das diversas componentes químicas analisadas e apresentadas no trabalho de referência, juntamente com a identificação da SGH correspondente. Sobre esse subconjunto dos dados será gerado um classificador, que posteriormente será avaliado à semelhança do primeiro experimento.

Adicionalmente é gerado um classificador que considera também a combinação entre as características macroscópicas, granulométricas e químicas das amostras, reunidas em um único *training set*, de maneira a permitir confrontar essa classificação teoricamente mais completa com aquelas parciais geradas anteriormente a partir de apenas um ou dois desses fatores (análise macroscópica, granulometria e composição química).

Experimento / Versão	***Análise Macroscópica***	***Análise Granulométrica***	***Análise Química***
1 a		X	
1 b	X	X	
2 a			X
2 b		X	X
2 c	X	X	X

Tabela 5: As diferentes versões dos experimentos e os dados utilizados

Capítulo 4: Resultados

A realização de cada experimento produziu como resultado um classificador específico, induzido pela ferramenta a partir de seu respectivo conjunto de treinamento. Este capítulo tem por objetivo descrever o classificador obtido em cada experimento juntamente com uma interpretação sucinta sobre o seu significado. A apresentação dos resultados de cada um dos experimentos constará de um rápido preâmbulo mencionando os dados que foram utilizados no treinamento; as saídas fornecidas pela ferramenta See5; uma figura que representa a árvore de decisão de uma maneira alternativa; uma descrição textual da árvore de decisão e do conjunto de regras induzido e, finalmente, um texto com uma interpretação mais ou menos livre sobre o classificador que foi obtido.

No relatório de saída gerado pelo See5 ao final do processo de indução consta, além do classificador em si, também uma análise dos erros verificados, apresentada na forma de uma matriz de confusão, que foi mantida neste material devido ao seu grande valor prático para o usuário final. Entretanto, quando no mesmo processo de indução é gerada tanto a árvore de decisão como o conjunto de regras de produção correspondente, apenas a matriz de confusão destas é fornecida pelo programa nesse relatório, o que pode ser inconveniente quando as taxas de erro forem distintas entre essas duas representações. Por esse motivo, optou-se por produzir a árvore de decisão separadamente do respectivo conjunto de regras, de maneira a poder apresentar a matriz de confusão específica de cada caso.

A figura que acompanha o resultado de cada experimento representa exatamente a mesma árvore de decisão indicada no relatório de saída do See5, tendo sido desenhada por meio de um processador de textos com um tratamento gráfico um pouco mais elaborado, visando uma maior clareza da estrutura do classificador.

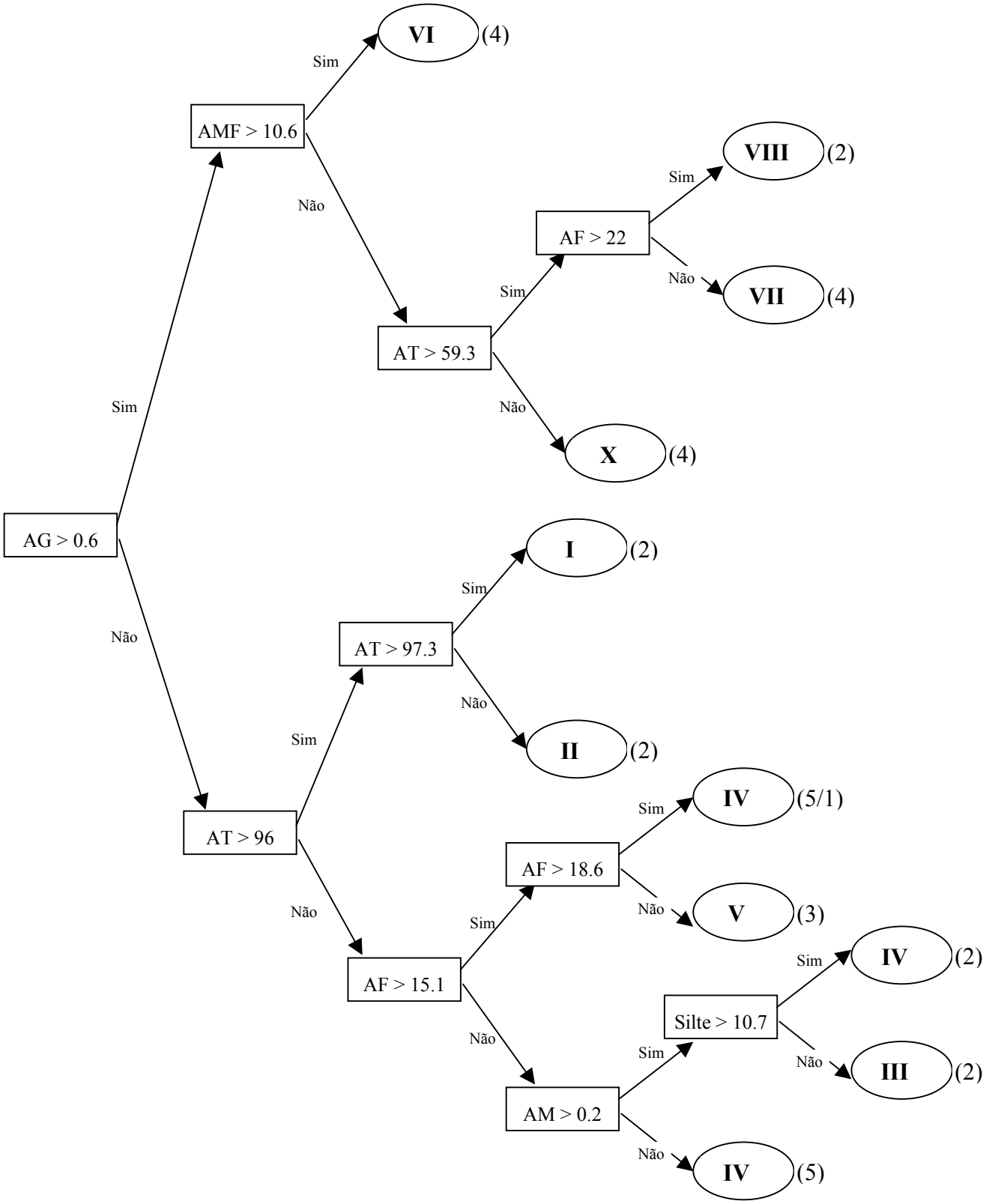
Como foi mencionado anteriormente, os dois experimentos propostos deram origem a um conjunto relativamente grande de versões, indicadas por letras. Assim, o experimento *1a* é primeira versão do experimento 1, e utiliza apenas os dados da granulometria para a determinação da SGH; a versão *1b*, por sua vez, considera também os dados da análise macroscópica, etc. Além dessas distinções entre as diversas versões de um mesmo experimento, baseada no conjunto de atributos utilizado; há também uma diferença com relação a quais os níveis de profundidade que foram considerados. Cada versão dos experimentos foi submetida ao mecanismo indutor de regras de duas maneiras: a primeira com dados coletados apenas nas profundidades entre 10cm e 20cm, teoricamente mais imunes à variação vertical dos diversos componentes no ponto de coleta, porém, com um menor volume de amostras; a segunda, com os dados coletados nas profundidades entre 10cm e 40cm, onde a possibilidade de variação vertical na composição das amostras é teoricamente mais significativa, porém com um volume de casos igualmente maior.

Com relação à análise dos resultados, são necessários alguns esclarecimentos adicionais sobre a abordagem adotada. Tanto o texto que descreve um classificador como aquele que o interpreta podem mencionar, quando necessário, certos valores de referência tais como a média de um atributo dentro do conjunto de treinamento considerado ou a descrição correspondente aos diversos códigos usados para os atributos que foram organizados como escalas ordenadas, como a Textura do material, por exemplo. O objetivo de tais colocações

é sempre o de conferir um maior significado para os valores de corte calculados pelo mecanismo indutor. Por esse motivo, um pequeno resumo estatístico sobre os dados utilizados em cada experimento é fornecido em anexo, no final deste trabalho, juntamente com as tabelas que descrevem as escalas ordenadas geradas para atributos como a Textura mencionada anteriormente. Em algumas poucas situações é feita referência ao valor de um caso ou amostra em particular, geralmente para comentar alguma anomalia ou curiosidade. Esse procedimento geralmente desaconselhável é permitido neste trabalho apenas devido ao muito pequeno volume de dados disponível, e justifica-se basicamente para tratar de situações caracterizadas por uma muito baixa prevalência numa classe. Por fim, em algumas situações, um juízo de valor é emitido com relação à qualidade de algumas regras induzidas pela ferramenta. Esse julgamento baseia-se tanto na taxa de erro da regra, como em seu grau de complexidade, como ainda em seu significado prático. Assim, taxas de erro elevadas indicam regras ruins; regras com muitas condições são consideradas piores que regras com poucas condições (Navalha de Occam); e regras onde um atributo é comparado com um valor ou faixa de valores pouco significativo dentro de seu contexto são consideradas de menor qualidade.

4.1 Experimento 1

4.1.1 Versão 1a (10-20cm)



Árvore

```
See5 [Release 1.12]     Sun Nov 05 11:29:32 2000
Class specified by attribute `SGH'
Read 35 cases (10 attributes) from Exp1a.data

Decision tree:
AG > 0.6:
:...AMF > 10.6: VI (4)
:   AMF <= 10.6:
:   :...AT <= 59.3: X (4)
:       AT > 59.3:
:       :...AF <= 22: VII (4)
:           AF > 22: VIII (2)
AG <= 0.6:
:...AT > 96:
    :...AT <= 97.3: II (2)
    :   AT > 97.3: I (2)
    AT <= 96:
    :...AF > 15.1:
        :...AF <= 18.6: V (3)
        :   AF > 18.6: IV (5/1)
        AF <= 15.1:
        :...AM <= 0.2: IV (5)
            AM > 0.2:
            :...S <= 10.7: III (2)
                S > 10.7: IV (2)


Evaluation on training data (35 cases):

           Decision Tree
         ----------------
         Size      Errors
           11    1( 2.9%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                          (a): class I
                  2                                                    (b): class II
                        2                                              (c): class III
                             11                                        (d): class IV
                              1     3                                  (e): class V
                                          4                            (f): class VI
                                                4                      (g): class VII
                                                      2                (h): class VIII
                                                                       (i): class IX
                                                                  4    (j): class X
```

Regras

```
See5 [Release 1.12]     Sun Nov 05 11:31:24 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 35 cases (10 attributes) from Exp1a.data

Extracted rules:
Rule 1: (2, lift 13.1)
        AT > 97.3
        ->  class I  [0.750]

Rule 2: (2, lift 13.1)
        AT > 96
        AT <= 97.3
        ->  class II  [0.750]

Rule 3: (2, lift 13.1)
        AT <= 96
        S <= 10.7
        AG <= 0.6
        AM > 0.2
        AF <= 15.1
        ->  class III  [0.750]

Rule 4: (5, lift 2.7)
        AT <= 96
        AM <= 0.2
        AF <= 15.1
        ->  class IV  [0.857]

Rule 5: (2, lift 2.4)
        S > 10.7
        AG <= 0.6
        ->  class IV  [0.750]

Rule 6: (5/1, lift 2.3)
        AG <= 0.6
        AF > 18.6
        ->  class IV  [0.714]

Rule 7: (3, lift 7.0)
        AG <= 0.6
        AF > 15.1
        AF <= 18.6
        ->  class V  [0.800]

Rule 8: (4, lift 7.3)
        AG > 0.6
        AMF > 10.6
        ->  class VI  [0.833]
```

```
Rule 9: (4, lift 7.3)
        AT > 59.3
        AF <= 22
        AMF <= 10.6
        ->  class VII  [0.833]

Rule 10: (2, lift 13.1)
        AF > 22
        AMF <= 10.6
        ->  class VIII  [0.750]

Rule 11: (4, lift 7.3)
        AT <= 59.3
        ->  class X  [0.833]

Default class: IV


Evaluation on training data (35 cases):

           Decision Tree           Rules
         ----------------     ----------------
         Size      Errors      No      Errors
           11    1( 2.9%)       11    1( 2.9%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                  2                                                     (b): class II
                        2                                               (c): class III
                             11                                         (d): class IV
                             1     3                                    (e): class V
                                         4                              (f): class VI
                                               4                        (g): class VII
                                                     2                  (h): class VIII
                                                                        (i): class IX
                                                                  4     (j): class X
```

Análise dos resultados

Este experimento utiliza apenas dados da análise granulométrica, nas profundidades de 10 a 20cm, perfazendo um total de 35 amostras (ou casos) com 10 atributos cada uma, sendo 9 deles significativos.

Árvore de decisão

Apenas um caso (ou amostra) foi classificado incorretamente, o que representa 2.9% do total. De saída foram induzidos dois grandes subconjuntos de SGHs, um deles composto por 4 regiões (VI, VII, VIII e X) e o outro correspondendo às 5 restantes, cuja principal diferença reside na quantidade de Areia Grossa observada.

As SGHs VI, VII, VIII e X, que formam o primeiro grupo e possuem uma idade relativa de antiga para média, caracterizam-se inicialmente pela presença de Areia Grossa em quantidades perceptíveis (chegando a cerca de 30% do total da amostra em alguns casos). Além disso, a área VI possui teores de Areia Muito Fina significativos, ainda que bem abaixo da média geral, destacando-se das regiões VII, VIII e X, que apresentam teores bastante baixos desse tipo de material. Adicionalmente, a região X destoa dentro deste subconjunto por possuir uma relativamente pequena proporção de Areia Total (em torno de 50% para uma média geral de 84%) e a região VIII destaca-se por possuir uma relativamente alta proporção de Areia Fina.

No segundo grupo, formado pelas SGHs I, II, III, IV e V, a característica dominante é uma quantidade extremamente baixa de Areia Grossa (e também de Areia Muito Grossa, se observarmos o banco de dados). As regiões I e II destacam-se por apresentar índices muito elevados de Areia Total (superiores a 96%, basicamente material fino e muito fino com pouco Silte ou Argila) e a região V caracteriza-se por possuir uma proporção de Areia Fina igual ou superior à média geral das amostras. Duas regiões não ficaram muito bem caracterizadas: a III, onde talvez a principal característica seja uma quantidade relativamente alta (em torno de 75%, contra 51% da média geral) de Areia Muito Fina, não apontada na árvore de decisão; e a IV, espalhada em três subgrupos: um com 5 casos onde percebe-se uma boa proporção de Areia Fina, outro com 2 casos onde há uma presença significativa de Silte e outro ainda, com 5 casos, caracterizado na árvore como possuindo proporções pequenas de Areia Fina e Areia Média, mas que apresentam-se com predomínio de Areia Muito Fina, Silte e Argila.

Regras de Decisão

A exemplo do que ocorreu na árvore de decisão, apenas um caso (ou amostra) foi classificado incorretamente, o que representa 2.9% do total, e foram geradas 11 regras, uma para cada classe, exceto a classe IV, cujos casos são definidos em 3 regras distintas.

A SGH I é definida pela regra 1 apenas em termos de sua quantidade de Areia Total: amostras com mais de 97.3% desse tipo de material pertencem a essa região.

O percentual de Areia Total também define a classe II: amostras com valores entre 96% e 97.3% para esse atributo são desta SGH, de acordo com a regra 2. Pouca Areia Grossa (inferior a 0.6%) e Areia Fina entre 15.1% e 18.6% indicam que o material coletado pertence à SGH V, conforme descreve a regra 7.

A SGH VI possui Areia Grossa superior a 0.6% e Areia Muito Fina superior a 10.6% (regra 8). Percentuais de Areia Fina e Areia Muito Fina abaixo da média geral correspondem a amostras da SGH VII, quando associados à teores de Areia Total acima de 59% (regra 9).

Quantidade significativa de Areia Fina acompanhada de pouca Areia Muito Fina (inferiores a 10.6% quando a média geral desse tipo de material é 51%) indicam amostras referentes à SGH VIII, segundo a regra 10.

Proporção pequena de Areia Total (inferior a 59%, para uma média de 84%) devem ser classificadas como pertencentes à SGH X pela regra 11.

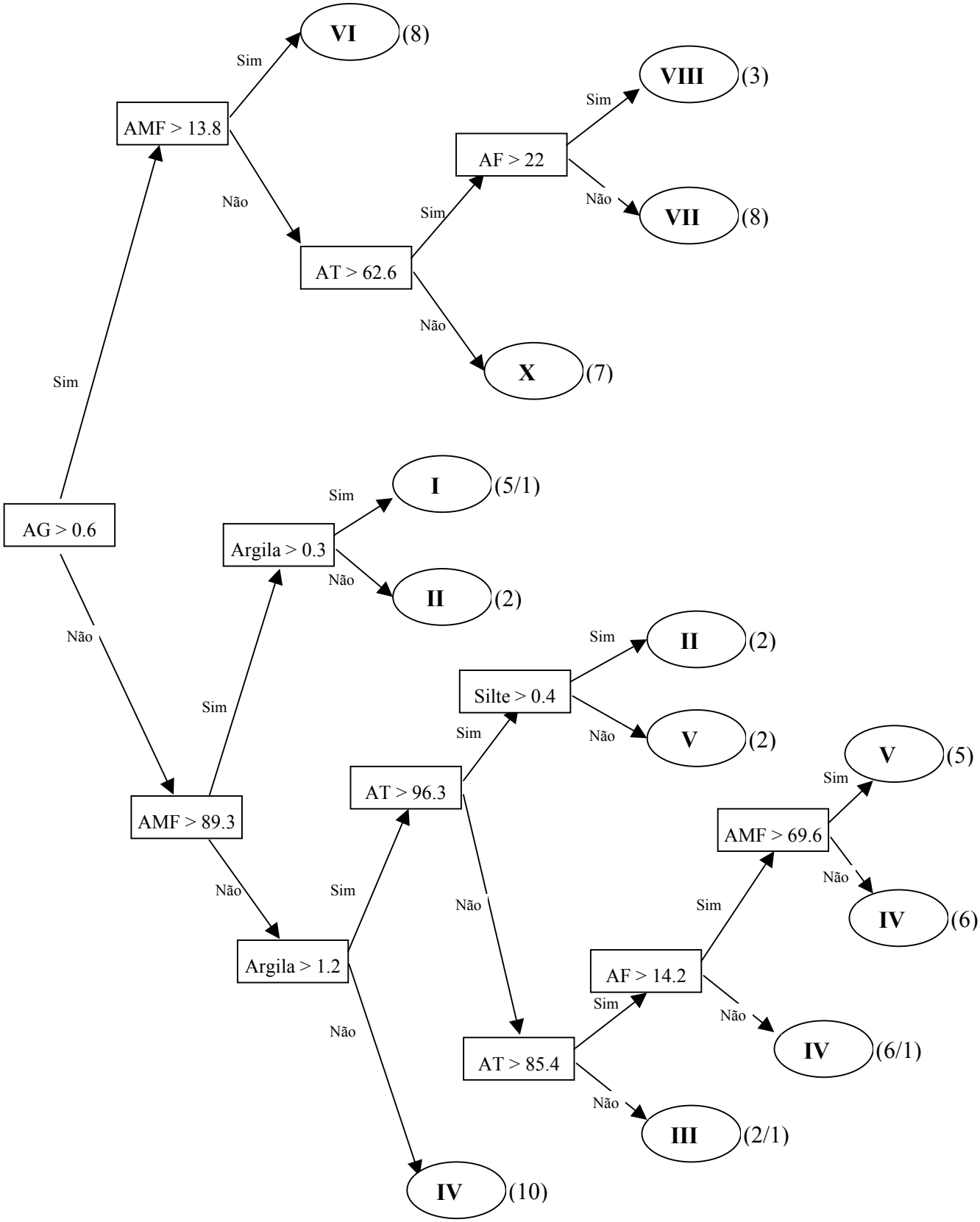
A região III é definida por uma regra relativamente complexa, que considera 5 atributos diferentes: muito pouca Areia Grossa; Areia Fina pouco abaixo da média geral; Silte abaixo de 10.7%; Areia Total abaixo de 96%, pouca Areia Média. Essa composição de atributos é de difícil interpretação prática, principalmente se levarmos em conta a grande variação dos valores observados em relação às médias gerais de cada atributo.

Finalmente, a região IV é definida pelas regras 4, 5 e 6. A primeira delas baseia-se em 5 casos, para os quais a Areia Total é um pouco acima da média (mas menor que 96%), a Areia Média é praticamente inexistente e a Areia Fina é também abaixo da média. Segundo a regra 5, que classifica 2 dos casos, a quantidade de Silte deve ser maior que a média e o teor de Areia Grossa deve ser muito pequeno. Finalmente, a regra 6 especifica que a Areia Grossa deve ser em quantidade muito pequena, com Areia Fina acima da média geral.

Interpretação dos resultados

A quantidade presente de Areia Grossa indica que existem dois grupos de SGHs principais: um com forte presença de material arenoso mais grosseiro e quantidades de Argila e Silte expressivamente acima da média geral das amostras, representado pelas regiões VI (denominada "Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos"), que correspondem a regiões relativamente antigas e a terrenos de composição mais diversificada; e outro, representado pelas SGHs I ("Planície de Maré"), II ("Berma"), III ("Duna"), IV ("Planície Litorânea de Cordões Regressivos") e V ("Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes"), onde há um forte predomínio de areias finas e muito finas e uma proporção relativamente pequena de Silte e Argila. Duas regiões em particular apresentam-se com contornos pouco precisos quanto ao aspecto granulométrico: a III, cujas definições apresentadas na árvore de decisão e na regra de decisão são idênticas, mas com complexidade relativamente alta e pouco significado prático; e a IV, talvez por possuir um maior número de casos e cobrir uma área bem maior que as demais, para a qual foi necessário dividir seus casos em 3 subconjuntos, um deles com expressiva quantidade de Areia Fina e outros dois onde esse material aparece em quantidades inferiores à média geral. Em um desses subconjuntos, composto por apenas 2 amostras, provavelmente coletadas no mesmo local, mas em profundidades diferentes, há uma grande proporção de Silte e uma quantidade relativamente pequena de Areia Total; em outro, há um predomínio de Areia Muito Fina e pouco Silte ou Argila.

4.1.2 Versão 1a (10-40cm)



Árvore

```
See5 [Release 1.12]     Sun Nov 05 11:37:46 2000
Class specified by attribute `SGH'
Read 66 cases (10 attributes) from Exp1a.data

Decision tree:
AG > 0.6:
:...AMF > 13.8: VI (8)
:   AMF <= 13.8:
:   :...AT <= 62.6: X (7)
:       AT > 62.6:
:       :...AF <= 22: VII (8)
:           AF > 22: VIII (3)
AG <= 0.6:
:...AMF > 89.3:
    :...Argila <= 0.3: II (2)
    :   Argila > 0.3: I (5/1)
    AMF <= 89.3:
    :...Argila <= 1.2: IV (10)
        Argila > 1.2:
        :...AT > 96.3:
            :...S <= 0.4: V (2)
            :   S > 0.4: II (2)
            AT <= 96.3:
            :...AT <= 85.4: III (2/1)
                AT > 85.4:
                :...AF <= 14.2: IV (6/1)
                    AF > 14.2:
                    :...AMF <= 69.6: IV (6)
                        AMF > 69.6: V (5)


Evaluation on training data (66 cases):

          Decision Tree
        ----------------
        Size      Errors
          13    3( 4.5%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           4                                                          (a): class I
                 4                                                    (b): class II
                       1     1                                        (c): class III
           1                21                                       (d): class IV
                       1           7                                  (e): class V
                                         8                            (f): class VI
                                               8                      (g): class VII
                                                     3                (h): class VIII
                                                                      (i): class IX
                                                           7          (j): class X
```

Regras

```
See5 [Release 1.12]     Sun Nov 05 11:38:52 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 66 cases (10 attributes) from Exp1a.data

Extracted rules:
Rule 1: (5/1, lift 11.8)
        Argila > 0.3
        AMF > 89.3
        ->  class I  [0.714]

Rule 2: (2, lift 12.4)
        AT > 96.3
        S > 0.4
        Argila > 1.2
        ->  class II  [0.750]

Rule 3: (2, lift 12.4)
        Argila <= 0.3
        AMF > 89.3
        ->  class II  [0.750]

Rule 4: (2/1, lift 16.5)
        AT <= 85.4
        Argila > 1.2
        AG <= 0.6
        ->  class III  [0.500]

Rule 5: (10, lift 2.7)
        Argila <= 1.2
        AG <= 0.6
        AMF <= 89.3
        ->  class IV  [0.917]

Rule 6: (7, lift 2.7)
        AT > 85.4
        AG <= 0.6
        AMF <= 69.6
        ->  class IV  [0.889]

Rule 7: (11/1, lift 2.5)
        AT > 85.4
        AT <= 96.3
        AG <= 0.6
        AF <= 14.2
        ->  class IV  [0.846]

Rule 8: (5, lift 7.1)
        Argila > 1.2
        AF > 14.2
        AMF > 69.6
        ->  class V  [0.857]
```

```
Rule 9: (2, lift 6.2)
        AT > 96.3
        S <= 0.4
        AMF <= 89.3
        ->  class V  [0.750]

Rule 10: (8, lift 7.4)
        AG > 0.6
        AMF > 13.8
        ->  class VI  [0.900]

Rule 11: (8, lift 7.4)
        AT > 62.6
        AF <= 22
        AMF <= 13.8
        ->  class VII  [0.900]

Rule 12: (3, lift 17.6)
        AF > 22
        AMF <= 13.8
        ->  class VIII  [0.800]

Rule 13: (7, lift 8.4)
        AT <= 62.6
        AMF <= 13.8
        ->  class X  [0.889]

Default class: IV


Evaluation on training data (66 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
           13    3( 4.5%)      13    2( 3.0%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            4                                                           (a): class I
                  4                                                     (b): class II
                        1     1                                         (c): class III
                              22                                        (d): class IV
                        1           7                                   (e): class V
                                          8                             (f): class VI
                                                8                       (g): class VII
                                                      3                 (h): class VIII
                                                                        (i): class IX
                                                                  7     (j): class X
```

Análise dos resultados

Este experimento utiliza apenas dados da análise granulométrica, nas profundidades de 10 a 40cm, perfazendo um total de 66 amostras (ou casos) com 10 atributos cada uma, sendo 9 deles significativos.

Árvore de decisão

Três casos (ou amostras) foram classificados incorretamente, o que representa 4.5% do total. De saída foram induzidos dois grandes subconjuntos de SGHs, um deles composto por 4 regiões (VI, VII, VIII e X) e o outro correspondendo às 5 restantes, cuja principal diferença reside na quantidade de Areia Grossa observada.

As SGHs VI, VII, VIII e X, que formam o primeiro grupo e possuem uma idade relativa de antiga para média, caracterizam-se inicialmente pela presença de Areia Grossa em quantidades perceptíveis (chegando a cerca de 40% do total da amostra em alguns casos). Além disso, a área VI possui teores de Areia Muito Fina significativos, ainda que bem abaixo da média geral, destacando-se das regiões VII, VIII e X, que apresentam teores bastante baixos desse tipo de material. Adicionalmente, a região X destoa dentro deste subconjunto por possuir uma relativamente pequena proporção de Areia Total (em torno de 50% para uma média geral de 86%, o que indica uma proporção elevada de Silte e Argila nessas amostras) e a região VIII destaca-se por possuir uma relativamente alta proporção de Areia Fina.

No segundo grupo, formado pelas SGHs I, II, III, IV e V, a característica dominante é uma quantidade extremamente baixa de Areia Grossa (e também de Areia Muito Grossa, se observarmos o banco de dados). A região I destaca-se por apresentar índices muito elevados de Areia Muito Fina (superior a 89%, com muito pouco Silte ou Argila) e a região III foi muito mal identificada pela ferramenta, já que possui apenas duas amostras e uma delas foi classificada com erro (na região IV), além do que o critério esboçado pela árvore leva a uma segunda incorreção, em que um caso da região V é classificado como sendo da região III.

As amostras da região II aparecem em duas folhas da árvore, e olhando melhor o banco de dados e o classificador, observa-se que referem-se a 4 observações de um mesmo ponto, sendo que em uma das folhas estão as amostras de camadas mais superficiais (retiradas a 10 e 20 cm) e em outra, as amostras das camadas mais profundas (em 30 e 40 cm). Estas últimas caracterizam-se por possuírem um altíssimo teor de Areia Muito Fina e pela inexistência de Argila; enquanto que as primeiras apresentam uma quantidade ligeiramente menor de Areia Muito Fina e uma proporção um pouco maior de Silte e Argila, embora ainda em valores relativamente pequenos. A proporção de Areia Total é muito alta para as quatro observações, sempre entre 96% e 99%, embora esse atributo seja utilizado como critério de classificação intermediário apenas para as camadas mais superficiais.

A SGH V possui suas amostras distribuídas em duas folhas da árvore, uma correspondendo a dois casos e outra, a cinco. Em linhas gerais, os dois conjuntos caracterizam-se pela ausência de Areia Grossa, forte predomínio de Areia Total e pela presença de alguma Argila na composição do material. A grande diferença entre os conjuntos se dá na proporção de Areia Fina, que é menor nas duas amostras da primeira folha, correspondentes

a profundidades de 30 e 40cm, em relação às 5 observações agrupadas na segunda folha. Isso significa que em geral as amostras da região V possuem um predomínio de material arenoso fino a muito fino, variando apenas a proporção entre esses materiais.

Mais uma vez a região IV, que possui o maior número de amostras (21), deu origem também ao maior número de folhas da árvore: três, sendo uma com 10 casos e outras duas com 6 casos cada uma. As características gerais dessa SGH quanto à granulometria são a quase ausência de Areia Grossa e um predomínio de material arenoso, entre fino e muito fino, com o Silte aparecendo regularmente, em pequenas quantidades. A folha com 10 casos destaca-se das demais por possuir quantidades muito pequenas de Argila (no máximo 1.2%, embora na maioria dos casos, esse valor seja zero). As duas outras folhas possuem características semelhantes entre si, com teor de Argila um pouco maior, ainda que geralmente abaixo da média geral, diferindo apenas nas proporções relativas de Areia Fina e Areia Muito Fina, a exemplo do que ocorreu com as amostras da região V.

Regras de Decisão

Neste experimento, o conjunto de regras de decisão gerado apresentou uma taxa de erros ligeiramente menor que a da árvore de decisão, já que foram verificados apenas 2 erros, que representam 3% dos casos. Essa diferença na taxa de erro pode ocorrer basicamente em decorrência da independência entre as regras, que afeta a forma como a ferramenta avalia os erros observados. Neste experimento, mais especificamente, um caso da SGH IV é classificado pela árvore como pertencendo à SGH I, enquanto que no conjunto de regras, ele é classificado tanto na SGH I (regra 1) como na SGH IV (regra 7). Isso significa que enquanto na árvore um caso é exclusivo de uma folha, ao se gerar regras ele pode ser coberto por mais de uma definição, devido à independência entre elas. Assim, na contagem final dos erros do conjunto de regras somente são contabilizados os casos que não são corretamente classificados por nenhuma das regras.

Algumas SGHs foram definidas por mais de uma regra, a exemplo do que ocorreu com a árvore de decisão e a SGH III novamente teve sua caracterização prejudicada, já que apresenta apenas dois casos, sendo um deles classificado incorretamente pelo conjunto de regras.

A SGH I é definida por apresentar Areia Muito Fina maior que 89.3% e Argila superior a 0.3%, o que pode ser interpretado como material arenoso muito fino com a presença de alguma argila observável.

As regras 2 e 3 referem-se à SGH II, que possui apenas um ponto amostrado em quatro profundidades diferentes. A regra 2 refere-se às amostras coletadas até a profundidade de 20cm e indicam, em linhas gerais, material arenoso muito fino com presença de Silte e Argila em pequenas quantidades. A regra 3 classifica as amostras coletadas entre 30 e 40cm para as quais a Argila é inexistente e o teor de Areia Muito Fina é ainda maior que nas camadas mais superficiais.

A regra 5 corresponde exatamente à descrição da folha com 10 casos da SGH IV discutida anteriormente, cuja principal característica em relação aos outros casos da mesma SGH é a quantidade de Argila, muito pequena ou inexistente. A regra 6 especifica que uma parcela das amostras da região IV possui Areia Total acima da média geral, quase nenhuma Areia Grossa e Areia Muito Fina mais ou menos em torno da média geral. A regra 7 indica que uma outra parcela das amostras possui uma proporção um pouco menor de Areia Fina (o

que indica um percentual maior de Areia Muito Fina em relação aos casos classificados pela regra 6).

A SGH V é tratada pelas regras 8 (que cobre 5 casos) e 9 (que cobre 2 casos, referentes a amostras coletadas a profundidades de 30 e 40cm). A primeira regra define o material como possuindo alguma Argila, Areia Fina ligeiramente acima da média e um predomínio significativo de Areia Muito Fina. Resumindo, trata-se de material arenoso fino e muito fino, com traços de Argila. A regra 9, por sua vez, indica altíssimas proporções de Areia Total, pouco ou nenhum Silte e uma proporção ligeiramente superior de Areia Muito Fina em relação ao restante do material desta SGH.

A regra 10 classifica o material da SGH VI, e o caracteriza como possuindo uma proporção razoável de Areia Grossa, próxima da média geral, e Areia Muito Fina acima de 13%. Se considerarmos que a média geral da Areia Muito Fina é de 53%, podemos esperar que um valor de corte assim tão baixo represente uma proporção relativamente pequena desse componente em relação ao restante das amostras. No caso, verificamos valores entre 31% e 50% desse material, o que permite dizer que nessas amostras o teor de Areia Muito Fina está claramente abaixo da média geral. Entretanto, a regra não menciona (e nem ao menos sugere) que há também uma forte presença de Silte, às custas de uma menor quantidade de Areia Total, o que talvez fosse uma informação mais significativa em termos práticos.

A região VII é definida pela regra número 11 como possuindo Areia Total acima de 62% (para uma média geral de 85%, o que não é muito esclarecedor), teores de Areia Fina abaixo da média geral e proporções muito reduzidas de Areia Muito Fina (abaixo de 14%, quando a média geral está em torno de 53%). Na realidade, observando-se o banco de dados, percebemos que trata-se de um conjunto de amostras bastante homogêneo, que retrata um terreno de composição variada, com predomínio de material arenoso grosso e muito grosso, com presença de razoáveis quantidades de Argila e Silte, o que fica levemente sugerido pela análise da regra de decisão.

Quantidade significativa de Areia Fina acompanhada de pouca Areia Muito Fina (inferiores a 14% quando a média geral desse tipo de material é 53%) indicam amostras referentes à SGH VIII, segundo a regra 12. Uma característica que não fica evidente na análise da regra é que trata-se de um terreno arenoso, com proporções mais ou menos próximas de Areia Grossa, Areia Média e Areia Fina e quase nenhuma Argila.

Proporção pequena de Areia Total (inferior a 63%, para uma média de 86%) e pouca Areia Muito Fina (abaixo de 14%) indicam amostras que devem ser classificadas pela regra 13 como pertencentes à SGH X (que, após uma rápida inspeção sobre o banco de dados, revelou-se como sendo a menos arenosa e mais argilosa de todas).

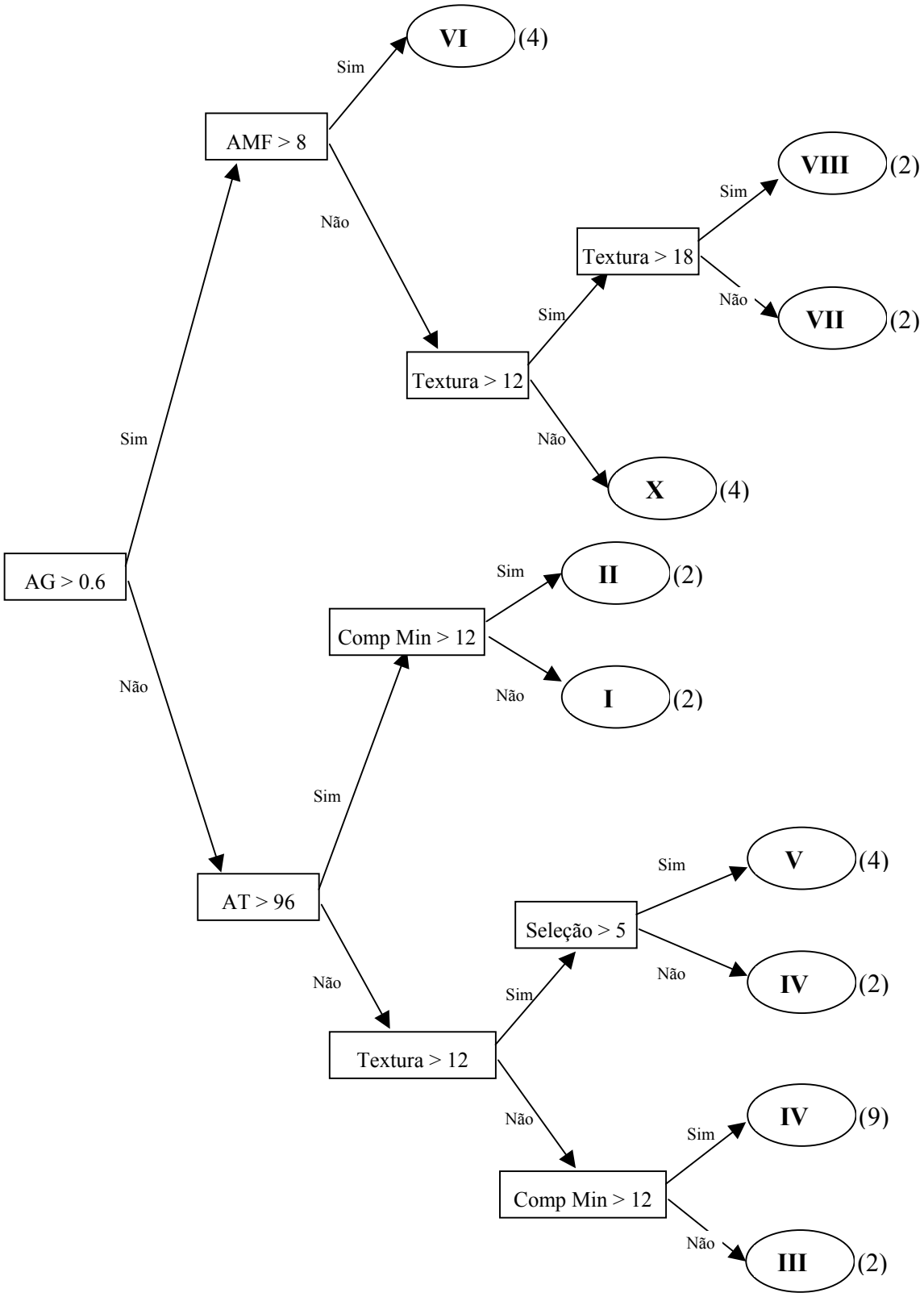
Interpretação dos resultados

A quantidade presente de Areia Grossa indica que existem dois grupos de SGHs principais: um com forte presença de material arenoso mais grosseiro e quantidades de Argila e Silte expressivamente acima da média geral das amostras, representado pelas regiões VI (denominada "Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos"), que correspondem a regiões relativamente antigas, com terrenos de composição mais diversificada; e outro, representado pelas SGHs I ("Planície de Maré"), II ("Berma"),

III (“Duna”), IV (“Planície Litorânea de Cordões Regressivos”) e V (“Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes”), onde há um forte predomínio de areias finas e muito finas e uma proporção relativamente pequena de Silte e Argila. A região III apresenta-se com contornos pouco precisos quanto ao aspecto granulométrico, em princípio por contar com um número muito reduzido de amostras, e não pode ser analisada com um mínimo de segurança. No geral observa-se também que não há muita diferença quanto ao aspecto granulométrico entre as SGHs I e II, que podem ser resumidas como possuindo material arenoso fino e muito fino, com presença de algum Silte e pouca Argila. Há também uma certa similaridade entre as SGHs IV e V, conforme indicam as suas descrições tanto via árvore de decisão como via conjunto de regras.

A diferença de profundidade entre os níveis de 10 a 40cm parece ter pouco efeito para as SGHs VI, VII, VIII e X quando comparamos suas respectivas definições nos classificadores gerados com amostras das profundidades de 10 a 20cm e 10 a 40cm. Nas SGHs restantes, entretanto, esse mesmo comportamento não ocorre, o que pode significar que uma parte da complexidade deste classificador seja devida a essa variação vertical.

4.1.3 Versão 1b (10-20cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 03 12:00:22 2000
Class specified by attribute `SGH'
Read 33 cases (16 attributes) from Exp1b_n20.data

Decision tree:
AG <= 0.6:
:...AT > 96:
:   :...CompMineral <= 12: I (2)
:   :   CompMineral > 12: II (2)
:   AT <= 96:
:   :...Textura <= 12:
:       :...CompMineral <= 12: III (2)
:       :   CompMineral > 12: IV (9)
:       Textura > 12:
:       :...Selecao <= 5: IV (2)
:           Selecao > 5: V (4)
AG > 0.6:
:...AMF > 8: VI (4)
    AMF <= 8:
    :...Textura <= 12: X (4)
        Textura > 12:
        :...Textura <= 18: VII (2)
            Textura > 18: VIII (2)


Evaluation on training data (33 cases):

           Decision Tree
         ----------------
         Size      Errors
           10    0( 0.0%)   <<

          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                  2                                                     (b): class II
                        2                                               (c): class III
                             11                                         (d): class IV
                                   4                                    (e): class V
                                         4                              (f): class VI
                                               2                        (g): class VII
                                                     2                  (h): class VIII
                                                                        (i): class IX
                                                                 4      (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 03 12:06:42 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 33 cases (16 attributes) from Exp1b_n20.data

Extracted rules:
Rule 1: (2, lift 12.4)
        CompMineral <= 12
        AT > 96
        ->  class I  [0.750]

Rule 2: (2, lift 12.4)
        CompMineral > 12
        AT > 96
        ->  class II  [0.750]

Rule 3: (2, lift 12.4)
        Textura <= 12
        CompMineral <= 12
        AT <= 96
        AG <= 0.6
        ->  class III  [0.750]

Rule 4: (9, lift 2.7)
        Textura <= 12
        CompMineral > 12
        AT <= 96
        ->  class IV  [0.909]

Rule 5: (2, lift 2.2)
        Selecao <= 5
        AT <= 96
        AG <= 0.6
        ->  class IV  [0.750]

Rule 6: (4, lift 6.9)
        Textura > 12
        Selecao > 5
        ->  class V  [0.833]

Rule 7: (4, lift 6.9)
        AG > 0.6
        AMF > 8
        ->  class VI  [0.833]

Rule 8: (2, lift 12.4)
        Textura > 12
        Textura <= 18
        AMF <= 8
        ->  class VII  [0.750]
```

```
Rule 9: (2, lift 12.4)
        Textura > 18
        ->  class VIII  [0.750]

Rule 10: (4, lift 6.9)
        Textura <= 12
        AMF <= 8
        ->  class X  [0.833]

Default class: IV


Evaluation on training data (33 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
           10    0( 0.0%)      10    0( 0.0%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----    ----------------
            2                                                          (a): class I
                  2                                                    (b): class II
                        2                                              (c): class III
                             11                                        (d): class IV
                                   4                                   (e): class V
                                         4                             (f): class VI
                                               2                       (g): class VII
                                                     2                 (h): class VIII
                                                                       (i): class IX
                                                           4           (j): class X
```

Análise dos resultados

Este experimento utiliza dados da análise granulométrica combinados com alguns atributos da análise macroscópica, nas profundidades de 10 a 20cm, perfazendo um total de 33 amostras (ou casos) com 16 atributos em cada uma, dos quais 12 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Todos os casos foram classificados corretamente pela árvore. De saída foram induzidos dois grandes subconjuntos de SGHs, um deles composto por 4 regiões (VI, VII, VIII e X) e o outro correspondendo às 5 restantes, cuja principal diferença reside na quantidade de Areia Grossa observada.

Na primeira subárvore, a principal característica é a quantidade muito pequena de Areia Muito Fina, inferior a 8% (para uma média geral de 54%) nas regiões VII, VIII e X, e entre 30% e 50% na região VI. A Textura do material é um outro discriminador adotado: para a SGH X, varia de "argiloso com forte agregação" até "muito fina"; para a SGH VII varia de "muito fina" a "fina"; e para a SGH VIII, é acima de "média" (observando no banco de dados, descobrimos que a textura desse material é definida como "grossa"), um resultado bastante previsível, dados os altos teores de Areia Grossa e Média observados para as amostras dessa SGH.

A segunda subárvore, por sua vez, possui um predomínio de material arenoso (nunca inferior a 75% e na maioria das vezes, superior a 90%) geralmente fino ou muito fino. As SGHs I e II possuem em comum uma percentagem muito alta de Areia Total e diferem basicamente quanto à Composição Mineral: na primeira o material é composto predominantemente por quartzo com a presença subordinada de mica e suas variedades; enquanto que na segunda, a Composição Mineral é apenas de quartzo.

A SGH III caracteriza-se por possuir uma Textura "muito fina" com Composição Mineral onde o quartzo aparece predominando, com traços de biotita e muscovita. A SGH V, por sua vez, é apresentada na árvore de decisão como sendo composta predominantemente por material arenoso, possuidor de uma Textura entre "fina" e "muito fina" e "muito bem selecionado" (atributo Grau de Seleção > 5).

Finalmente, a SGH IV origina duas folhas na árvore, uma com 2 casos (provavelmente obtidos em um mesmo ponto) e outra com os 9 restantes. A característica comum entre esses subconjuntos é a quase ausência de Areia Grossa e também a ocorrência de proporções elevadas de Areia Total, ainda que nunca superiores a 96%. A principal diferença se dá com relação à Textura observada: o subconjunto menor apresenta Textura "fina" (com um Grau de Seleção "bem selecionado") enquanto que o restante apresenta Textura "muito fina" e quartzo (sem traços de mica, argila ou feldspato) como Composição Mineral.

Regras de Decisão

O conjunto de regras de decisão gerado também classificou corretamente todos os casos, e acompanha em suas linhas gerais os critérios estabelecidos na árvore de decisão, sendo que apenas a SGH IV possui duas regras para descrevê-la.

A regra 1 especifica que as amostras da região I possuem taxa de Areia Total superior a 96% e Composição Mineral de quartzo onde aparece alguma mica e quase nenhum feldspato.

A SGH II por sua vez possui Areia Total superior a 96% e, ao contrário da SGH I, um predomínio de quartzo (puro, sem feldspato, mica ou argila) em sua Composição Mineral.

A SGH III é descrita na regra III exatamente da mesma forma que na árvore de decisão correspondente: ausência de Areia Grossa; percentual de Areia Total próximo da média geral; Textura "muito fina"; Composição Mineral onde o quartzo aparece predominando, com traços de biotita e muscovita. Não é uma descrição muito significativa, na medida em que requer uma combinação relativamente grande de condições (4 testes) e nenhuma particularidade mais interessante foi capturada, além do que esta regra baseia-se em apenas 2 casos.

A regra 4 trata da maioria das amostras da SGH IV, caracterizadas por uma Composição Mineral de quartzo puro, Textura "muito fina" e Areia Total abaixo de 96%, mas ainda assim em quantidade acima da média geral desse atributo. A regra 5 define que as 2 amostras não cobertas pela regra anterior possuem apenas traços de Areia Grossa, grau de seleção "bem selecionado" e grandes quantidades de Areia Total, mesmo que abaixo de 96%.

A regra 6 estabelece que a SGH V pode ser caracterizada apenas em função de uma Textura "fina" a "muito fina" e um Grau de Seleção "muito bem selecionado".

A SGH VI é descrita vagamente pela regra 7 como possuindo Areia Grossa em percentuais superiores a 0.6% (o que não é muito esclarecedor, visto que a média geral desse atributo é 6%) e Areia Muito Fina superior a 8% (para uma média geral de 54%).

A SGH VII é caracterizada pela regra 8 como possuindo Textura entre "muito fina" e "fina à média" e percentuais muito pequenos de Areia Muito Fina.

Amostras com Textura de "média" a "muito grossa" são características da SGH VIII, conforme a regra 9.

A regra 10 estabelece que Textura entre "argiloso, com forte agregação da argila" e "muito fina", associadas com percentuais muito pequenos de Areia Muito Fina (menores que 8% quando a média geral desse atributo é 54%) indicam amostras da SGH X.

Interpretação dos resultados

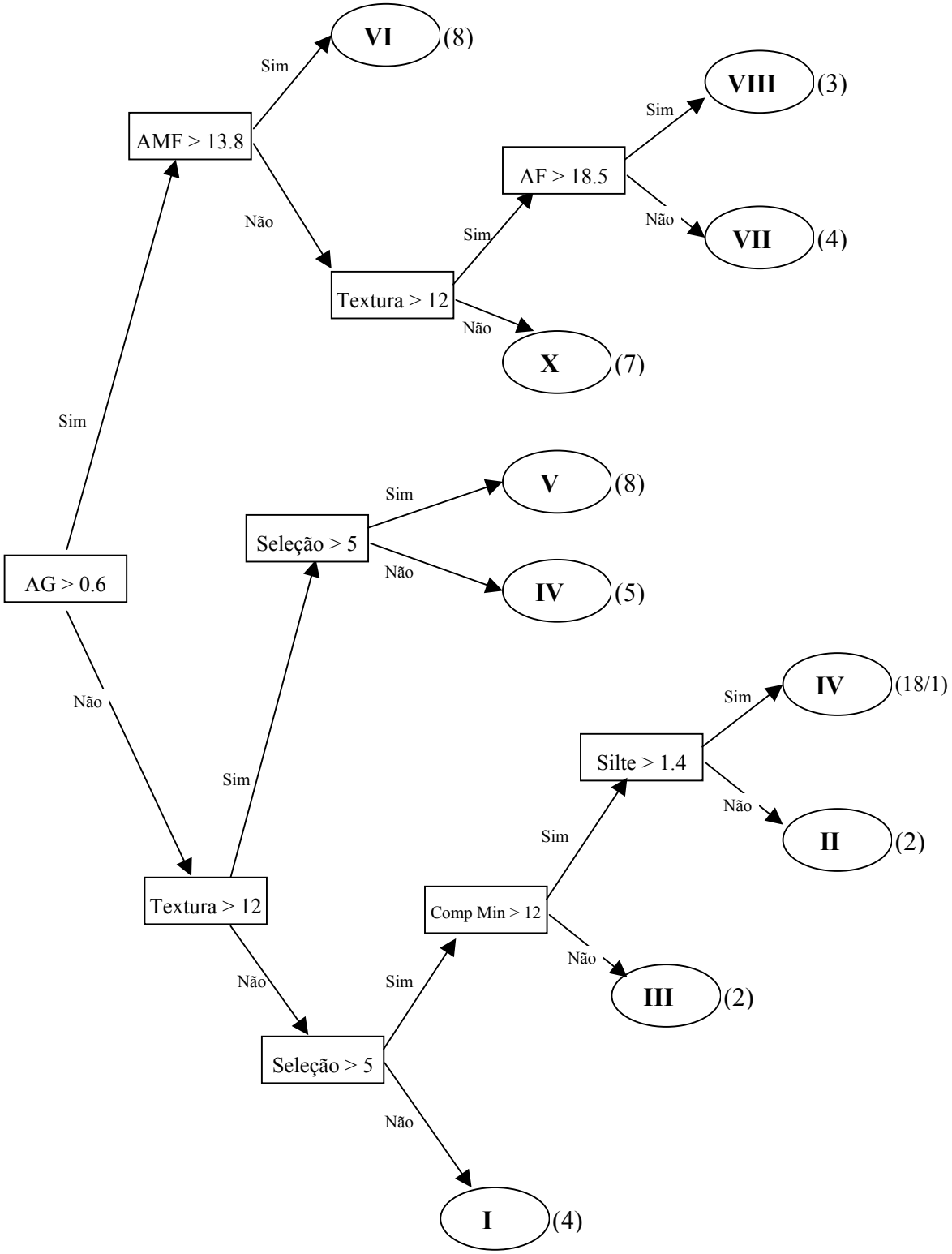
Com base na árvore de decisão, fica bastante clara a existência de três grandes grupos de SGHs. O primeiro deles é formado por terrenos mais diversificados, com menor proporção de material arenoso, geralmente mais grosseiro, e presença significativa de argila e silte e composto pelas regiões VI (denominada "Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos"). A importância dos atributos Areia Muito Fina e, principalmente, Textura para a subdivisão desse conjunto é bastante coerente com esse perfil, assim como o conjunto de regras induzido para cada classe, simples e de fácil interpretação.

O segundo grupo é formado pelas SGHs I e II, respectivamente “Planície de Maré” e “Berma”, e muito bem caracterizado pela ocorrência de terrenos quase que totalmente arenosos (acima de 96%) muito finos, predominantemente de quartzo.

Por fim, o terceiro grupo é também tipicamente arenoso, variando de fino a muito fino, subdividindo-se secundariamente com base no Grau de Seleção e Composição Mineral. Esse grupo é integrado pelas SGHs III (“Duna”), IV (“Planície Litorânea de Cordões Regressivos”) e V (“Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes”), sendo que a primeira delas é muito mal caracterizada tanto pela árvore de decisão como pelo conjunto de regras, em parte devido ao seu reduzido número de amostras, mas também parecendo significar que os aspectos granulométricos e macroscópicos utilizados no experimento talvez não sejam suficientes (ou adequados) para sua correta individualização.

De uma maneira geral, observa-se que os atributos originários da análise macroscópica parecem desempenhar um papel de simplificação dos critérios de classificação em relação ao resultado obtido no experimento 1a, reduzindo a complexidade de algumas regras e diminuindo também o tamanho da árvore de decisão.

4.1.4 Versão 1b (10-40cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 03 11:31:35 2000
Class specified by attribute `SGH'
Read 61 cases (16 attributes) from Exp1b_n40.data

Decision tree:
AG > 0.6:
:...AMF > 13.8: VI (8)
:   AMF <= 13.8:
:   :...Textura <= 12: X (7)
:       Textura > 12:
:       :...AF <= 18.5: VII (4)
:           AF > 18.5: VIII (3)
AG <= 0.6:
:...Textura > 12:
    :...Selecao <= 5: IV (5)
    :   Selecao > 5: V (8)
    Textura <= 12:
    :...Selecao <= 5: I (4)
        Selecao > 5:
        :...CompMineral <= 12: III (2)
            CompMineral > 12:
            :...S <= 1.4: II (2)
                S > 1.4: IV (18/1)


Evaluation on training data (61 cases):

          Decision Tree
        ----------------
        Size      Errors
          10    1( 1.6%)   <<

         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           4                                                           (a): class I
                 2           1                                         (b): class II
                       2                                             (c): class III
                            22                                       (d): class IV
                                  8                                  (e): class V
                                        8                            (f): class VI
                                              4                      (g): class VII
                                                    3                (h): class VIII
                                                                     (i): class IX
                                                          7          (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 03 12:14:36 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 61 cases (16 attributes) from Exp1b_n40.data

Extracted rules:
Rule 1: (4, lift 12.7)
        Textura <= 12
        Selecao <= 5
        AG <= 0.6
        ->  class I  [0.833]

Rule 2: (2, lift 15.2)
        Textura <= 12
        Selecao > 5
        S <= 1.4
        ->  class II  [0.750]

Rule 3: (2, lift 22.9)
        Textura <= 12
        CompMineral <= 12
        Selecao > 5
        AG <= 0.6
        ->  class III  [0.750]

Rule 4: (18/1, lift 2.5)
        Textura <= 12
        CompMineral > 12
        S > 1.4
        ->  class IV  [0.900]

Rule 5: (5, lift 2.4)
        Textura > 12
        Selecao <= 5
        AG <= 0.6
        ->  class IV  [0.857]

Rule 6: (8, lift 6.9)
        Textura > 12
        Selecao > 5
        ->  class V  [0.900]

Rule 7: (8, lift 6.9)
        AG > 0.6
        AMF > 13.8
        ->  class VI  [0.900]

Rule 8: (4, lift 12.7)
        Textura > 12
        AF <= 18.5
        AMF <= 13.8
        ->  class VII  [0.833]
```

```
Rule 9: (3, lift 16.3)
        Textura > 12
        AF > 18.5
        AMF <= 13.8
        ->  class VIII  [0.800]

Rule 10: (7, lift 7.7)
        Textura <= 12
        AMF <= 13.8
        ->  class X  [0.889]

Default class: IV


Evaluation on training data (61 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
           10    1( 1.6%)      10    1( 1.6%)   <<

        (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
       ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
          4                                                           (a): class I
                2           1                                         (b): class II
                      2                                             (c): class III
                           22                                       (d): class IV
                                 8                                  (e): class V
                                       8                            (f): class VI
                                             4                      (g): class VII
                                                   3                (h): class VIII
                                                                    (i): class IX
                                                         7          (j): class X
```

Análise dos resultados

Este experimento utiliza dados da análise granulométrica combinados com alguns atributos da análise macroscópica, nas profundidades de 10 a 40cm, perfazendo um total de 61 amostras (ou casos) com 16 atributos em cada uma, dos quais 12 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Um caso foi classificado incorretamente pela árvore, o que representa 1.6% do total. De saída foram induzidos dois grandes subconjuntos de SGHs, um deles composto por 4 regiões (VI, VII, VIII e X) e o outro correspondendo às 5 restantes, cuja principal diferença reside na quantidade de Areia Grossa observada.

No primeiro subconjunto (ou subárvore) a região VI destaca-se das outras 3 por possuir o maior teor de Areia Muito Fina, ainda que em quantidades inferiores à média geral desse atributo. A região X destaca-se por possuir uma Textura mais fina (observando-se no banco de dados, verifica-se que varia de "argilosa com material endurecido" a "argilo-arenosa") e as duas SGHs restantes diferem basicamente quanto ao teor de Areia Fina, apresentando a região VII um material arenoso bem mais grosseiro que a região VIII.

O segundo subconjunto é composto por terrenos arenosos mais finos e, com base na Textura, é subdivido em duas partes. Na primeira delas estão as amostras de Textura entre fina e muito fina, que caso o Grau de Seleção seja "muito bem selecionado", indica material da SGH V, e caso seja "bem selecionado", indica SGH IV. A segunda parte dessa subárvore procura classificar amostras das SGHs I, II, III e IV, que tipicamente possuem uma Textura mais fina que a anterior. Se o material deste subconjunto não apresentar Grau de Seleção "muito bem selecionado", ele pertence à região I. À região III correspondem as amostras onde a Composição Mineral é quartzo com traços de muscovita e biotita. As amostras cuja Composição Mineral é tipicamente quartzo puro correspondem às SGHs II e IV, sendo que nesta última a quantidade verificada de Silte é superior à primeira.

Regras de Decisão

O conjunto de regras de decisão gerado também classificou incorretamente um dos casos, e acompanha em suas linhas gerais os critérios estabelecidos na árvore de decisão, sendo que apenas a SGH IV possui duas regras para descrevê-la.

Conforme a regra 1, a SGH I possui uma Textura "muito fina", um Grau de Seleção "bem selecionado" e quantidades mínimas de Areia Grossa.

A SGH II é definida na regra 2, que estipula como critério de classificação uma Textura "muito fina", um Grau de Seleção "muito bem selecionado" e alguma presença de Silte.

A regra 3 é a mais complexa do conjunto, e caracteriza o material da SGH III como possuindo Textura muito fina (no caso, é "muito fina argilosa"); Composição Mineral onde o quartzo predominante possui traços de mica e nenhum feldspato; Grau de Seleção "muito bem selecionado" e ausência de Areia Grossa.

As SGH IV é tratada pelas regras 4 e 5, que referem-se respectivamente a 18 e 5 amostras. O maior desses conjuntos é caracterizado por Texturas muito finas, material quartzoso com ausência de micas em sua Composição Mineral e presença de pequena quantidade de Silte.

A regra 5 define Textura finas e médias, Grau de Seleção "bem selecionado" e quantidades muito pequenas de Areia Grossa para os casos restantes dessa SGH.

A regra 6 estabelece que a SGH V pode ser caracterizada apenas em função de uma Textura "fina" a "muito fina" e um Grau de Seleção "muito bem selecionado".

A SGH VI é descrita vagamente pela regra 7 como possuindo Areia Grossa em percentuais superiores a 0.6% (o que não é muito esclarecedor, visto que a média geral desse atributo é 6%) e Areia Muito Fina superior a 13% (para uma média geral de 54%).

Texturas finas e médias e quantidades pequenas de Areia Fina e de Areia Muito Fina são típicas da SGH VII, conforme a regra 8.

A regra 9 classifica as amostras da SGH VIII como possuindo Texturas finas e médias, muito pouca Areia Muito Fina e quantidades de Areia Fina bastante acima da média geral (entre 27% e 42%, para a média de 16%), refletindo um material arenoso mais grosseiro.

Finalmente, a regra 10 estabelece que Textura entre "argiloso, com forte agregação da argila" e "muito fina", associadas com percentuais muito pequenos de Areia Muito Fina (menores que 8% quando a média geral desse atributo é 54%) indicam amostras da SGH X.

Interpretação dos resultados

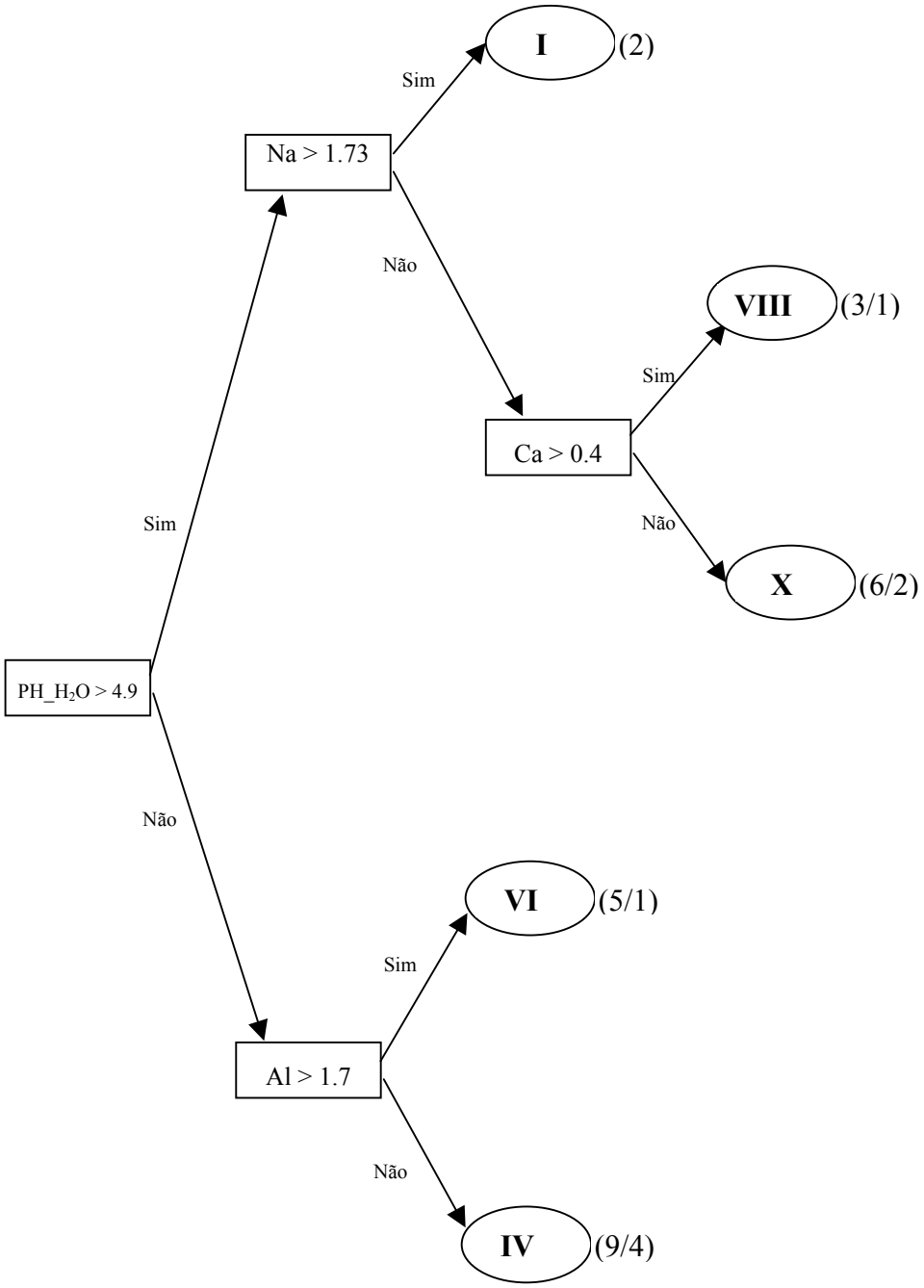
Com base na árvore de decisão, fica bastante clara a existência de dois grandes grupos de SGHs. O primeiro deles é formado por terrenos mais diversificados, com menor proporção de material arenoso, geralmente mais grosseiro, e presença significativa de argila e silte e composto pelas regiões VI (denominada "Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos"). A importância dos atributos Areia Muito Fina e Textura para a subdivisão desse conjunto é coerente com esse perfil, assim como o conjunto de regras induzido para cada classe, simples e de fácil interpretação.

O segundo grupo é formado pelas SGHs I ("Planície de Maré"), II ("Berma"), III ("Duna"), IV ("Planície Litorânea de Cordões Regressivos") e V ("Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes"), onde há um predomínio de terreno arenoso mais fino com Texturas finas e muito finas e, em geral, material bem selecionado. As caracterizações das regiões III não ficaram muito claras, na medida em que são produto de regras relativamente complexas e que revelam muito pouca especificidade para o material analisado, assim como há uma razoável sobreposição nas definições entre as SGHs II, IV e V.

Uma comparação entre os resultados das versões 1a e 1b indica novamente que os dados da análise macroscópica simplificam os classificadores obtidos, principalmente no que se refere às SGHs I, II, III, IV e V, que possuem uma maior homogeneidade granulométrica. Com relação à variação vertical das amostras, as classificações geradas para as SGHs I, II e VIII são bastante diferentes conforme se considera os níveis de profundidade entre 10 e 20cm e entre 10 e 40cm, permanecendo sem maiores alterações nos demais casos, ou seja, aparentemente a variação na composição do terreno nos diversos níveis não é igual em todas as SGHs.

4.2 Experimento 2

4.2.1 Versão 2a (10-20cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 03 21:18:34 2000
Class specified by attribute `SGH'
Read 25 cases (13 attributes) from Exp2a.data

Decision tree:
Ph_H2O <= 4.9:
:...Al <= 1.7: IV (9/4)
:   Al > 1.7: VI (5/1)
Ph_H2O > 4.9:
:...Na > 1.73: I (2)
    Na <= 1.73:
    :...Ca <= 0.4: X (6/2)
        Ca > 0.4: VIII (3/1)


Evaluation on training data (25 cases):

           Decision Tree
         ----------------
         Size      Errors
            5    8(32.0%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                              2                                         (b): class II
                                                    1           1       (c): class III
                              5                                         (d): class IV
                              1                                   1     (e): class V
                                          4                             (f): class VI
                              1           1                             (g): class VII
                                                    2                   (h): class VIII
                                                                        (i): class IX
                                                                  4     (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 03 21:45:18 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 25 cases (13 attributes) from Exp2a.data

Extracted rules:
Rule 1: (2, lift 9.4)
        Na > 1.73
        Ph_H2O > 4.9
        ->  class I  [0.750]

Rule 2: (9/4, lift 2.7)
        Al <= 1.7
        Ph_H2O <= 4.9
        ->  class IV  [0.545]

Rule 3: (5/1, lift 4.5)
        Al > 1.7
        Ph_H2O <= 4.9
        ->  class VI  [0.714]

Rule 4: (3/1, lift 7.5)
        Ca > 0.4
        Ph_H2O > 4.9
        ->  class VIII  [0.600]

Rule 5: (6/2, lift 3.9)
        Ca <= 0.4
        Na <= 1.73
        Ph_H2O > 4.9
        ->  class X  [0.625]

Default class: IV

Evaluation on training data (25 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
            5    8(32.0%)       5    8(32.0%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                              2                                         (b): class II
                                                    1           1       (c): class III
                              5                                         (d): class IV
                              1                                   1     (e): class V
                                          4                             (f): class VI
                              1           1                             (g): class VII
                                                    2                   (h): class VIII
                                                                        (i): class IX
                                                                  4     (j): class X
```

Análise dos resultados

Este experimento utiliza dados da análise química, nas profundidades de 10 a 20cm, perfazendo um total de 25 amostras (ou casos) com 13 atributos em cada uma, dos quais 12 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

A árvore induzida, de tamanho bastante reduzido, apresentou uma taxa de erros de 32% (8 casos foram classificados incorretamente), muito elevada para fins analíticos. Apenas as SGHs I, IV, VI, VIII e X possuem suas respectivas folhas, ainda que em algumas delas a taxa de erro chegue próximo de 50%. As amostras da região II aparecem juntamente com a SGH IV, o que poderia indicar que possuem um perfil químico semelhante. As amostras da SGH III, V, VI e VII aparecem dispersas por mais de uma folha, inviabilizando uma análise mais detalhada, dado o pequeno volume de amostras de cada uma. O principal critério de classificação utilizado na árvore é o Ph da Água, que gera um grupo onde esse indicador está acima da média geral e outro onde está abaixo dessa média.

A única folha que não apresenta erro é a correspondente à SGH I, caracterizada pelos maiores índices de Ph da Água dentre todas as amostras e pela grande quantidade de Sódio, bastante superior à média geral desse atributo.

As amostras com alto Ph da Água e menores quantidades de Sódio correspondem, em princípio, às SGHs VIII e X que distinguem-se levemente entre si pela quantidade de Cálcio verificada em sua composição. A SGH VIII possui esse componente em taxas ligeiramente acima da média, enquanto que na SGH X esses valores são um pouco menores.

Finalmente, as amostras onde o Ph da Água é mais reduzido são subdivididas com base no teor de Alumínio: as maiores concentrações, claramente acima da média desse componente são atribuídas à região VI e as amostras com menores concentrações são atribuídas à região IV. Nessa última folha ocorre a maior taxa de erro, pois apenas 5 das 9 amostras classificadas como classe IV são de fato dessa SGH, sendo 2 das amostras restantes pertencentes à SGH II.

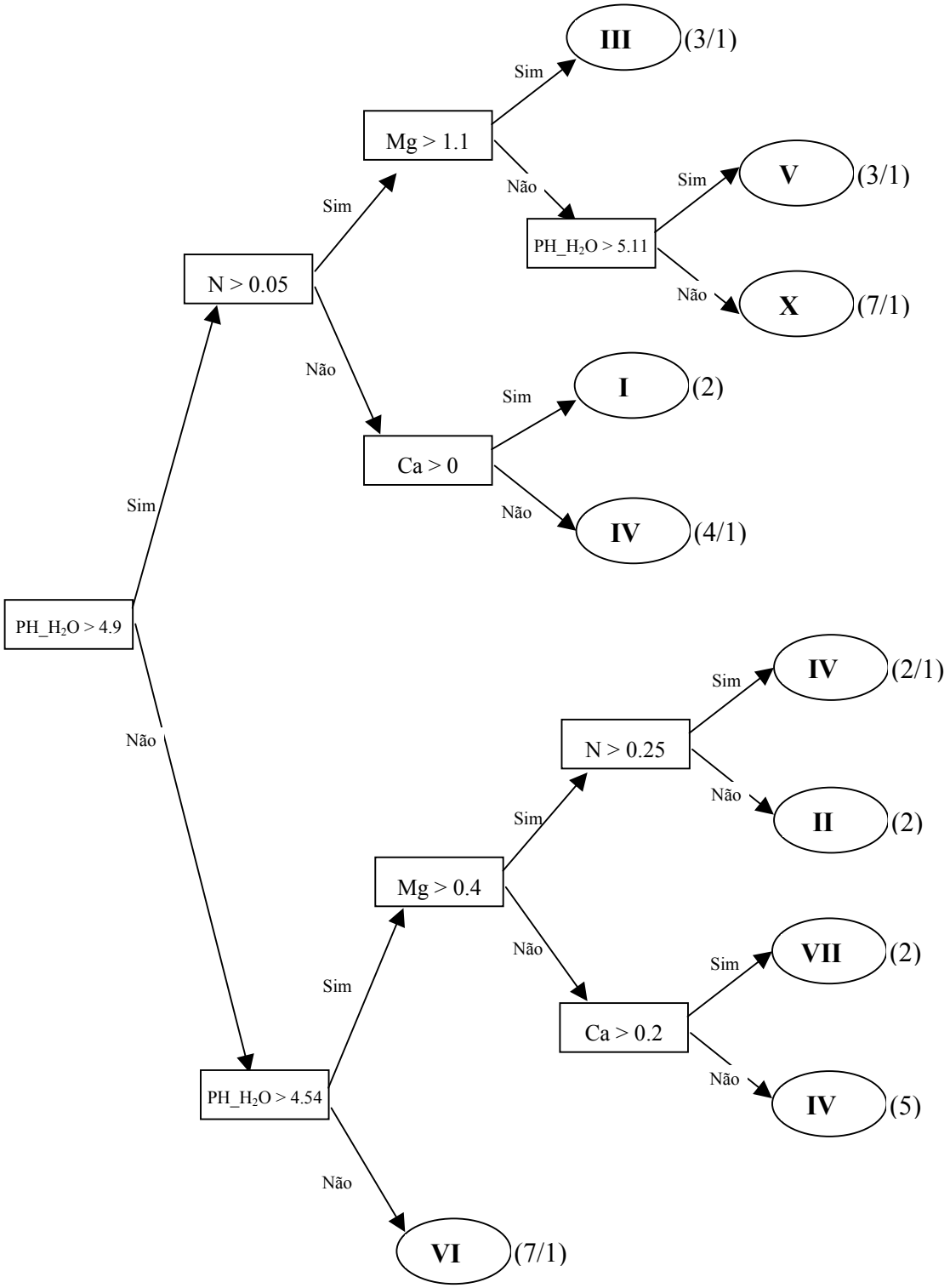
Regras de Decisão

O conjunto de regras de decisão gerado também classificou incorretamente oito casos, e acompanha quase que exatamente os critérios estabelecidos na árvore de decisão, sendo que apenas para a SGH VIII há uma pequena simplificação, representada pelo fato de que o teor de Sódio não é considerado, ao contrário do que ocorreu com a árvore correspondente. Assim, uma amostra é classificada pela regra 4 como pertencendo à SGH VIII se apresentar alto Ph da Água e teores de Cálcio acima da média geral das amostras.

Interpretação dos resultados

O resultado final foi claramente prejudicado pelo pequeno volume de amostras, e qualquer julgamento mais apurado que se faça é de validade duvidosa, dadas as altas taxas de erro verificadas sobre o conjunto de treinamento. Pode-se aceitar como significativo o papel do Ph da Água como elemento divisor entre alguns tipos de terreno e, secundariamente, talvez a importância do Sódio para caracterizar a SGH I ("Planície de Maré") e do Alumínio para a SGH VI ("Planície de Retrabalhamento Fluvial-Marinho"). O fato de as 2 amostras da SGH II ("Berma") terem sido classificadas como terrenos de baixo Ph da Água e pequena concentração de Alumínio juntamente com as 5 amostras da SGH IV ("Planície Litorânea de Cordões Regressivos") pode eventualmente merecer alguma atenção, mas em princípio não parece ser relevante.

4.2.2 Versão 2a (10-40cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 03 21:22:13 2000
Class specified by attribute `SGH'
Read 37 cases (13 attributes) from Exp2a.data

Decision tree:
Ph_H2O <= 4.9:
:...Ph_H2O <= 4.54: VI (7/1)
:   Ph_H2O > 4.54:
:   :...Mg <= 0.4:
:       :...Ca <= 0.2: IV (5)
:       :   Ca > 0.2: VII (2)
:       Mg > 0.4:
:       :...N <= 0.25: II (2)
:           N > 0.25: IV (2/1)
Ph_H2O > 4.9:
:...N <= 0.05:
    :...Ca <= 0: IV (4/1)
    :   Ca > 0: I (2)
    N > 0.05:
    :...Mg > 1.1: III (3/1)
        Mg <= 1.1:
        :...Ph_H2O <= 5.11: X (7/1)
            Ph_H2O > 5.11: V (3/1)


Evaluation on training data (37 cases):

          Decision Tree
        ----------------
        Size      Errors
          10    6(16.2%)   <<


       (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
      ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
        2                                                           (a): class I
              2                                                     (b): class II
                    2     1                                         (c): class III
                          9           1                           (d): class IV
                          1     2                                 (e): class V
                                      6                           (f): class VI
                                            2               1     (g): class VII
                    1           1                                 (h): class VIII
                                                                  (i): class IX
                                                              6   (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 03 21:26:09 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 37 cases (13 attributes) from Exp2a.data

Extracted rules:
Rule 1: (2, lift 13.9)
        N <= 0.05
        Ca > 0
        ->  class I  [0.750]

Rule 2: (2, lift 13.9)
        N <= 0.25
        Mg > 0.4
        Ph_H2O > 4.54
        Ph_H2O <= 4.9
        ->  class II  [0.750]

Rule 3: (3/1, lift 7.4)
        Mg > 1.1
        Ph_H2O > 4.9
        ->  class III  [0.600]

Rule 4: (4/1, lift 2.5)
        N <= 0.05
        Ca <= 0
        ->  class IV  [0.667]

Rule 5: (11/5, lift 2.0)
        Ph_H2O > 4.54
        Ph_H2O <= 4.9
        ->  class IV  [0.538]

Rule 6: (3/1, lift 7.4)
        N > 0.05
        Mg <= 1.1
        Ph_H2O > 5.11
        ->  class V  [0.600]

Rule 7: (7/1, lift 4.8)
        Ph_H2O <= 4.54
        ->  class VI  [0.778]

Rule 8: (2, lift 9.3)
        Ca > 0.2
        Mg <= 0.4
        Ph_H2O <= 4.9
        ->  class VII  [0.750]
```

```
Rule 9: (7/1, lift 4.8)
        N > 0.05
        Mg <= 1.1
        Ph_H2O > 4.9
        Ph_H2O <= 5.11
        ->  class X  [0.778]

Default class: IV

Evaluation on training data (37 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
           10    6(16.2%)       9    6(16.2%)    <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                  2                                                     (b): class II
                        2     1                                         (c): class III
                              9     1                                   (d): class IV
                              1     2                                   (e): class V
                                    6                                   (f): class VI
                                          2                       1     (g): class VII
                        1           1                                   (h): class VIII
                                                                        (i): class IX
                                                                  6     (j): class X
```

Análise dos resultados

Este experimento utiliza dados da análise química, nas profundidades de 10 a 40cm, perfazendo um total de 37 amostras (ou casos) com 13 atributos em cada uma, dos quais 12 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Foram computados 6 erros de classificação no total, o que representa uma taxa de cerca de 16% dos casos, relativamente alta mas significativamente melhor que a verificada apenas com as amostras de 10 a 20cm. Apesar disso, uma folha apresenta uma taxa de erro de 50% (2 casos e um deles foi classificado incorretamente) e as SGHs III, V, VII e VIII tiveram algumas de suas amostras dispersas em mais de uma folha, o que não permite que sejam analisadas com maior segurança, dada a pequena quantidade de amostras de cada uma. Além disso, ocorre de o valor referente a Nitrogênio faltar para a maioria das amostras da SGH V, o que inviabiliza ainda mais a consideração dessas amostras, na medida em que o teor desse elemento desempenha um papel importante dentro do critério de classificação induzido. Observa-se ainda que o elemento mais importante para a classificação foi o Ph da Água, que deu origem a dois grupos de amostras, um com valores acima da média geral desse atributo e outro com valores mais baixos.

O grupo das amostras com maior Ph da Água é subdividido com base no teor de Nitrogênio e depois, segundo outros elementos secundários. Os terrenos com muito pouco Nitrogênio e alguma ocorrência de Cálcio pertencem à SGH I e aqueles onde não há presença desse último elemento, são classificados como SGH IV. As amostras com maior taxa de Nitrogênio e Ph da Água mais alto pertencem à região X.

No segundo grupo, as amostras com Ph da Água mais baixo entre todas pertencem à SGH VI enquanto que aquelas que possuem o valor desse atributo apenas ligeiramente abaixo de sua média geral são depois subdivididas com base no teor de Magnésio. No subconjunto de menor teor de Magnésio e quantidade muito pequena de Cálcio estão as amostras da SGH IV; maiores quantidades de Magnésio e volumes de Nitrogênio razoavelmente próximos da média geral são característicos das amostras da SGH II.

Regras de Decisão

O conjunto de regras de decisão gerado também classificou incorretamente 6 casos, apresentando as mesmas características da árvore de decisão quanto à distribuição dos erros.

A regra 1 define as amostras da SGH I como possuindo quantidades de Nitrogênio muito pequenas (inferiores a 0.05% quando a média desse atributo é 0.18%) e pela presença de Cálcio, em qualquer quantidade.

As amostras da SGH II são descritas como possuindo um Ph da Água entre 4.54 e 4.9 (portanto, ligeiramente abaixo da média de 4.89), Magnésio acima da média geral e quantidades médias de Nitrogênio.
As regras 4 e 5 descrevem as amostras da SGH IV e apresentam um fator de confiança pequeno, entre 0.5 e 0.7, devido a grande volume de erros que apresentam: dos 15 casos classificados por essas regras, 6 não pertencem à região IV. Entretanto, se considerarmos que 9 dos 10 casos dessa SGH encaixam-se nas definições dessas regras, é possível que elas sejam de alguma utilidade prática para delinear algumas características químicas gerais desse tipo de terreno. A regra 4 estipula níveis muito baixos de Nitrogênio e ausência de Cálcio, e classifica corretamente 3 casos da SGH; e a regra 5 baseia-se apenas no Ph da Água, que encontra-se entre 4.54 e 4.9 (ou seja, ligeiramente abaixo da média geral), para identificar os 6 casos restantes da região IV.

Conforme indicado na regra 7, os terrenos da SGH VI caracterizam-se por apresentar o menor Ph da Água dentre todas as regiões analisadas.

A regra 8 estipula que as amostras da SGH VII possuem Ph da Água relativamente baixo, concentrações muito pequenas de Magnésio e teores de Cálcio acima da média geral, mas tem sua confiança prejudicada pelo fato de apenas 2 dos 3 casos dessa região se enquadrarem nessa definição.

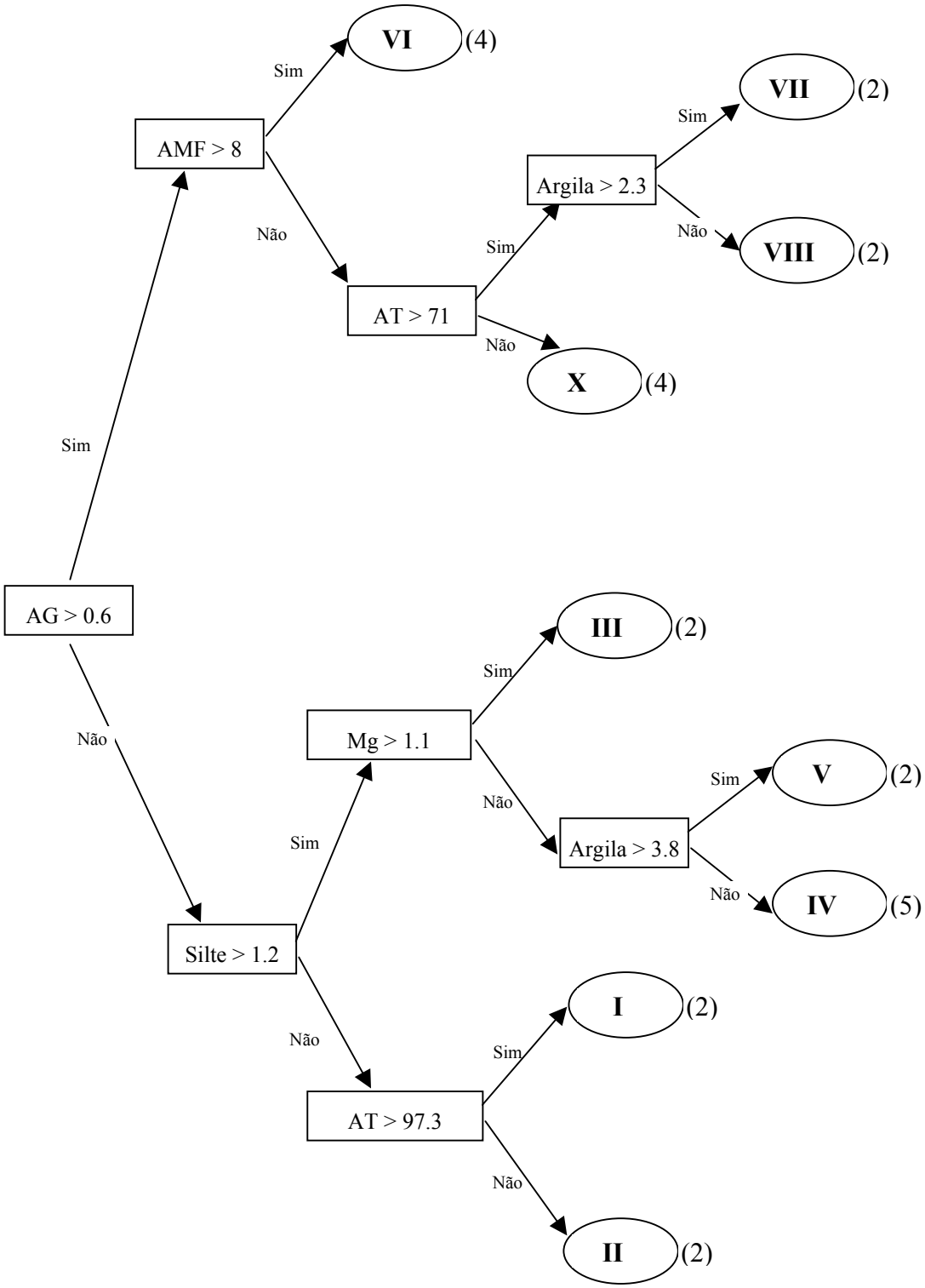
Finalmente, a regra 9 indica que pertencem à SGH X os terrenos com Ph da Água ligeiramente acima da média geral (entre 4.9 e 5.11, para uma média de 4.89), quantidades de Nitrogênio próximas da média geral e teores de Magnésio razoáveis, ainda que nunca superiores a 1.1.

Interpretação dos resultados

Mesmo com um aumento no número de casos utilizados no experimento, as taxas de erro ainda são relativamente altas. A exemplo do que ocorreu quando se considerou apenas as amostras coletadas mais próximas à superfície, o principal critério de classificação adotado foi o Ph da Água, cuja faixa de valores verificados foi dividida em 4 segmentos que foram referenciados pela maioria das regras, tendo ficado claro que a SGH VI (denominada "Planície de Retrabalhamento Fluvial-Marinho") é a que apresenta os menores valores. Enquanto que na análise das amostras até 20cm de profundidade os elementos químicos secundários utilizados foram o Sódio, o Alumínio e o Cálcio, quando se consideram também as amostras até 40cm são importantes, segundo o classificador induzido pelo See5, apenas os elementos Cálcio, Nitrogênio e Magnésio.

Outras características que pode-se deduzir são que a SGH IV ("Planície Litorânea de Cordões Regressivos") tipicamente apresenta muito pouco Cálcio e que a SGH VII ("Planície Colúvio-Aluvionar") destaca-se pelo pouco Magnésio.

4.2.3 Versão 2b (10-20cm)



Árvore

```
See5 [Release 1.12]     Sat Nov 04 11:03:47 2000
Class specified by attribute `SGH'
Read 25 cases (21 attributes) from Exp2b.data

Decision tree:
AG <= 0.6:
:...S <= 1.2:
:   :...AT <= 97.3: II (2)
:   :   AT > 97.3: I (2)
:   S > 1.2:
:   :...Mg > 1.1: III (2)
:       Mg <= 1.1:
:       :...Argila <= 3.8: IV (5)
:           Argila > 3.8: V (2)
AG > 0.6:
:...AMF > 8: VI (4)
    AMF <= 8:
    :...AT <= 71: X (4)
        AT > 71:
        :...Argila <= 2.3: VIII (2)
            Argila > 2.3: VII (2)


Evaluation on training data (25 cases):

          Decision Tree
        ----------------
        Size      Errors
           9    0( 0.0%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           2                                                           (a): class I
                 2                                                     (b): class II
                       2                                               (c): class III
                             5                                         (d): class IV
                                   2                                   (e): class V
                                         4                             (f): class VI
                                               2                       (g): class VII
                                                     2                 (h): class VIII
                                                                       (i): class IX
                                                           4           (j): class X
```

Regras

```
See5 [Release 1.12]     Sat Nov 04 11:06:08 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 25 cases (21 attributes) from Exp2b.data

Extracted rules:
Rule 1: (2, lift 9.4)
        AT > 97.3
        ->  class I  [0.750]

Rule 2: (2, lift 9.4)
        AT <= 97.3
        S <= 1.2
        ->  class II  [0.750]

Rule 3: (2, lift 9.4)
        AG <= 0.6
        Mg > 1.1
        ->  class III  [0.750]

Rule 4: (5, lift 4.3)
        S > 1.2
        Argila <= 3.8
        AG <= 0.6
        Mg <= 1.1
        ->  class IV  [0.857]

Rule 5: (2, lift 9.4)
        Argila > 3.8
        AG <= 0.6
        Mg <= 1.1
        ->  class V  [0.750]

Rule 6: (4, lift 5.2)
        AG > 0.6
        AMF > 8
        ->  class VI  [0.833]

Rule 7: (2, lift 9.4)
        AT > 71
        Argila > 2.3
        AMF <= 8
        ->  class VII  [0.750]

Rule 8: (2, lift 9.4)
        Argila <= 2.3
        AMF <= 8
        ->  class VIII  [0.750]

Rule 9: (4, lift 5.2)
        AT <= 71
        AMF <= 8
        ->  class X  [0.833]

Default class: IV


Evaluation on training data (25 cases):

            Decision Tree           Rules
          ----------------    ----------------
```

Size	Errors	No	Errors	
9	0(0.0%)	9	0(0.0%)	<<

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	<-classified as
2										(a): class I
	2									(b): class II
		2								(c): class III
			5							(d): class IV
				2						(e): class V
					4					(f): class VI
						2				(g): class VII
							2			(h): class VIII
										(i): class IX
									4	(j): class X

Análise dos resultados

Este experimento utiliza dados das análises granulométrica e química, nas profundidades de 10 a 20cm, perfazendo um total de 25 amostras (ou casos) com 21 atributos em cada uma, dos quais 20 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Não foram computados erros de classificação, cabendo a cada SGH uma folha na árvore. Observa-se que prevaleceram os atributos oriundos da granulometria (apenas o teor de Magnésio é utilizado, secundariamente, dentre aqueles produzidos pela análise química) e por esse motivo a estrutura geral do classificador é semelhante à apresentada no experimento 1a. Três subconjuntos são formados, o primeiro deles com Areia Grossa em volumes bastante destacados, correspondendo a terrenos bastante diversificados, com ocorrência de materiais arenosos (geralmente grosseiros), argilosos e silte; o segundo composto quase que exclusivamente por material arenoso muito fino e com quantidade muito pequena de Silte; e o terceiro, igualmente arenoso, mas de fino a muito fino, com subdivisões com base no teor de Magnésio e de Argila.

Dentre as SGHs classificadas no primeiro grupo, a região VI é definida vagamente como possuindo Areia Grossa em níveis mais perceptíveis e quantidades de Areia Muito Fina acima de 8% (o que não é muito esclarecedor se considerarmos que a média geral desse material gira em torno de 48%). A SGH X é caracterizada como possuindo as menores proporções de Areia Total e quantidades muito pequenas de Areia Muito Fina. A SGH VII é definida como possuindo muito pouca Areia Muito Fina e quantidades de Argila próximas da média geral e a região VIII como possuindo também muito pouca Areia Muito Fina e Argila (que na verdade, não foi detectada em nenhuma de suas amostras).

As SGHs I e II fazem parte do segundo grupo e diferem entre si basicamente quanto à proporção de Areia Total, maior na primeira ainda que apenas ligeiramente.

No terceiro grupo, as amostras com as maiores quantidades de Magnésio são classificadas como SGH III e as restantes, divididas com base na quantidade de Argila indicam as SGHs V (onde esse material aparece em quantidades acima da média geral) e IV.

Regras de Decisão

O conjunto de regras de decisão gerado também não apresentou erros de classificação sobre os dados de treinamento, tendo sido gerada uma única regra para cada SGH, em geral bastante simplificadas.

De acordo com a regra 1, apenas com base no índice de Areia Total é possível identificar os materiais da SGH I: se essa quantidade for muito grande (superior a 97%), o material é típico dessa região.

Quantidades de Areia Total inferiores a 97.3% associadas com proporções muito pequenas de Silte indicam terrenos da SGH II, segundo a regra 2.

A regra 3 classifica os terrenos da SGH III como possuindo quase nenhuma Areia Grossa e as maiores quantidades de Magnésio. É um regra interessante pela sua simplicidade, na medida em que as amostras da SGH III sempre foram problemáticas quanto à classificação nos outros experimentos, onde aparecia associada com regras muito complexas ou com altas taxas de erro. Talvez o Magnésio seja um indicador importante da identidade dos terrenos desta SGH.

A SGH IV é descrita por uma regra relativamente complexa: quantidades mínimas de Areia Grossa; pouca Argila (abaixo da média geral); algum Silte e, eventualmente, proporções moderadas de Magnésio.

A regra 5 especifica que a SGH V possui muito pouca Areia Grossa, alguma Argila e Magnésio.

Conforme a regra 6, os terrenos da SGH VI possuem Areia Grossa em quantidades um pouco maiores, juntamente com quantidades significativas, embora pequenas, de Areia Muito Fina.

Quantidades ligeiramente abaixo da média de Areia Total, associadas com um volume muito pequena de Areia Muito Fina e alguma Argila são características da SGH VII, segundo indica a regra 7.

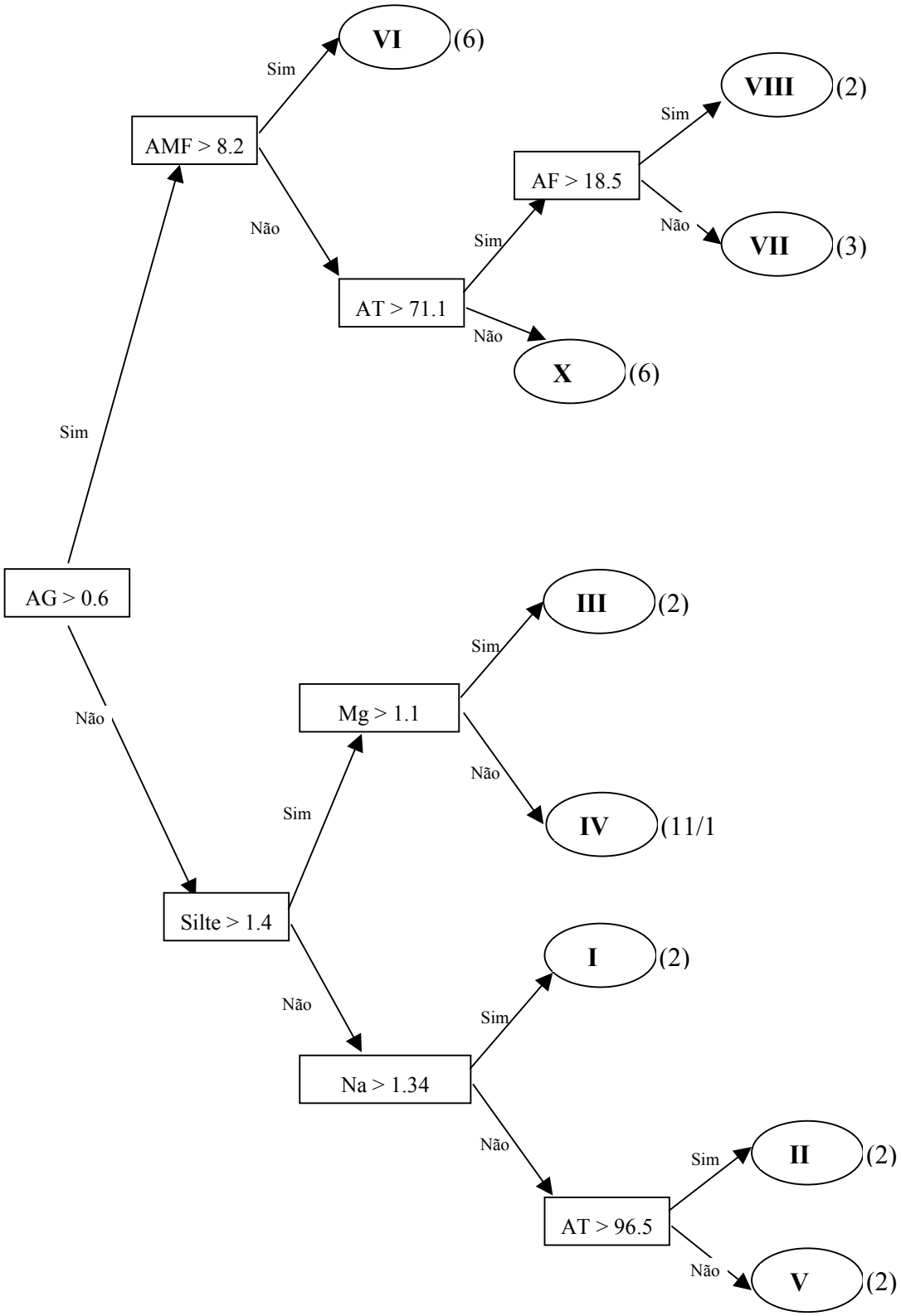
Pouca Argila e pouca Areia Muito Fina são típicas das amostras da SGH VIII, conforme a regra 8, enquanto que pouca Areia Total e pouca Areia Muito Fina são as características principais identificadas pela regra 9 para a SGH X.

Interpretação dos resultados

De uma maneira geral, observa-se que o aspecto granulométrico é mais importante para a caracterização das SGHs do que o químico. A única grande contribuição dos dados químicos foi na caracterização da SGH III (“Duna”), que sempre foi muito difícil de ser estabelecida com base apenas nos valores das análises macroscópica e granulométrica. Essa contribuição pode mesmo ter sido decisiva para a maior simplicidade da árvore quando comparada ao resultado do experimento 1a, mas essa afirmação é de difícil comprovação prática, uma vez que algumas amostras utilizadas no primeiro experimento não puderam ser utilizadas neste, visto que não possuiam dados da análise química.

Ficaram bem claras as definições das SGHs I (“Planície de Maré”) e II (“Berma”), com sua altíssima quantidade de Areia Total e X (“Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos”), com sua baixa proporção de areias, principalmente das mais finas. As SGHs IV (“Planície Litorânea de Cordões Regressivos”) e V (“Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes”) não ficaram muito bem caracterizadas, mas em linhas gerais pode-se dizer que tratam-se de terrenos arenosos finos a muito finos com presença de Argila ou Silte.

4.2.4 Versão 2b (10-40cm)



Árvore

```
See5 [Release 1.12]     Sat Nov 04 11:10:27 2000
Class specified by attribute `SGH'
Read 36 cases (21 attributes) from Exp2b.data

Decision tree:
AG <= 0.6:
:...S > 1.4:
:   :...Mg <= 1.1: IV (11/1)
:   :   Mg > 1.1: III (2)
:   S <= 1.4:
:   :...Na > 1.34: I (2)
:       Na <= 1.34:
:       :...AT <= 96.5: V (2)
:           AT > 96.5: II (2)
AG > 0.6:
:...AMF > 8.2: VI (6)
    AMF <= 8.2:
    :...AT <= 71.1: X (6)
        AT > 71.1:
        :...AF <= 18.5: VII (3)
            AF > 18.5: VIII (2)

Evaluation on training data (36 cases):

          Decision Tree
        ----------------
        Size      Errors
           9    1( 2.8%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           2                                                           (a): class I
                 2                                                     (b): class II
                       2                                               (c): class III
                            10                                         (d): class IV
                           1      2                                    (e): class V
                                        6                              (f): class VI
                                              3                        (g): class VII
                                                    2                  (h): class VIII
                                                                       (i): class IX
                                                                 6     (j): class X
```

Regras

```
See5 [Release 1.12]     Sat Nov 04 11:11:42 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 36 cases (21 attributes) from Exp2b.data

Extracted rules:
Rule 1: (2, lift 13.5)
        S <= 1.4
        Na > 1.34
        ->  class I  [0.750]

Rule 2: (2, lift 13.5)
        AT > 96.5
        S <= 1.4
        Na <= 1.34
        ->  class II  [0.750]

Rule 3: (2, lift 13.5)
        AG <= 0.6
        Mg > 1.1
        ->  class III  [0.750]

Rule 4: (11/1, lift 3.0)
        S > 1.4
        AG <= 0.6
        Mg <= 1.1
        ->  class IV  [0.846]

Rule 5: (2, lift 9.0)
        AT <= 96.5
        S <= 1.4
        ->  class V  [0.750]

Rule 6: (6, lift 5.2)
        AG > 0.6
        AMF > 8.2
        ->  class VI  [0.875]

Rule 7: (3, lift 9.6)
        AT > 71.1
        AF <= 18.5
        AMF <= 8.2
        ->  class VII  [0.800]

Rule 8: (2, lift 13.5)
        AT > 71.1
        AF > 18.5
        AMF <= 8.2
        ->  class VIII  [0.750]
```

```
Rule 9: (6, lift 5.2)
        AT <= 71.1
        AMF <= 8.2
        ->  class X  [0.875]

Default class: IV

Evaluation on training data (36 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
            9    1( 2.8%)       9    1( 2.8%)   <<


          (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
         ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
            2                                                           (a): class I
                  2                                                     (b): class II
                        2                                               (c): class III
                             10                                         (d): class IV
                             1     2                                    (e): class V
                                         6                              (f): class VI
                                               3                        (g): class VII
                                                     2                  (h): class VIII
                                                                        (i): class IX
                                                                 6      (j): class X
```

Análise dos resultados

Este experimento utiliza dados das análises granulométrica e química, nas profundidades de 10 a 40cm, perfazendo um total de 36 amostras (ou casos) com 21 atributos em cada uma, dos quais 20 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Não foram computados erros de classificação, cabendo a cada SGH uma folha na árvore. Observa-se que prevaleceram os atributos oriundos da granulometria (apenas os teores de Magnésio e de Cálcio são utilizados, dentre aqueles produzidos pela análise química) e por esse motivo a estrutura geral do classificador é, em parte, semelhante à apresentada no experimento 1a. Dois subconjuntos são formados, o primeiro deles com Areia Grossa em volumes mais destacados, correspondendo a terrenos bastante diversificados, com ocorrência de materiais arenosos (geralmente grosseiros), argilosos e silte; o segundo composto por material predominantemente arenoso de fino a muito fino, com subdivisões com base nos valores de Silte, Magnésio, Cálcio e Areia Total.

As SGHs VI, VII, VIII e X, que formam o primeiro grupo e possuem uma idade relativa de antiga para média, caracterizam-se inicialmente pela presença de Areia Grossa em quantidades perceptíveis (chegando a cerca de 40% do total da amostra em alguns casos). Além disso, a área VI possui teores de Areia Muito Fina significativos, ainda que bem abaixo da média geral, destacando-se das regiões VII, VIII e X, que apresentam teores bastante baixos desse tipo de material. Adicionalmente, a região X destoa dentro deste subconjunto por possuir uma relativamente pequena proporção de Areia Total (abaixo de 71% para uma média geral de 82%) e a região VIII destaca-se por possuir uma relativamente alta proporção de Areia Fina.

No segundo grupo, proporções pequenas mas significativas de Silte caracterizam as SGHs III e IV, que distinguem-se entre si pelo teor de Magnésio observado, muito maior na região III. As amostras da SGH I são classificadas pela árvore como contendo muito pouco Silte (menos de 1.4%, quando a média geral desse material é de 11%) e taxas muito grandes de Sódio (acima de 1.34 ppm, quando a média é 0.95 ppm, sendo que verificou-se no banco de dados estarem os valores dessas amostras entre 2.74 ppm e 3.93 ppm). Os terrenos com menores teores de Sódio nessa subárvore pertencem às SGHs II e V, que diferem quanto ao percentual de Areia Total, maior na primeira região.

Regras de Decisão

O conjunto de regras de decisão gerado apresentou um único erro de classificação sobre os dados de treinamento, que representa 2.8% do total. Foi gerada uma única regra, em geral bastante simplificada, para cada SGH.

A regra 1 caracteriza a SGH I como possuindo muito Sódio e pouco Silte, o que é até certo ponto surpreendente dada sua altíssima quantidade de Areia Total e de Areia Muito Fina, ignoradas pelo classificador.
A SGH II é caracterizada na regra 2 como possuindo uma porcentagem muito grande de Areia Total, muito pouco Silte e quantidades de Sódio inferiores à da SGH I.

Muito pouca Areia Grossa e teor muito alto de Magnésio são a principal característica da SGH III, segundo a regra 3.

De acordo com a regra 4, a SGH IV possui algum Silte, muito pouca Areia Grossa e algum Magnésio.

A SGH V é definida em termos de uma pequena quantidade de Silte e proporções de Areia Total inferiores a 97%, o que é um critério pouco específico para fins analíticos, assim como o critério definido na regra 6 para a SGH VI: Areia Grossa superior a 0.6% (quando a média geral desse atributo é de quase 8%) e Areia Muito Fina acima de 8% (sendo que a média geral desse componente é de 50%).

A regra 7 indica que a SGH VII possui uma quantidade razoável de Areia Total (acima de 70%) mas proporções bastante reduzidas de Areia Fina e Areia Muito Fina, significando a predominância de material arenoso mais grosseiro.

A regra 8 classifica as amostras da SGH VIII como possuindo apenas uma proporção de Areia Fina um pouco superior à da SGH VII.

Finalmente, a SGH X é classificada na regra 9 como possuindo proporções relativamente pequenas de Areia Total e muito pouca Areia Muito Fina, indicando com isso que trata-se de terrenos com maior ocorrência de Argila e Silte e material arenoso mais grosseiro.

Interpretação dos resultados

Algumas conclusões interessantes podem ser tiradas da análise do classificador produzido e da comparação de seus resultados com o dos experimentos anteriores. Sem dúvida, o aspecto granulométrico parece prevalecer sobre o químico num primeiro momento da classificação, e fica clara a distinção entre o grupo das SGHs VI ("Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos") em relação às demais devido às suas características bastante diversificadas de areias, tipicamente mais grossas que a média, e grandes percentuais de argila e silte.

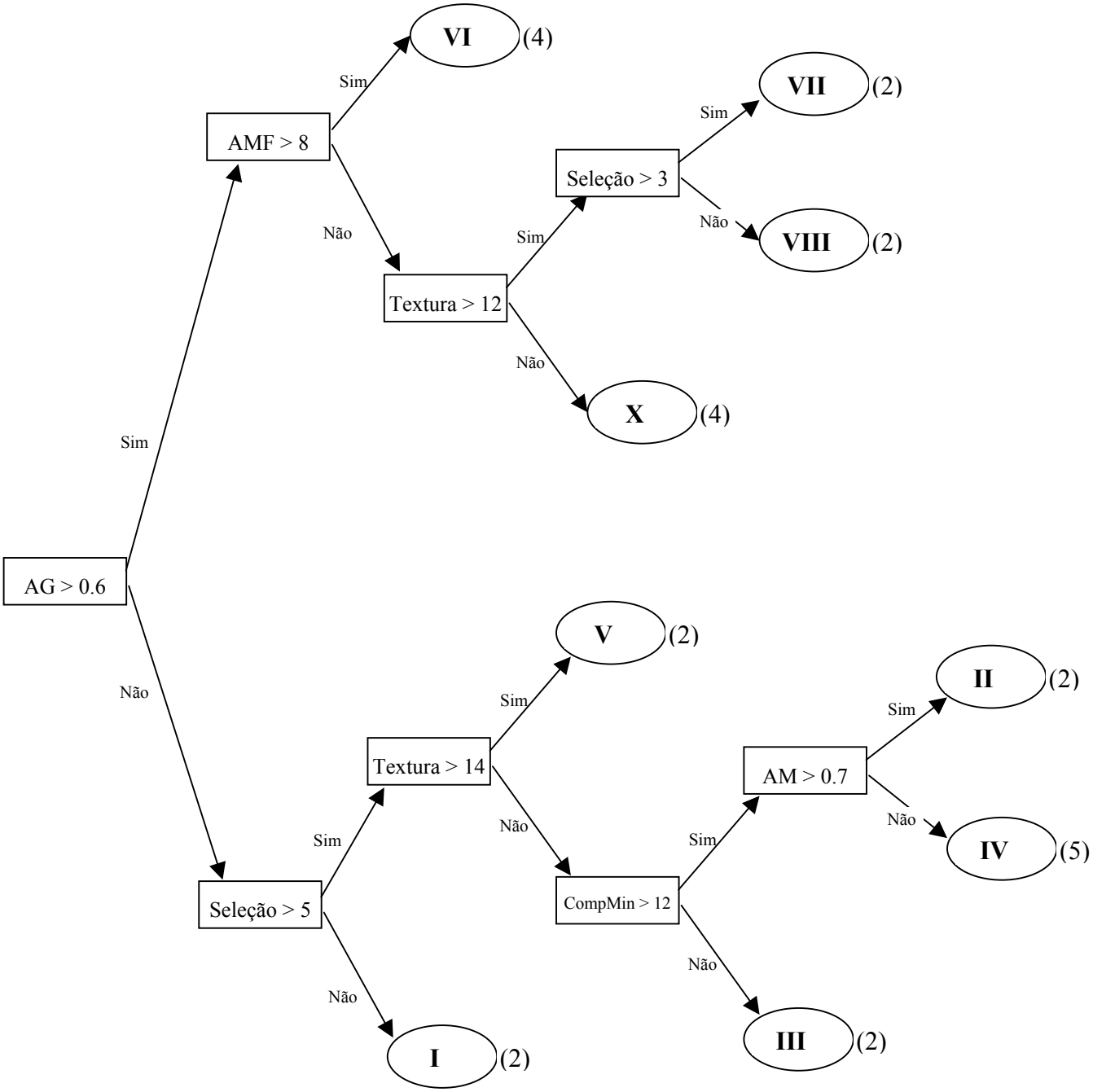
As SGHs do segundo grupo, I ("Planície de Maré"), II ("Berma"), III ("Duna"), IV ("Planície Litorânea de Cordões Regressivos") e V ("Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes") possuem uma maior uniformidade granulométrica (terrenos arenosos finos e muito finos, com pouca ou nenhuma Argila ou Silte), o que torna um pouco mais difícil (e até mesmo inviável) sua individualização com base apenas nesses critérios. É então que parece ter uma maior importância a composição

química, que consegue evidenciar algumas distinções muito mal resolvidas anteriormente. Assim, as SGHs I e II possuem em comum um altíssimo percentual de Areia Total e de Areia Muito Fina, que não permite diferenciar com segurança essas amostras com base apenas nessas características, mas observa-se que a primeira possui uma grande quantidade de Sódio, não verificada na segunda. A SGH III também é de muito difícil caracterização com base em sua granulometria, mas a introdução do atributo teor de Magnésio como um dos critérios de classificação confere um perfil único entre todas as SGHs, que pode ser definido com simplicidade como "predomínio de areias finas e muito finas com muito altas quantidades de Magnésio".

As SGHs IV e V, por sua vez, não ficaram muito bem caracterizadas, diferindo entre si principalmente pela menor quantidade de Silte presente na segunda.

Uma comparação entre os resultados obtidos com os dados dos níveis de 10 a 20cm e de 10 a 40cm revela uma grande semelhança na estrutura dos classificadores obtidos, mas que não deve ser generalizada visto que dois terços das amostras são comuns aos dois conjuntos de treinamento.

4.2.5 Versão 2c (10-20cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 24 10:27:08 2000
Class specified by attribute `SGH'
Read 25 cases (24 attributes) from Exp2c.data

Decision tree:
AG > 0.6:
:...AMF > 8: VI (4)
:   AMF <= 8:
:   :...Textura <= 12: X (4)
:       Textura > 12:
:       :...Selecao <= 3: VIII (2)
:           Selecao > 3: VII (2)
AG <= 0.6:
:...Selecao <= 5: I (2)
    Selecao > 5:
    :...Textura > 14: V (2)
        Textura <= 14:
        :...CompMineral <= 12: III (2)
            CompMineral > 12:
            :...AM <= 0.7: IV (5)
                AM > 0.7: II (2)


Evaluation on training data (25 cases):

          Decision Tree
        ----------------
        Size      Errors
           9    0( 0.0%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           2                                                           (a): class I
                 2                                                     (b): class II
                       2                                               (c): class III
                             5                                         (d): class IV
                                   2                                   (e): class V
                                         4                             (f): class VI
                                               2                       (g): class VII
                                                     2                 (h): class VIII
                                                                       (i): class IX
                                                           4           (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 24 10:30:19 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 25 cases (24 attributes) from Exp2c.data

Extracted rules:
Rule 1: (2, lift 9.4)
        Selecao <= 5
        AG <= 0.6
        ->  class I  [0.750]

Rule 2: (2, lift 9.4)
        Textura <= 14
        AG <= 0.6
        AM > 0.7
        ->  class II  [0.750]

Rule 3: (2, lift 9.4)
        Selecao > 5
        CompMineral <= 12
        AG <= 0.6
        ->  class III  [0.750]

Rule 4: (5, lift 4.3)
        Textura <= 14
        CompMineral > 12
        AM <= 0.7
        ->  class IV  [0.857]

Rule 5: (2, lift 9.4)
        Textura > 14
        AG <= 0.6
        ->  class V  [0.750]

Rule 6: (4, lift 5.2)
        AG > 0.6
        AMF > 8
        ->  class VI  [0.833]

Rule 7: (2, lift 9.4)
        Selecao > 3
        Textura > 12
        AMF <= 8
        ->  class VII  [0.750]

Rule 8: (2, lift 9.4)
        Selecao <= 3
        Textura > 12
        AMF <= 8
        ->  class VIII  [0.750]
```

```
Rule 9: (4, lift 5.2)
        Textura <= 12
        AMF <= 8
        ->  class X  [0.833]

Default class: IV


Evaluation on training data (25 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
            9    0( 0.0%)       9    0( 0.0%)   <<


        (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
       ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
          2                                                           (a): class I
                2                                                     (b): class II
                      2                                               (c): class III
                            5                                         (d): class IV
                                  2                                   (e): class V
                                        4                             (f): class VI
                                              2                       (g): class VII
                                                    2                 (h): class VIII
                                                                      (i): class IX
                                                               4      (j): class X
```

Análise dos resultados

Este experimento utiliza dados das análises macroscópica, granulométrica e química, nas profundidades de 10 a 20cm, perfazendo um total de 25 amostras (ou casos) com 24 atributos em cada uma, dos quais 23 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta.

Árvore de decisão

Não foram computados erros de classificação, cabendo a cada SGH uma folha na árvore. Observa-se que prevaleceram os atributos oriundos das análises granulométrica e macroscópica, não havendo a participação de qualquer atributo referente aos dados químicos das amostras. Dois subconjuntos são formados, com base na quantidade de Areia Grossa verificada: o primeiro deles, onde esse atributo possui maiores valores, é formado pelas SGHs VI, VII, VIII e X e corresponderiam a terrenos bastante diversificados, com ocorrência de materiais arenosos geralmente mais grosseiros e Textura e grau de Seleção mais variados; e o segundo, composto pelas outras 5 SGHs, apresentando terrenos arenosos mais finos e bem selecionados.

Além da maior quantidade relativa de Areia Grossa, as SGHs do primeiro subconjunto possuem algumas características particulares, segundo o classificador induzido: a SGH VI se destaca por possuir os maiores teores de Areia Muito Fina desse grupo, em valores próximos à média geral de 48% para esse atributo, enquanto que nas 3 regiões restantes esse valor é muito baixo, sempre inferior a 8%. A SGH X, por sua vez, apresenta uma Textura mais fina, argilosa; e as regiões VII e VIII se distinguem pela Textura mais grosseira e também quanto ao grau de Seleção, que é "medianamente selecionado" na primeira e "mal selecionado" na segunda.

As SGHs do segundo conjunto são subdivididas principalmente com base nos dados da análise macroscópica. A região I destaca-se nesse grupo por ser a única que não possui grau de Seleção "Muito bem selecionada". Dentre as SGH restantes, a V apresenta uma Textura um pouco mais grosseira que as demais; a III caracteriza-se por não possuir um predomínio de quartzo puro em sua Composição Mineral e a SGH IV difere da SGH II por possuir uma quantidade menor de Areia Média.

Regras de Decisão

O conjunto de regras de decisão gerado também não apresentou erros de classificação sobre os dados de treinamento, tendo sido gerada uma única regra para cada SGH.

A regra 1 caracteriza a SGH I como possuindo quantidades ínfimas de Areia Grossa e grau de Seleção de "muito mal selecionada" até "bem selecionada" o que, pela definição desse atributo no banco de dados deve ser interpretado como "diferente de muito bem selecionada".

A regra 2 define as amostras da SGH II como possuindo quantidades muito pequenas de Areia Grossa, Textura entre "argilosa" e "fina argilosa" e alguma ocorrência de Areia Média.

Segundo a definição da regra 3, a SGH III caracteriza-se por apresentar um grau de Seleção "muito bem selecionado"; uma Composição Mineral de quartzo com presença de mica e suas variedades subordinadas e muito pouca Areia Grossa.

A regra 4 indica que na SGH IV a Textura encontra-se entre "argilosa" e "fina argilosa"; na Composição Mineral há um predomínio de quartzo mais puro, sem outros materiais subordinados, e existe muito pouca Areia Média.

A SGH V é caracterizada pela regra 5 como possuindo Textura de "fina argilosa" até "muito grossa" e quantidades mínimas de Areia Grossa.

Conforme a regra 6, a SGH VI apresenta, dentre aquelas que possuem teores mais perceptíveis de Areia Grossa (superiores a 0.6%) uma maior proporção de Areia Muito Fina.

A regra 7 define a SGH VII como apresentando grau de Seleção de "medianamente selecionada" até "muito bem selecionada"; Textura não "muito fina" e quantidades muito pequenas de Areia Muito Fina, inferiores a 8% quando a média geral desse atributo é de 48%, indicando o predomínio de terrenos arenosos mais grosseiros.

A SGH VIII é classificada pela regra 8 como possuindo grau de Seleção inferior a "medianamente selecionado" e, a exemplo da SGH VII, Textura não "muito fina" e quantidades muito pequenas de Areia Muito Fina, inferiores a 8%, novamente significando um predomínio de terrenos arenosos mais grosseiros.

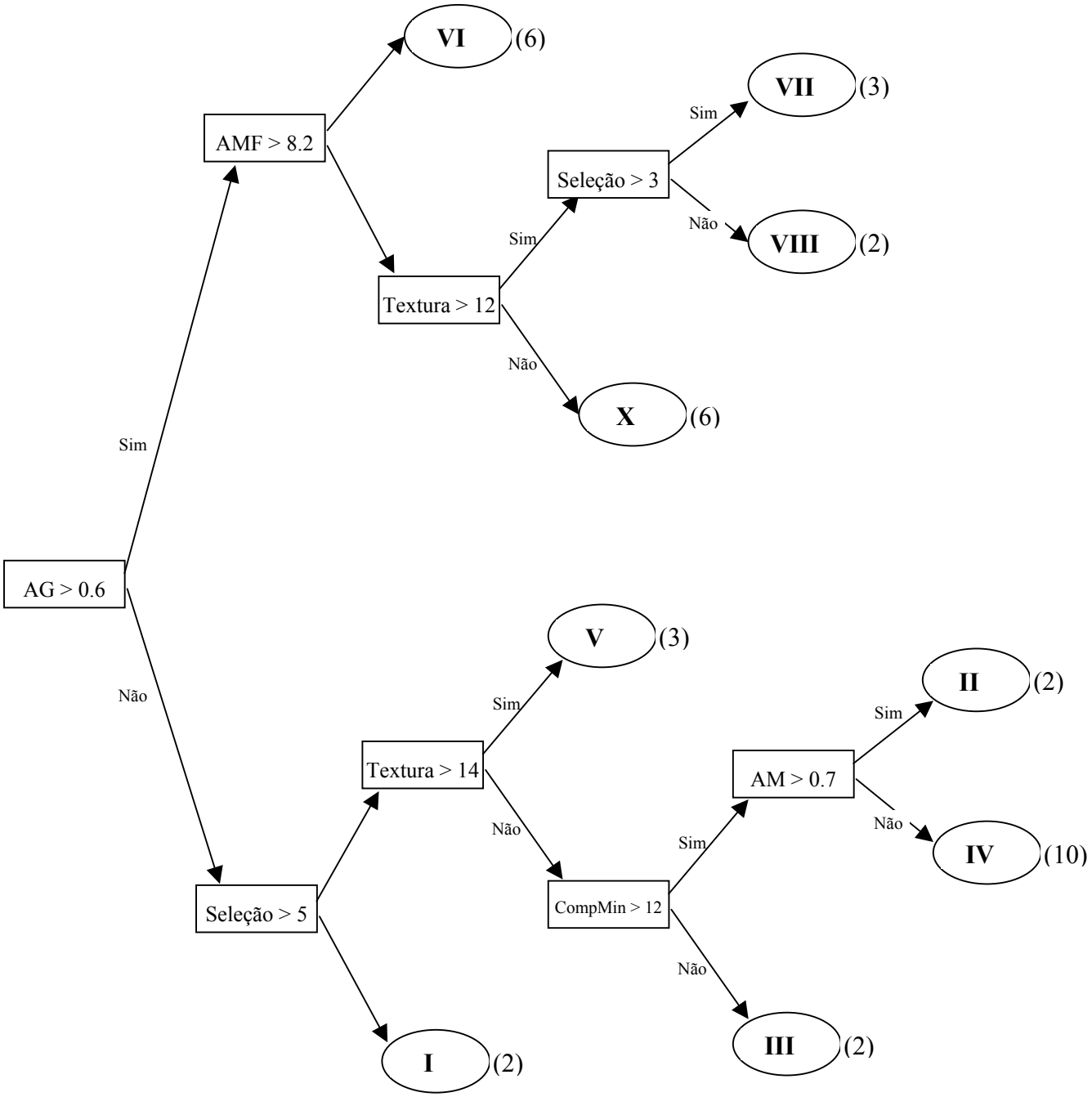
Finalmente, a regra 9 define a SGH X como caracterizando-se por Textura entre "argilosa" e "muito fina", com pequena ocorrência de Areia Muito Fina, o que indica uma maior presença de Silte e Argila na composição desses terrenos.

Interpretação dos resultados

A principal conclusão que pode ser tirada deste experimento é que os dados das análises macroscópica e granulométrica parecem ser suficientes para, com base nos dados fornecidos para o treinamento do classificador, diferenciar os terrenos das 9 SGHs, já que nenhum dado da análise química foi selecionado para compor a árvore ou o conjunto de regras. Seria de se esperar então que o classificador induzido fosse idêntico ao gerado no experimento 1b, que combina apenas os dados granulométricos e macroscópicos, mas isso não ocorre porque o conjunto de treinamento é diferente em ambos os experimentos, já que apenas uma parte das amostras deste nível de profundidade possui dados da análise química, necessários neste experimento. As linhas gerais dos dois classificadores são, entretanto, semelhantes e permitem atingir mais ou menos as mesmas conclusões.

Novamente, a exemplo do que ocorreu nos experimentos anteriores, definiu-se inicialmente uma divisão das regiões com base no teor de Areia Grossa, formando-se os mesmos dois conjuntos já vistos nas análises precedentes: um com terrenos mais diversificados e outro com terrenos arenosos muito finos. O primeiro é formado pelas SGHs VI ("Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos") e o segundo formado pela "Planície de Maré" (SGH I), "Berma" (II), "Duna" (III), "Planície Litorânea de Cordões Regressivos" (IV) e "Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes" (V).

4.2.6 Versão 2c (10-40cm)



Árvore

```
See5 [Release 1.12]     Fri Nov 24 10:31:48 2000
Class specified by attribute `SGH'
Read 36 cases (24 attributes) from Exp2c.data

Decision tree:
AG > 0.6:
:...AMF > 8.2: VI (6)
:   AMF <= 8.2:
:   :...Textura <= 12: X (6)
:       Textura > 12:
:       :...Selecao <= 3: VIII (2)
:           Selecao > 3: VII (3)
AG <= 0.6:
:...Selecao <= 5: I (2)
    Selecao > 5:
    :...Textura > 14: V (3)
        Textura <= 14:
        :...CompMineral <= 12: III (2)
            CompMineral > 12:
            :...AM <= 0.7: IV (10)
                AM > 0.7: II (2)


Evaluation on training data (36 cases):

          Decision Tree
        ----------------
        Size      Errors
           9    0( 0.0%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
        ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
           2                                                           (a): class I
                 2                                                     (b): class II
                       2                                               (c): class III
                            10                                         (d): class IV
                                  3                                    (e): class V
                                        6                              (f): class VI
                                              3                        (g): class VII
                                                    2                  (h): class VIII
                                                                       (i): class IX
                                                               6       (j): class X
```

Regras

```
See5 [Release 1.12]     Fri Nov 24 10:32:43 2000
    Options:
        Generating rules
Class specified by attribute `SGH'
Read 36 cases (24 attributes) from Exp2c.data

Extracted rules:
Rule 1: (2, lift 13.5)
        Selecao <= 5
        AG <= 0.6
        ->  class I  [0.750]

Rule 2: (2, lift 13.5)
        Textura <= 14
        AG <= 0.6
        AM > 0.7
        ->  class II  [0.750]

Rule 3: (2, lift 13.5)
        Selecao > 5
        Textura <= 14
        CompMineral <= 12
        AG <= 0.6
        ->  class III  [0.750]

Rule 4: (10, lift 3.3)
        Textura <= 14
        CompMineral > 12
        AM <= 0.7
        ->  class IV  [0.917]

Rule 5: (3, lift 9.6)
        Textura > 14
        AG <= 0.6
        ->  class V  [0.800]

Rule 6: (6, lift 5.2)
        AG > 0.6
        AMF > 8.2
        ->  class VI  [0.875]

Rule 7: (3, lift 9.6)
        Selecao > 3
        Textura > 12
        AMF <= 8.2
        ->  class VII  [0.800]

Rule 8: (2, lift 13.5)
        Selecao <= 3
        Textura > 12
        AMF <= 8.2
        ->  class VIII  [0.750]
```

```
Rule 9: (6, lift 5.2)
        Textura <= 12
        AMF <= 8.2
        ->  class X  [0.875]

Default class: IV


Evaluation on training data (36 cases):

           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
            9    0( 0.0%)       9    0( 0.0%)   <<


        (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
       ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
          2                                                           (a): class I
                2                                                     (b): class II
                      2                                               (c): class III
                           10                                         (d): class IV
                                 3                                    (e): class V
                                       6                              (f): class VI
                                             3                        (g): class VII
                                                   2                  (h): class VIII
                                                                      (i): class IX
                                                               6      (j): class X
```

Análise dos resultados

Este experimento utiliza dados das análises macroscópica, granulométrica e química, nas profundidades de 10 a 40cm, perfazendo um total de 36 amostras (ou casos) com 24 atributos em cada uma, dos quais 23 são efetivamente considerados pelo mecanismo indutor de regras da ferramenta. O classificador induzido a partir desses dados é totalmente equivalente ao produzido com os dados de 10 a 20cm, sendo a única diferença observada no valor de corte (*split value*) calculado para o atributo Areia Muito Fina, que passou de 8% para 8.2%. Essa semelhança é devida à diferença muito pequena na composição das amostras em um experimento e no outro, já que o conjunto de treinamento aqui utilizado possui todas as 25 amostras de 10 a 20cm mais 9 outras referentes às profundidades de 20 a 40cm, coletadas nos mesmos pontos amostrais das primeiras. Torna-se portanto desnecessário repetir as análises feitas para o nível de 10 a 20cm, uma vez que as conclusões que podem ser tiradas são rigorosamente as mesmas.

Resultados gerais dos experimentos

Uma clara distinção entre dois grupos de SGHs ficou evidente ao longo de todos os experimentos: de um lado os terrenos arenosos finos das regiões I ("Planície de Maré"), II ("Berma"), III ("Duna"), IV ("Planície Litorânea de Cordões Regressivos") e V ("Planície Litorânea de Cordões Regressivos com Micro-Canais Interligantes"), cuja individualização nem sempre é muito fácil por meios granulométricos e de outro lado, os terrenos mais diversificados das SGHs VI ("Planície de Retrabalhamento Fluvial-Marinho"), VII ("Planície Colúvio-Aluvionar"), VIII ("Planície Colúvio-Aluvionar com Micro-Canais Interligantes") e X ("Planície Litorânea de Cordões Regressivos Recobertos por Sedimentos Continentais Finos"), onde predomina material arenoso mais grosseiro e, conforme o local, Silte e Argila. Essa divisão se manteve sistematicamente em praticamente todos os experimentos, à exceção do 2a, que foi o mais prejudicado pela pequena quantidade de material disponível para os testes.

A introdução, no experimento 1b, dos dados da análise macroscópica mostrou uma sensível melhora na qualidade dos classificadores induzidos em relação àqueles gerados apenas com base nos dados granulométricos (experimento 1a), tanto por simplificar a estrutura da árvore de decisão como por diminuir a taxa de erro. Isso significa que os atributos Grau de Seleção, Textura e Composição Mineral do material possuem a capacidade de melhorar a classificação das amostras em função das SGHs em que foram recolhidas.

Algumas alterações puderam ser percebidas, e as principais encontram-se devidamente registradas nas análises específicas de cada experimento, ao se variar os níveis de profundidade das amostras consideradas em cada conjunto de treinamento. Isso pode ser apenas um efeito da pequena quantidade de amostras em cada experimento ou então constituir-se num indício de variações estruturais na composição dos terrenos em função do aumento da profundidade do terreno.

Os dados químicos não apresentaram um resultado importante nas análises, provavelmente por que seu volume era o mais reduzido dentre todas as categorias. Os classificadores induzidos com base apenas neles foram de péssima qualidade e a sua combinação com os dados macroscópicos e granulométricos levou a árvores de decisão onde apenas o Magnésio teve alguma participação importante. O principal indício levantado foi de que talvez esse elemento químico mereça alguma atenção do especialista ao se estudar a SGH III ("Duna"), que por sua própria natureza sempre foi a região de classificação mais difícil e na qual a presença do Magnésio pareceu representar um elemento simplificador.

Capítulo 5: Conclusões

Este trabalho procurou avaliar, através de um estudo de caso, as possibilidades de utilização de técnicas de classificação automática em pesquisas na área ambiental, tendo permitido chegar a algumas conclusões mais genéricas e a outras mais específicas para o caso estudado.

Dentre as conclusões mais gerais, a primeira é que a coleta de material para os experimentos deve levar em conta, de alguma forma, a posterior utilização de ferramentas de mineração de dados. Esse cuidado é necessário tanto para permitir um melhor aproveitamento dos recursos das ferramentas adotadas como também para dimensionar as expectativas quanto às possibilidades oferecidas por seu emprego no contexto daquele trabalho em particular. Em segundo lugar, ferramentas indutoras de regras de classificação permitem acelerar o processo de análise nas pesquisas, uma vez que conseguem extrair e documentar com rapidez estruturas lógicas potencialmente interessantes a partir de volumes consideráveis de dados, em situações impraticáveis para um ser humano. A representação dos classificadores na forma de árvores de decisão e de regras de produção parece bastante intuitiva para viabilizar sua utilização mesmo por pesquisadores com conhecimento muito pequeno sobre aprendizagem de máquina ou quanto aos detalhes internos do funcionamento do *software* adotado. Entretanto, tais ferramentas devem ser encaradas apenas como elementos de suporte à decisão dentro de um projeto, uma vez que o papel do especialista é fundamental, tanto para conduzir o processo interativo de ajustes dos experimentos como para identificar as estruturas lógicas de maior interesse em função dos objetivos da pesquisa como, ainda, para atribuir um maior significado para essas estruturas nos termos do domínio do problema focalizado. Assim, a ferramenta apenas levanta indícios sugeridos pelos dados disponíveis, e cabe ao especialista interpretá-los, avaliá-los e decidir se merecem algum tipo de atenção especial ou não. Essa capacidade de julgamento é também necessária para a avaliação do impacto sobre as conclusões que podem ser obtidas devido a eventuais limitações presentes na base de dados ou às características técnicas do mecanismo de indução utilizado, por exemplo.

Com relação, mais especificamente, à utilização da Dissertação de Mestrado "Análise Geomorfológica e Distribuição Espacial da Vegetação na Planície Litorânea de Picinguaba (Ubatuba - SP)" de José Paulo Marsola Garcia (1995) como pano de fundo para a avaliação da ferramenta, foram identificados uma série de aspectos que merecem consideração. Em primeiro lugar, verificou-se que as restrições decorrentes do volume muito reduzido de dados impedem qualquer afirmação categórica a partir dos dados em si. Além do volume absoluto dos dados, fatores como a existência de amostras provenientes de vários níveis de profundidade distintos, o relativamente alto número de classes e de atributos e a falta de dados para muitas amostras (principalmente da análise química) serviram para potencializar o problema, que parece ser típico de trabalhos dessa natureza, freqüentemente condicionados à fortes limitações orçamentárias. Apesar disso, ainda foi possível a adoção de uma abordagem interessante, uma vez que os mesmos poucos dados que permitiram ao especialista chegar, por outros meios, a determinadas conclusões no trabalho original, foram aqueles utilizados nos experimentos com o See5. E nesse aspecto, o principal resultado obtido pela indução automática de regras é que os dados das análises

granulométrica e macroscópica parecem ser suficientes para validar as unidades ambientais (SGHs) identificadas, como afirma a conclusão do trabalho de referência.

As adaptações realizadas para alguns atributos qualitativos que encontravam-se relativamente mal-estruturados no banco de dados, e para os quais foram definidas escalas ordenadas, permitiram a utilização dos dados da análise macroscópica nos experimentos. Essa abordagem merece especial atenção quando do emprego da ferramenta See5 e pode ter larga utilidade em trabalhos que a apliquem.

Alguns tipos de análises que envolvem classificação apresentam uma maior complexidade e oferecem espaço para futuros aprofundamentos. É o caso, por exemplo, da consideração das variedades de orquídeas em conjunto com os aspectos físicos do terreno, que poderia ser feita com auxílio de alguma categorização adicional das espécies observadas ou com o estabelecimento de uma escala de diversidade baseada no número de variedades encontradas por região, ponderada eventualmente por elementos como o tipo de orquídea ou a área da SGH. Também a caracterização física do terreno considerando a variação dos atributos em função do nível de profundidade poderia ser mais explorada, requerendo entretanto um maior volume de dados.

Referências bibliográficas

BAUER, Eric; KOHAVI, Ron. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. **Machine Learning**, v. 36 (1/2), p. 105-139, july 1999.

BERSON, Alex; SMITH, Stephen J. **Data Warehousing, Data Mining & OLAP**. New York (NY): McGraw-Hill, 1997. 612 p.

BRESLOW, Leonard A.; AHA, David W. Simplifying Decision Trees: A Survey. **NCARAI Technical Report Nº AIC-96-014**. 1996. [online]. Disponível em: <http://www.aic.nrl.navy.mil/papers/1996/AIC-96-014.ps.Z >. Acessado em: 11 dez. 1999.

CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S. Data Mining: An Overview from a Database Perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 866-883, dec. 1996.

DIETTERICH, Thomas G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, **Machine Learning**, v. 40 (2), p. 139-157, aug. 2000.

DYMOND, John R.; LUCKMEN, Paul G. Direct induction of compact rule-based classifiers for resource mapping. **International Journal of Geographical Information Systems**, v. 8, n. 4, p. 357-367, july-aug. 1994.

ELDER IV, John F.; PREGIBON, Daryl. A Statistical Perspective on Knowledge Discovery in Databases. *In*: Fayyad, U.M. *et al.* (eds.), **Advances in Knowledge Discovery and Data Mining**, Menlo Park (CA): AAAI/MIT Press, 1996. Cap 4, p. 83-116.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Fundamentals of database systems**. 3rd ed. Reading (MA): Addison-Wesley, 2000. 955 p.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, Fall 1996, p. 37-54.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, nov. 1996.

FREUND, Yoav; SCHAPIRE, Robert E. A Short Introduction to Boosting. **Journal of Japanese Society for Artificial Intelligence**, v. 14, n. 5, p. 771-780, sep. 1999.

GANTI, Venkatesh; GEHRKE, Johannes; RAMAKRISHNAN, Raghu. Mining Very Large Databases. **Computer**, v. 32, n. 8, p. 38-45, aug. 1999.

GARCIA, J.P.M. **Análise Geomorfológica e Distribuição Espacial da Vegetação na Planície Litorânea de Picinguaba (Ubatuba - SP)**. 1995. 176 p. Dissertação (Mestrado em Geografia Física). Universidade de São Paulo.

GUERRA, A.J.T.; CUNHA, S.B. (Org.). **Geomorfologia e Meio Ambiente**. Rio de Janeiro: Bertrand Brasil, 1966.

GUERRA, A.J.T.; CUNHA, S.B. (Org.). **Geomorfologia: Uma Atualização de Bases e Conceitos**. Rio de Janeiro: Bertrand Brasil, 1998.

GUPTA, S.K. *et al*. Scalable Classifiers with Dynamic Pruning. *In*: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, 9th, 1998, Viena, Áustria. **Proceedings ...**, IEEE Computer Society Press, 1998.

HAN, Jiawei. From Database Systems To Knowledge-Base Systems: A Evolutionary Approach. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 11th, 1995, Taipei, Taiwan, **Conference Tutorial**.

HAND, David J. Data Mining: Statistics and More? **The American Statistician**, v. 52, n. 2, p. 112-118, may 1998.

HART, Anna. **Knowledge Acquisition for Expert Systems**. 2nd ed. London: Kogan Page, 1986.

HUANG, Yueh-Min; LIN, Shian-Hua. An Efficient Inductive Learning Method for Object-Oriented Database Using Attribute Entropy. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 946-951, dec. 1996.

HYAFIL, L.; RIVEST, R. Constructing optimal binary decision trees is NP-complete. **Information Processing Letters**, v. 5, n. 1, p. 15-17, 1976.

JENSEN, David ; OATES, Tim; COHEN, Paul R. Building Simple Models: A Case Study with Decision Trees. *In*: ADVANCES IN INTELLIGENT DATA ANALYSIS: REASONING ABOUT DATA, 2nd, 1997, London, UK, **Proceedings ...**, Springer, 1997, p. 211-222.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. 4th ed. Upper Saddle River (NJ): Prentice-Hall, 1998. 816 p.

LIM, Tjen-Sien; LOH, Wei-Yin; SHIH, Yu-Shan. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms, **Machine Learning**, v. 40 (3), p. 203-228, sep. 2000.

LU, Hongjun; SETIONO, Rudy; LIU, Huan. Effective Data Mining Using Neural

Networks. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 957-961, dec. 1996.

MAGEE, Bryan. **História da Filosofia**. São Paulo: Edições Loyola, 1999.

MANNILA, Heikki. Data mining: machine learning, statistics, and databases. *In*: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, 8th, 1996, Stockholm, Sweden, **Proceedings ...**, IEEE Computer Society Press, 1996, p. 2-9.

MEHTA, Manish; AGRAWAL, Rakesh; RISSANEN, Jorma. SLIQ: A Fast Scalable Classifier for Data Mining. *In*: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, 5th, 1996, Avignon, France, **Proceedings ...**., 1996, p. 18-32.

MICHALSKI, Ryszard S.; BRATKO, Ivan; KUBAT, Miroslav (Ed.). **Machine Learning and Data Mining: Methods and Applications**. Baffins Lane (UK): John Wiley & Sons, 1998. 456 p.

MURPHY, Patrick M.; PAZZANI, Michael J. Exploring the Decision Forest: An Empirical Investigation of Occam's Razor in Decision Tree Induction. **Journal of Artificial Intelligence Research**, v. 1, p. 257-275, mar. 1994.

OATES, T.; JENSEN, D. Toward a Theoretical Understanding of Why and When Decision Tree Pruning Algorithms Fail. *In*: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 6th, 1999, Orlando (FL), **Proceedings ...**. AAAI/MIT Press, 1999, p. 372-378.

OLIVER, Martyn. **História Ilustrada da Filosofia: Os Grandes Filósofos, de 2000 a.C. aos Dias de Hoje**. São Paulo: Manole, 1998.

PEREIRA, Júlio Cesar Rodrigues. **Análise de Dados Qualitativos: Estratégias Metodológicas para as Ciências da Saúde, Humanas e Sociais**. São Paulo: Editora da Universidade de São Paulo, 1999.

PIRES NETO, A.G. **As Abordagens Sintético-Histórica e Analítico-Dinâmica, Uma Proposição Metodológica Para a Geomorfologia**. 1991. 302 p. Tese (Doutorado em Geografia). Universidade de São Paulo.

PYLE, Dorian. **Data preparation for data mining**. San Francisco (CA): Morgan Kaufmann, 1999. 540 p.

QUINLAN, John Ross. **C4.5: Programs for machine learning**. San Mateo (CA): Morgan Kaufmann, 1993. 302 p.

QUINLAN, J.Ross. Improved Use of Continuous Attributes in C4.5. **Journal of Artificial Intelligence Research**, v. 4, p. 77-90, june 1996.

RAMAKRISHNAN, Naren; GRAMA, Anand Y. Data Mining: From Serendipity to Science. **Computer**, v. 32, n. 8, p. 34-37, aug. 1999.

RICH, Elaine; KNIGHT, Kevin. **Inteligência Artificial**. 2. ed. São Paulo: Makron Books, 1993. 722 p.

RULEQUEST RESEARCH. **See5: An Informal Tutorial.** Disponível em: <http://www.rulequest.com/download.html>. Acesso em: 21 jun. 2000.

RUSSEL, Stuart J.; NORVIG, Peter. **Artificial intelligence: a modern approach**. Upper Saddle River (NJ): Prentice-Hall, 1995. 932 p.

SHAFER, John C.; AGRAWAL, Rakesh; MEHTA, Manish. SPRINT: A Scalable Parallel Classifier for Data Mining. *In*: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 22th, 1996, Mumbai, India, **Proceedings ...**. 1996, p. 544-555.

SILBERSCHATZ, Avi; KORTH, Henry; SUDARSHAN S. **Sistema de Banco de Dados**. 3. ed. São Paulo: Makron Books, 1999. 778 p.

TUKEY, John W. **Exploratory Data Analysis**. Reading (MA): Addison Wesley, 1977. 688 p.

WEBB, Geoffrey I. Further Experimental Evidence against the Utility of Occam's Razor. **Journal of Artificial Intelligence Research**, v. 4, p. 397-417, june 1996.

WEISS, Sholom M.; INDURKHYA, Nitin. **Predictive Data Mining: A Practical Guide**. San Francisco (CA): Morgan Kaufmann, 1998. 228 p.

WINSTON, Patrick Henry. **Artificial Intelligence**. 3rd ed. Reading (MA): Addison-Wesley, 1992.

Anexo 1: Resumo estatístico dos experimentos

Experimento: 1a **Profundidade: 10 a 20cm**

Atributo ⇒	Prof.	AT	S	Argila	AMG	AG	AM	AF	AMF	SGH
Média	-	84,3	10,3	5,1	3,9	6,6	6,1	16,3	51,5	-
Máximo	-	98,8	33,2	18,5	61,2	39,2	34,2	36,6	92,2	-
Mínimo	-	48,9	0,4	0,0	0,0	0,0	0,0	3,5	2,7	-
Desvio Padrão	-	13,8	9,4	5,4	10,8	11,1	8,8	9,4	32,8	-

Experimento: 1a **Profundidade: 10 a 40cm**

Atributo ⇒	Prof.	AT	S	Argila	AMG	AG	AM	AF	AMF	SGH
Média	-	85,7	9,2	4,9	4,5	6,3	5,6	15,9	53,5	-
Máximo	-	99,3	40,5	24,8	61,2	42,3	34,2	46,5	93,3	-
Mínimo	-	36,6	0,0	0,0	0,0	0,0	0,0	1,5	1,5	-
Desvio Padrão	-	15,0	9,9	6,0	11,7	10,9	8,4	10,4	33,6	-

Experimento: 1b **Profundidade: 10 a 20cm**

Atributo ⇒	Prof.	Cor	MatOrg	Textura	Comp. Mineral	Selecao	Lencol Freatico
Média	-	-	-	13,0	12,4	5,0	-
Máximo	-	-	-	23,0	20,0	6,0	-
Mínimo	-	-	-	4,0	0,0	2,0	-
Desvio Padrão	-	-	-	4,3	7,7	1,4	-

Atributo ⇒	AT	S	Argila	AMG	AG	AM	AF	AMF	SGH
Média	84,6	10,2	4,9	1,7	6,2	6,0	16,6	54,2	-
Máximo	98,8	33,2	18,5	8,3	39,2	34,2	36,6	92,2	-
Mínimo	48,9	0,4	0,0	0,0	0,0	0,0	3,5	2,7	-
Desvio Padrão	14,0	9,6	5,4	2,9	11,2	9,1	9,5	31,8	-

Experimento: 1b **Profundidade: 10 a 40cm**

Atributo ⇒	Prof.	Cor	MatOrg	Textura	Comp. Mineral	Selecao	Lencol Freatico
Média	-	-	-	12,4	11,4	5,0	-
Máximo	-	-	-	23,0	20,0	6,0	-
Mínimo	-	-	-	0,0	0,0	2,0	-
Desvio Padrão	-	-	-	4,8	7,5	1,4	-

Atributo ⇒	AT	S	Argila	AMG	AG	AM	AF	AMF	SGH
Média	85,5	9,4	4,9	1,9	5,9	5,6	16,4	55,9	-
Máximo	98,8	40,5	24,8	12,1	42,3	34,2	46,5	93,3	-
Mínimo	36,6	0,0	0,0	0,0	0,0	0,0	1,5	1,5	-
Desvio Padrão	15,4	10,2	6,1	3,3	11,1	8,7	10,6	32,3	-

Experimento: 2a **Profundidade: 10 a 20cm**

Atributo ⇒	Prof	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	-	0,2	1,0	1,1	0,3	0,7	1,4	1,0	2,5	4,3	4,9	3,9	-
Máximo	-	0,5	1,5	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	-	0,1	0,4	0,2	0,0	0,0	0,2	0,2	0,6	1,0	4,0	3,1	-
Desvio Padrão	-	0,1	0,3	0,8	0,4	0,5	0,9	1,1	1,1	1,9	0,4	0,3	-

Experimento: 2a **Profundidade: 10 a 40cm**

Atributo ⇒	Prof	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	-	0,2	1,0	0,9	0,3	0,6	1,4	0,9	2,1	3,7	4,9	3,9	-
Máximo	-	0,5	1,8	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	-	0,0	0,3	0,1	0,0	0,0	0,1	0,1	0,3	0,6	4,0	3,1	-
Desvio Padrão	-	0,1	0,4	0,7	0,3	0,5	1,0	1,1	1,2	2,0	0,4	0,3	-

Experimento: 2b **Profundidade: 10 a 20cm**

Atributo ⇒	Prof.	AT	S	Argila	AMG	AG	AM	AF	AMF
Média	-	82,3	11,9	5,5	2,3	8,1	7,7	15,6	48,4
Máximo	-	98,8	33,2	18,5	8,3	39,2	34,2	36,6	92,2
Mínimo	-	48,9	0,4	0,0	0,0	0,0	0,0	6,3	2,7
Desvio Padrão	-	15,4	10,5	6,0	3,1	12,3	9,9	8,6	34,2

Atributo ⇒	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	0,2	1,0	1,1	0,3	0,7	1,4	1,0	2,5	4,3	4,9	3,9	-
Máximo	0,5	1,5	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	0,1	0,4	0,2	0,0	0,0	0,2	0,2	0,6	1,0	4,0	3,1	-
Desvio Padrão	0,1	0,3	0,8	0,4	0,5	0,9	1,1	1,1	1,9	0,4	0,3	-

Experimento: 2b **Profundidade: 10 a 40cm**

Atributo ⇒	Prof.	AT	S	Argila	AMG	AG	AM	AF	AMF
Média	-	82,4	11,3	6,0	2,4	7,9	7,2	14,7	50,1
Máximo	-	98,8	33,2	24,8	12,0	42,3	34,2	36,6	92,2
Mínimo	-	48,9	0,0	0,0	0,0	0,0	0,0	6,3	2,7
Desvio Padrão	-	16,1	10,3	7,0	3,6	12,4	9,3	7,6	34,7

Atributo ⇒	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	0,2	1,0	0,9	0,3	0,6	1,4	0,9	2,2	3,7	4,9	3,9	-
Máximo	0,5	1,8	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	0,0	0,3	0,1	0,0	0,0	0,1	0,1	0,4	0,6	4,0	3,1	-
Desvio Padrão	0,1	0,4	0,7	0,3	0,5	1,0	1,1	1,2	2,0	0,4	0,3	-

Experimento: 2c **Profundidade: 10 a 20cm**

Atributo ⇒	Prof.	Selecao	Textura	Comp. Mineral
Média	-	4,8	12,8	12,6
Máximo	-	6,0	23,0	20,0
Mínimo	-	2,0	4,0	0,0
Desvio Padrão	-	1,6	4,9	7,5

Atributo ⇒	AT	S	Argila	AMG	AG	AM	AF	AMF
Média	82,3	11,9	5,5	2,3	8,1	7,7	15,6	48,4
Máximo	98,8	33,2	18,5	8,3	39,2	34,2	36,6	92,2
Mínimo	48,9	0,4	0,0	0,0	0,0	0,0	6,3	2,7
Desvio Padrão	15,4	10,5	6,0	3,1	12,3	9,9	8,6	34,2

Atributo ⇒	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	0,2	1,0	1,1	0,3	0,7	1,4	1,0	2,5	4,3	4,9	3,9	-
Máximo	0,5	1,5	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	0,1	0,4	0,2	0,0	0,0	0,2	0,2	0,6	1,0	4,0	3,1	-
Desvio Padrão	0,1	0,3	0,8	0,4	0,5	0,9	1,1	1,1	1,9	0,4	0,3	-

Experimento: 2c **Profundidade: 10 a 40cm**

Atributo ⇒	Prof.	Selecao	Textura	Comp. Mineral
Média	-	4,9	11,7	12,4
Máximo	-	6,0	23,0	20,0
Mínimo	-	2,0	0,0	0,0
Desvio Padrão	-	1,5	5,4	7,6

Atributo ⇒	AT	S	Argila	AMG	AG	AM	AF	AMF
Média	82,4	11,3	6,0	2,4	7,9	7,2	14,7	50,1
Máximo	98,8	33,2	24,8	12,0	42,3	34,2	36,6	92,2
Mínimo	48,9	0,0	0,0	0,0	0,0	0,0	6,3	2,7
Desvio Padrão	16,1	10,3	7,0	3,6	12,4	9,3	7,6	34,7

Atributo ⇒	N	P	K	Ca	Mg	Al	Na	C	Mat Org2	pH H_2O	pH KCl	SGH
Média	0,2	1,0	0,9	0,3	0,6	1,4	0,9	2,2	3,7	4,9	3,9	-
Máximo	0,5	1,8	2,4	1,6	1,8	3,8	3,9	5,9	10,1	5,9	4,7	-
Mínimo	0,0	0,3	0,1	0,0	0,0	0,1	0,1	0,4	0,6	4,0	3,1	-
Desvio Padrão	0,1	0,4	0,7	0,3	0,5	1,0	1,1	1,2	2,0	0,4	0,3	-

Anexo 2: Escalas ordenadas para atributos discretos

Composição Mineral

Cód.	Descrição original	Ordem
1		
2	Argila	0
3	Argila sub quartzo	1
24	Quartzo, mica e argila	2
6	Predomínio de quartzo e muita mica subordinada	3
4	Pred.quartzo;sub.mica	4
9	Predomínio de quartzo;subordinadamente mica	4
31	Subordinado mica	4
10	Predomínio de quartzo;subordinadamente mica (grãos maiores)	5
14	Quartzo e sub quantidade relativa alta de muscovita, aumentando	5
17	Quartzo sub +++musc	5
18	Quartzo sub ++musc	6
28	Quartzo, sub ++musc	6
12	Quartzo com biotita e muscovita aumentando em profundidade	7
19	Quartzo sub +musc	7
13	Quartzo e sub muscovita	8
20	Quartzo sub biotita	8
21	Quartzo sub musc	8
7	Predomínio de quartzo, com pouca mica sub.; biotita e muscuvita (grãos	9
8	Predomínio de quartzo, muito pouca mica	9
23	Quartzo, com mica e feldspato subordinado; em maiores proporções	9
26	Quartzo, sub + musc e felds	9
29	Quartzo, sub ++musc felds	9
22	Quartzo+Feuds+Mica	10
25	Quartzo, mica e feldspato	10
27	Quartzo, sub ++felds, musc	10
16	Quartzo m feldspato sub com aumento de biotita em profundidade	11
15	Quartzo m feldspato sub	12
30	Subordinado feldspato	12
5	Predomínio de quartzo	15
11	Quartzo	20

Textura

Cód.	Descrição	Ordem
1		
27	Sem dados	
7	Argiloso, forte agreg arg	0
8	Argiloso; forte agreg	1
2	Arg fina; forte agreg	2
3	Arg. Amf fraca agregação	3
5	Argilosa com material endurecido	4
6	Argiloso a muito fino	5
4	Argilo-arenosa	6
22	Muito fina argilosa	7
24	Muito fina, sub argiloso	8
20	Muito fina a argiloso	9
23	Muito fina, agreg fr arg	10
21	Muito fina agreg	11
19	Muito fina	12
11	Fina a muito fina	13
13	Fina arg; média agreg	14
12	Fina arg; fraca agreg	15
14	Fina com agregados. Teor de agregad decr. Até 90cm	16
9	Fina	17
10	Fina a média	18
16	Média	19
18	Média grossa	20
26	Sem agregados, textura média	21
17	Média a grossa	22
15	Grossa	23
25	Muito grossa	24

Grau de Seleção

Cód.	Descrição	Ordem
1	Muito mal selecionada	0
6	Mal selecionado a muito mal selecionado com aumento em prof	1
2	Mal selecionada	2
7	Média seleção piorando com o aumento da profundidade	3
3	Medianamente selecionada	4
4	Bem selecionada	5
5	Muito bem selecionada	6