Tratamento Bayesiano de Interações entre Atributos de Alta Cardinalidade

Este exemplar corresponde à redação final da Tese devidamente corrigida e defendida por Jorge Eduardo de Schoucair Jambeiro Filho e aprovada pela Banca Examinadora.

Campinas, 19 de setembro de 2007.

Jacques Wainer (Orientador)

Tese apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Bibliotecária: Maria Júlia Milani Rodrigues – CRB8a / 2116

Jambeiro Filho, Jorge Eduardo de Schoucair

J226t Tratamento bayesiano de interações entre atributos de alta cardinalidade / Jorge Eduardo de Schoucair Jambeiro Filho -- Campinas, [S.P. :s.n.], 2007.

Orientador : Jacques Wainer

Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Computação.

1. Inteligência artificial. 2. Teoria bayesiana de decisão estatística. 3. Aprendizado do computador. I. Wainer, Jacques. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Handling interactions among high cardinality attributes.

Palavras-chave em inglês (Keywords): 1. Artificial intelligence. 2. Bayesian statistical decision theory. 3. Machine learning.

Área de concentração: Sistemas de Informação

Titulação: Doutor em Ciência da Computação

Banca examinadora: Prof. Dr. Fabio Gagliardi Cozman (Poli-USP)
Prof. Dr. André Ponce de Leon F. de Carvalho (ICMC-USP)
Prof. Dr. Siome Klein Goldenstein (IC-UNICAMP)
Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho (IC-UNICAMP)

Data da defesa: 07/11/2007

Programa de Pós-Graduação: Doutorado em Ciência da Computação

TERMO DE APROVAÇÃO

Tese Defendida e Aprovada em 07 de novembro de 2007, pela Banca examinadora composta pelos Professores Doutores:

Prof. Dr. Fábio Gagliardi Cozman
EP - USP.

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
ICMC - USP.

Prof. Dr. Siome Klein Goldenstein
IC - UNICAMP.

Profª. Drª. Ariadne Maria Brito Rizzoni Carvalho
IC - UNICAMP.

Prof. Dr. Jacques Wainer
IC - UNICAMP.

Instituto de Computação
Universidade Estadual de Campinas

Tratamento Bayesiano de Interações entre Atributos de Alta Cardinalidade

Jorge Eduardo de Schoucair Jambeiro Filho

Setembro de 2007

Banca Examinadora:

- Jacques Wainer (Orientador)
- Fabio Gagliardi Cozman
  Escola Politécnica da Universidade de São Paulo
- André Ponce de Leon F. de Carvalho
  Instituto de Ciências Matemáticas e de Computação de Universidade de São Paulo
- Siome Klein Goldenstein
  Instituto de Computação da Universidade de Campinas
- Ariadne Maria Brito Rizzoni Carvalho
  Instituto de Computação da Universidade de Campinas

Resumo

Analisamos o uso de métodos Bayesianos em um problema de classificação de padrões de interesse prático para a Receita Federal do Brasil que é caracterizado pela presença de atributos de alta cardinalidade e pela existência de interações relevantes entre eles. Mostramos que a presença de atributos de alta cardinalidade pode facilmente gerar tantas subdivisões no conjunto de treinamento que, mesmo tendo originalmente uma grande quantidade de dados, acabemos obtendo probabilidades pouco confiáveis, inferidas a partir de poucos exemplos. Revisamos as estratégias usualmente adotadas para lidar com esse problema dentro do universo Bayesiano, exibindo sua dependência em suposições de não interação inaceitáveis em nosso domínio alvo. Mostramos empiricamente que estratégias Bayesianas mais avançadas para tratamento de atributos de alta cardinalidade, como pré-processamento para redução de cardinalidade e substituição de tabelas de probabilidades condicionais (CPTs) de redes Bayesianas (BNs) por tabelas default (DFs), árvores de decisão (DTs) e grafos de decisão (DGs) embora tragam benefícios pontuais não resultam em ganho de desempenho geral em nosso domínio alvo. Propomos um novo método Bayesiano de classificação, chamado de *hierarchical pattern Bayes* (HPB), que calcula probabilidades posteriores para as classes dado um padrão W combinando as observações de W no conjunto de treinamento com probabilidades prévias que são obtidas recursivamente a partir das observações de padrões estritamente mais genéricos que W. Com esta estratégia, ele consegue capturar interações entre atributos de alta cardinalidade quando há dados suficientes para tal, sem gerar probabilidades pouco confiáveis quando isso não ocorre. Mostramos empiricamente que, em nosso domínio alvo, o HPB traz benefícios significativos com relação a redes Bayesianas com estruturas populares como o *naïve Bayes* e o *tree augmented naïve Bayes*, com relação a redes Bayesianas (BNs) onde as tabelas de probabilidades condicionais foram substituídas pelo *noisy-OR*, por DFs, por DTs e por DGs, e com relação a BNs construídas, após uma fase de redução de cardinalidade usando o *agglomerative information bottleneck*. Além disso, explicamos como o HPB, pode substituir CPTs e mostramos com testes em outro problema de interesse prático que esta substituição pode trazer ganhos significativos. Por fim, com testes em vários conjuntos de dados públicos da UCI, mostramos que a utilidade do HPB ser bastante ampla.

Abstract

In this work, we analyze the use of Bayesian methods in a pattern classification problem of practical interest for Brazil's Federal Revenue which is characterized by the presence of high cardinality attributes and by the existence of relevant interactions among them. We show that the presence of high cardinality attributes can easily produce so many subdivisions in the training set that, even having originally a great amount of data, we end up with unreliable probability estimates, inferred from small samples. We cover the most common strategies to deal with this problem within the Bayesian universe and show that they rely strongly on non interaction assumptions that are unacceptable in our target domain. We show empirically that more advanced strategies to handle high cardinality attributes like cardinality reduction by preprocessing and conditional probability tables replacement with default tables, decision trees and decision graphs, in spite of some restricted benefits, do not improve overall performance in our target domain. We propose a new Bayesian classification method, named hierarchical pattern Bayes (HPB), which calculates posterior class probabilities given a pattern W combining the observations of W in the training set with prior class probabilities that are obtained recursively from the observations of patterns that are strictly more generic than W. This way, it can capture interactions among high cardinality attributes when there is enough data, without producing unreliable probabilities when there is not. We show empirically that, in our target domain, HPB achieves significant performance improvements over Bayesian networks with popular structures like naïve Bayes and tree augmented naïve Bayes, over Bayesian networks where traditional conditional probability tables were substituted by noisy-OR gates, default tables, decision trees and decision graphs, and over Bayesian networks constructed after a cardinality reduction preprocessing phase using the agglomerative information bottleneck method. Moreover, we explain how HPB can replace conditional probability tables of Bayesian Networks and show, with tests in another practical problem, that such replacement can result in significant benefits. At last, with tests over several UCI datasets we show that HPB may have a quite wide applicability.

Agradecimentos

Agradeço a minha esposa por ter suportado a minha ausência pelo longo tempo tomado pela escrita desta tese e a nossos cães por lhe ter feito companhia em meu lugar.

Agradeço a meus pais por terem me dado a educação que tornou possível um dia fazer um trabalho como este. Também agradeço a meus professores e colegas de classe que me ensinaram em diferentes etapas da minha vida acadêmica.

Também agradeço a meu orientador, Jacques, por me introduzir ao universo Bayesiano, que se tornou a base não apenas desta tese, mas de quase todo o meu trabalho na Receita Federal do Brasil.

Agradeço a Marcos por ter feito a dissertação de mestrado que apresentou o problema que serve como motivação para esta tese e a Antonella por ter criado o projeto Harpia dentro do qual a tese foi desenvolvida.

Por fim, agradeço a meus colegas do projeto Harpia, AFRFs, professores, alunos e pós-docs pelos vários debates que me ajudaram e ajudam a desenvolver meu conhecimento de inteligência artificial.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A Receita Federal do Brasil (RFB) não possui auditores suficientes para acompanhar todos os processos de importação. Conseqüentemente, apenas uma amostra das mercadorias importadas é verificada por especialistas humanos. A administração da RFB não pretende aumentar o contingente de auditores empenhados na tarefa de conferir processos de importação, mas sim reduzir, ainda mais, os percentuais de conferência. É evidente a necessidade de adoção de um sistema de gerenciamento de risco que proporcione a seleção de operações com maior probabilidade de fraude para conferência, de modo a permitir esta redução sem comprometer a segurança dos controles aduaneiros [50].

A legislação brasileira determina que todas as mercadorias importadas em regime normal sejam enquadradas em uma das cerca de 9700 posições de uma tabela denominada Nomenclatura Comum do Mercosul (NCM). É muito importante que esta classificação, chamada de *classificação fiscal*, esteja correta, pois ela determina as alíquotas dos impostos que incidirão sobre a importação, e define a aplicação de diferentes exigências administrativas, sanitárias, militares e de segurança. Por esse motivo, embora sejam várias as infrações que podem ser cometidas em um procedimento de importação, tomamos como motivação para este trabalho apenas o erro de classificação fiscal.

O domínio da detecção desses erros de classificação envolve atributos que podem assumir muitos valores distintos (alta cardinalidade). Os principais atributos já vem sendo empregados pela RFB desde o trabalho realizado por Ferreira [32]. Eles são:

- *classificação fiscal declarada* (NCMD): a posição da NCM na qual o importador declara que uma mercadoria se enquadra. Esse atributo pode assumir cerca de 9700 valores distintos.

- *importador* (IMP): o identificador (CPF/CNPJ) do importador. No momento, esse atributo pode assumir cerca de 20000 valores distintos.

- *país de origem* (PAIS): o país onde a mercadoria foi produzida. Esse atributo pode assumir 194 valores distintos.
- *unidade de entrada* (URF): unidade alfandegária da RFB através da qual a mercadoria ingressou no Brasil. Esse atributo pode assumir mais de 100 valores distintos.

Temos o objetivo de desenvolver uma ferramenta que, considerando os atributos citados, estime para cada exemplo novo a probabilidade de que ele envolva um erro de classificação. Estas estimativas serão usadas por um sistema maior que aloca recursos humanos para diferentes tipos de operações anti-fraude[1].

Nosso principal conjunto de dados tem 682226 exemplos de classificação correta (que chamamos de exemplos negativos) e 6460 exemplos de erros de classificação (exemplos positivos). O conjunto de dados é desbalanceado, com apenas 0.93% de exemplos positivos. Dados com esta característica normalmente são tratados com diferentes estratégias de *resampling* [14]. Porém, em geral, *resampling* requer o retreinamento dos classificadores para diferentes atribuições de custo para falsos positivos e falsos negativos. No nosso contexto, tais custos não são conhecidos a priori (as prioridades mudam de acordo com outras demandas de combate a fraude) e podem variar de exemplo para exemplo (nem todos os falsos negativos tem o mesmo custo). Isso torna o uso de *resampling* pouco atraente.

Por outro lado, se pudermos produzir estimativas de probabilidade confiáveis diretamente a partir do conjunto de dados original, o trabalho do alocador de recursos se tornará bem mais fácil. A qualquer momento, após considerar todas as outras demandas de combate a fraude, ele poderá definir uma taxa de seleção correspondente aos recursos humanos disponíveis para a tarefa específica de detectar erros de classificação fiscal. Se a taxa de seleção for de, por exemplo, 10%, as mercadorias a ser verificadas serão as 10% que tiverem maior probabilidade de conter um erro de classificação de acordo com as estimativas produzidas. Alternativamente, o sistema de alocação de recursos poderá combinar as estimativas de probabilidade produzidas com custos que variam de exemplo para exemplo sem qualquer retreinamento.

Com o objetivo de produzir boas estimativas de probabilidade em mente, decidimos nos concentrar nas técnicas que têm nas probabilidades a sua essência, as técnicas Bayesianas.

Especialistas da Receita Federal sabem que certas combinações de valores de atributos formam padrões que tornam a probabilidade de presença de erro significativamente

[1]Desde o início do trabalho de doutorado que resultou na presente tese, a RFB, através do projeto Harpia, que envolve a Universidade de Campinas e o Instituto Tecnológico da Aeronáutica, vem intensificando o uso de sistemas de inteligência artificial. Hoje sistemas sendo desenvolvidos no projeto Harpia utilizam processamento de informação textual, aprendizado não supervisionado, redes de relacionamento e sistemas especialistas na detecção de vários tipos de infração, incluindo o erro de classificação fiscal. Contudo, tratar os quatro atributos de alta cardinalidade listados acima permanece como um importante subproblema e toda esta tese se foca neste ponto.

maior que o que poderia ser esperado caso os atributos fossem observados separadamente. Chamamos esses padrões de *padrões críticos*.

O número de padrões críticos é desconhecido e variável. Além disso, nenhum especialista individualmente conhece um grande percentual deles. Por isso, eles precisam ser aprendidos de forma automática.

O conhecimento da existência de padrões críticos indica que a adequada captura das interações entre atributos pode trazer ganhos de desempenho. Entretanto, sabemos que esta tarefa não é trivial, pois, em um trabalho sobre aplicação de redes Bayesianas (descritas na Seção 2.6) na seleção de declarações de importação para conferência, Ferreira [32] não pode levá-las em consideração.

Para aproveitar os padrões críticos gostaríamos de usar uma rede Bayesiana (BN) [63] com a estrutura apresentada na Figura 1.1, onde todos os atributos explanatórios são pais do atributo classe.

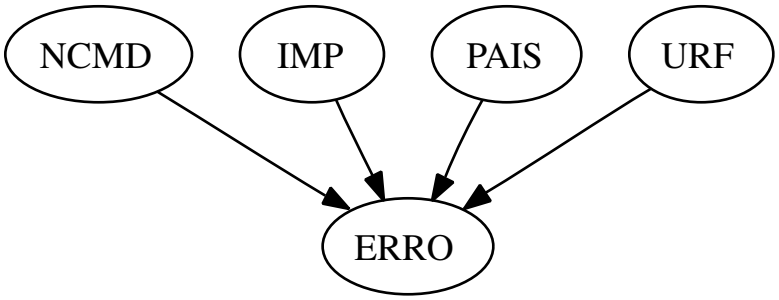

Figura 1.1: estrutura causal direta para erros de classificação fiscal

Em uma rede Bayesiana, considerando que x_{ji} é um possível valor para o nó X_j e π_{jk} é uma combinação completa de valores para Π_j, o conjunto de pais do nó X_j, o vetor, θ_{jk}, tal que $\theta_{jki} = P(x_{ji}|\pi_{jk})$, é armazenado em uma tabela chamada de *tabela de probabilidades condicionais* (CPT[2]) do nó X_j e é inferido a partir das freqüências dos valores de X_j entre as instâncias de treinamento onde $\Pi_j = \pi_{jk}$. As distribuições de X_j dadas duas combinações de valores para Π_j são consideradas independentes e uma distribuição de probabilidade prévia de Dirichlet para θ_{jk} é normalmente adotada. Aplicando-se a regra de Bayes e integrando-se ao longo de todos os valores possíveis para θ_{jk} encontra-se a relação

$$E(\theta_{jki}) = P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + \alpha_{jki}}{N_{jk} + \alpha_{jk}}, \tag{1.1}$$

onde N_{jki} é o número de observações simultâneas de x_{ji} e π_{jk} no conjunto de treinamento,

[2]Nesta tese, mantivemos as siglas com origem na língua inglesa em sua forma original, como pode ser visto na Tabela A.1

$N_{jk} = \sum_{\forall i} N_{jki}$ é o tamanho da amostra, α_{jki} é o valor de um dos parâmetros da distribuição de probabilidade prévia de Dirichlet e $\alpha_{jk} = \sum_{\forall i} \alpha_{jki}$ é o tamanho de amostra equivalente da distribuição de probabilidade prévia.

Normalmente, a distribuição de probabilidade prévia de Dirichlet adotada é não informativa, logo

$$P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + \lambda}{N_{jk} + \lambda M_j}, \tag{1.2}$$

onde todos os parâmetros da distribuição de Dirichlet são iguais a uma pequena constante de suavização, λ, e M_j é o número de valores possíveis para o nó X_j. Nós chamamos esse procedimento de *estimativa direta* (DE).

A tabela de probabilidades condicionais do nó classe de uma rede Bayesiana com a estrutura apresentada na Figura 1.1 conteria mais de 1.8×10^{12} linhas. É claro que para combinações raras de atributos tal estrutura em conjunto com a Equação (1.2) tenderia a produzir probabilidades não confiáveis cujas estimativas seriam dominadas pela distribuição de probabilidade prévia não informativa.

Redes Bayesianas com tabelas de probabilidades condicionais muito grandes quando comparadas aos conjuntos de treinamento são um exemplo de modelo de representação do comportamento de um conjunto de dados com excesso de parâmetros. Modelos com parâmetros demais, tendem a ter bom desempenho quando testados sobre as próprias instâncias de treinamento, mas desempenho insuficiente sobre instâncias novas [38]. Esse problema é denominado super-ajuste [59].

Uma alternativa para contornar esse problema é a adoção de estruturas de rede onde o número máximo de pais por nó seja pequeno, como o *naïve Bayes* [27] e o *tree augmented naïve Bayes* [34]. No entanto, limitar artificialmente o número de pais por nó limita a capacidade de representação da rede [56] tornando-a menos capaz de refletir interações entre atributos.

Uma estratégia natural para lidar com atributos de alta cardinalidade é agrupar valores considerados similares, tornando, assim, a cardinalidade mais baixa. O agrupamento de valores pode ser feito de forma automática buscando valores que tenham influência semelhante sobre o atributo classe de acordo com algum critério. Porém, em geral, estes critérios tratam cada atributo de forma isolada [58, 72, 9], e não podemos esperar que tragam benefícios significativos quanto a captura de padrões críticos, que dependem simultaneamente de vários atributos. Alternativamente, o agrupamento pode ser definido por regras que explorem a estrutura dos atributos. Por razões que apresentaremos no contexto das abordagens hierárquicas, também não desejamos seguir este caminho.

Quando o número de parâmetros a ser estimados para preenchimento das CPTs é muito grande se comparado ao número de instâncias de treinamento, Pearl [63] recomenda sua substituição por modelos que assumam não interação entre as variáveis e com isso

empreguem menos parâmetros, como, por exemplo, o *noisy-OR* [39]. Foi exatamente isso o que foi feito por Ferreira [32] na construção de uma rede Bayesiana para seleção de declarações de importação para conferência.

O número de parâmetros requeridos pelo *noisy-OR* é proporcional não ao produto da cardinalidade dos pais do nó ao qual está associado, mas à sua soma. Com isso, usando a estrutura na Figura 1.1 e o *noisy-OR*, apenas cerca de 30300 parâmetros precisam ser estimados. Um número aceitável considerando os conjuntos de dados disponíveis.

Por outro lado, como o *noisy-OR* assume a ausência de interação entre as variáveis, ele não tem a capacidade de capturar o efeito de padrões críticos.

Uma alternativa é empregar modelos mais flexíveis em lugar de CPTs ou do *noisy-OR* como tabelas default (DFs) [37], árvores de decisão (DTs) [37] e grafos de decisão (DGs) [16]. De acordo com Friedman e Goldszmidt [37], o uso desses modelos em conjunto com procedimentos de aprendizado adequados induz estruturas de rede mais complexas em termos de arcos, mas que requerem menos parâmetros e assim emulam melhor a real complexidade das interações presentes nos dados. Mais arcos significam mais pais por nó, o que aumenta a capacidade de captura de interações. Menos parâmetros significam menos problemas com probabilidades poucos confiáveis.

Usando CPTs, assumimos que as distribuições de probabilidade condicionais (CPDs) de um nó dadas duas atribuições distintas de valores para seus pais, são independentes entre-si. Se algumas destas distribuições forem, na verdade, idênticas, DTs, DFs e DGs podem refletir esta situação e representar a CPD do nó dados os seus pais usando um número de parâmetros que é proporcional apenas ao número de distribuições verdadeiramente distintas.

Usando DTs, DFs ou DGs para representar a distribuição de probabilidade condicional de um nó dados os seus pais, assumimos que as CPDs de um nó dadas duas atribuições distintas de valores para seus pais são ou completamente independentes ou idênticas. É possível que nenhuma das duas suposições seja verdadeira.

Gelman et al. [38] afirmam que usar modelos não hierárquicos para representar dados hierárquicos leva a resultados pobres. Com poucos parâmetros, eles não conseguem se ajustar aos dados de forma precisa. Com muitos parâmetros, eles se ajustam bem aos dados existentes, mas levam a previsões inferiores para dados novos. Em outras palavras eles sofrem de problemas de super-ajuste. Em contraste, modelos hierárquicos podem se ajustar bem aos dados sem incorrer em super-ajuste. Eles podem refletir similaridades entre distribuições de probabilidade sem assumir igualdade.

Os atributos de nosso problema alvo possuem algum tipo de hierarquia natural: importadores podem ser agrupados por área de atividade econômica, classificações fiscais podem ser agrupadas por capítulo, países por continente e unidades da Receita Federal por região fiscal. Estas hierarquias poderiam ser exploradas usando técnicas como a

apresentada em [71]. Porém, nem o conhecimento de especialistas, nem alguns primeiros experimentos indicaram que estas hierarquias naturais tenham um potencial, quanto a melhorias nas previsões a respeito de erros de classificação fiscal, elevado o suficiente para justificar sua exploração imediata.

Neste trabalho, mostramos como construir uma hierarquia entre padrões formados por valores de atributos não naturalmente hierárquicos, e aplicamos um novo modelo Bayesiano à hierarquia construída. Chamamos esse modelo de *hierarchical pattern Bayes* (HPB).

O HPB calcula probabilidades posteriores para as classes dado um padrão W combinando as observações de W no conjunto de treinamento com probabilidades prévias que são obtidas recursivamente a partir das observações de padrões estritamente mais genéricos que W. Com isso, ele consegue capturar interações entre atributos de alta cardinalidade quando há dados suficientes para tal, sem gerar probabilidades pouco confiáveis quando isso não ocorre.

Mostramos empiricamente que, em nosso domínio alvo, o HPB traz benefícios significativos com relação a redes Bayesianas com estruturas populares como o *naïve Bayes* e o *tree augmented naïve Bayes*, com relação a redes Bayesianas (BNs) onde as tabelas de probabilidades condicionais foram substituídas pelo *noisy-OR*, por DFs, por DTs e por DGs, e com relação a BNs construídas, após uma fase de pré-processamento para redução de cardinalidade usando o *agglomerative information bottleneck*. Além disso, explicamos como o HPB, pode substituir CPTs e mostramos com testes em outro problema de interesse prático que esta substituição pode trazer ganhos significativos. Por fim, com testes em vários conjuntos de dados da UCI, mostramos que a utilidade do HPB ser bastante ampla.

Modelos Bayesianos hierárquicos são amplamente utilizados na comunidade de marketing [3, 53] sob o nome de *hierarchical Bayes*. Eles também tem sido empregados em outras áreas como a medicina [4] e a robótica [74]. Contudo, não temos conhecimento de nenhum desses modelos seja aplicável a problemas de classificação de padrões envolvendo interações entre atributos de alta cardinalidade. Além disso, o HPB difere de outros modelos por lidar com uma hierarquia de múltiplos níveis recursivamente e também por lidar com o fato de que a população de instâncias associada a cada padrão está contida em várias e não em apenas uma superpopulação.

No Capítulo 2, apresentamos a fundamentação teórica desta tese. O leitor familiarizado com técnicas Bayesianas pode preferir saltar diretamente para o Capítulo 3, onde apresentamos o modelo hierárquico criado neste trabalho de doutoramento. Por outro lado, o Capítulo 2 apresenta uma revisão das técnicas Bayesianas mais utilizadas na área de inteligência artificial e pode ser útil como uma introdução mesmo para o leitor que não esteja interessado em nenhuma técnica nova.

No Capítulo 3, mostramos o HPB, descrevendo a construção de sua hierarquia de padrões e a forma como ele utiliza as distribuições de probabilidade posteriores de Dirichlet relativas aos níveis mais genéricos da hierarquia, na geração de distribuições de probabilidade prévias de Dirichlet informativas para seus níveis mais específicos.

No Capítulo 4, apresentamos resultados experimentais e no Capítulo 5, mostramos nossas conclusões e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, apresentamos uma revisão das técnicas Bayesianas mais utilizadas na área de inteligência artificial. Mostramos como inferir os parâmetros de uma distribuição multinomial a partir de amostras planas ou hierárquicas e como aplicar os mecanismos de inferência descritos a problemas de classificação de padrões.

Indicamos as dificuldades encontradas nesta aplicação quando atributos de alta cardinalidade são envolvidos, as estratégias usuais para contorno destas dificuldades e as limitações das estratégias indicadas.

Na Seção 2.1, mostramos como estimar os parâmetros de uma distribuição multinomial a partir de eventos, apresentando a abordagem freqüentista tradicional, denominada máxima verossimilhança, e a abordagem Bayesiana para o mesmo problema. Na abordagem Bayesiana descrevemos o uso de distribuições de probabilidade prévia de Dirichlet. Posteriormente (Capítulo 3), o leitor perceberá que a obtenção de distribuições de probabilidade prévia de Dirichlet informativas é nossa estratégia central para geração de probabilidades confiáveis.

Na Seção 2.2, mostramos modelos Bayesianos que utilizam distribuições de probabilidade prévia hierárquica. Esses modelos de dois níveis têm a capacidade de se ajustar bem a conjuntos de dados aos quais um modelo plano ou não conseguiria ajustar-se ou, ao fazê-lo, produziriam probabilidades pouco confiáveis para dados novos [38].

Na Seção 2.3, descrevemos o problema de classificação de padrões e a forma com que a explosão combinatória entre os possíveis valores para os atributos discretos reduz a confiabilidade das probabilidades inferidas a partir do conjunto de treinamento.

Na Seção 2.4, descrevemos um modelo Bayesiano que supõe independência entre todos os atributos dada a classe, o *naïve Bayes*. Na Seção 2.5, mostramos que o *naïve Bayes* é um exemplo de modelo de não interação e descrevemos o *noisy-OR*, um outro modelo de não interação que foi utilizado em [32].

Na Seção 2.6, descrevemos as redes Bayesianas, um mecanismo que permite a utilização

de suposições de independência na medida da conveniência do problema. Mostramos também que, ao tratar variáveis discretas, estas redes, em geral, utilizam tabelas de probabilidades que crescem a medida em que mais dependências são consideradas e que o crescimento destas tabelas leva a probabilidades pouco confiáveis.

2.1 Probabilidades a partir de experimentos com resultados discretos

Suponhamos que um experimento v possua M resultados possíveis representados pelo vetor $V = V_1, V_2, \ldots, V_M$, com probabilidades desconhecidas representadas pelo vetor $\theta = \theta_1, \theta_2, \ldots, \theta_M$ respectivamente. Suponhamos ainda que em uma amostra de N execuções independentes do experimento, os números de observações dos resultados sejam dados pelo vetor $K = K_1, K_2, \ldots, K_M$. Se quisermos estimar a probabilidade de que o resultado de um novo experimento independente seja V_i, por definição precisamos estimar o valor de θ_i. Existem duas abordagens comuns para fazer esta estimativa. Ambas baseiam-se no fato de que $P(K|\theta)$ segue uma distribuição multinomial:

$$P(K|\theta) = Multinomial(K, \theta) = \frac{N!}{\prod_{i=1}^{M} K_i!} \prod_{i=1}^{M} \theta_i^{K_i}.$$

2.1.1 Abordagem de máxima verossimilhança

A primeira abordagem, conhecida como freqüentista, ou de máxima verossimilhança [33], consiste em escolher para o vetor θ o valor que maximiza a probabilidade de termos observado na amostra aquilo que efetivamente observamos. Assim,

$$\begin{aligned}
\theta^{ML} &= \arg\max_{\theta} \left(P(K|\theta)\right) \\
\theta^{ML} &= \arg\max_{\theta} \left(\frac{N!}{\prod_{i=1}^{M} K_i!} \prod_{i=1}^{M} \theta_i^{K_i} \right) \\
\theta^{ML} &= \arg\max_{\theta} \left(\prod_{i=1}^{M} \theta_i^{K_i} \right), \qquad (2.1)
\end{aligned}$$

onde θ^{ML} é o vetor θ estimado via máxima verossimilhança. O valor de θ^{ML} calculado a partir da Equação 2.1 é sabidamente dado por

$$\theta_j^{ML} = \frac{K_j}{N} = \frac{K_j}{\sum_{i=1}^{M} K_i}, \quad \forall j.$$

A abordagem de máxima verossimilhança funciona bem se N for grande. Quando o tamanho da amostra tende a infinito, as estimativas de máxima verossimilhança tendem às probabilidades reais, ou seja, a abordagem de máxima verossimilhança é assintoticamente correta.

Por outro lado, se N for pequeno, ela leva a resultados de qualidade muito ruim. Em particular, se $K_i = 0$, θ_i é estimado como sendo também igual a zero, o que significa que o resultado V_i é considerado impossível. Intuitivamente, nenhum número finito de execuções do experimento poderia nos dar certeza de que o evento V_i não pode ocorrer. Se N fosse grande, um valor igual a zero para K_i significaria que o resultado V_i é, se não impossível, pelo menos muito improvável. Porém, sendo N pequeno, como por exemplo $N = 1$, considerar que $\theta_i = 0$ não faz qualquer sentido.

Uma exceção ao problema acima é quando estamos estimando não a probabilidade de que uma nova execução do experimento resulte no valor V_i, mas sim a probabilidade de que escolhendo ao acaso uma das execuções do experimento já realizadas encontremos um resultado igual a V_i. Nesse caso, $K_i = 0$ realmente torna impossível que venhamos a encontrar o valor V_i e a abordagem de máxima verossimilhança nos dá precisamente esse resultado.

É fácil perceber que usando máxima verossimilhança, os resultados obtidos para previsões dentro do conjunto de dados original (que chamaremos de conjunto de treinamento) são exatos. O ajuste da abordagem de máxima verossimilhança à amostra é portanto completo. Como já vimos, através do exemplo em que $K_i = 0$, estimar probabilidades de forma a refletir exatamente as observações presentes na amostra pode levar a resultados irrealistas para novas execuções do experimento. Esse ajuste excessivo é chamado de super-ajuste.

Mitchell [59] fornece uma definição formal para super-ajuste que é amplamente adotada na literatura [48, 25, 62, 31, 55].

Definição 2.1 *Dado um espaço de hipóteses H, diz-se que a hipótese $h \in H$ está cometendo super-ajuste ao conjunto de treinamento se existir uma hipótese h' pertencente a H tal que h tem erro menor que h' sobre os exemplos de treinamento e h' tem um erro menor que h sobre a distribuição completa de instâncias [59].*

O conjunto de treinamento referido na definição 2.1 é simplesmente a amostra de experimentos cujos resultados são conhecidos.

Uma definição menos formal, porém bastante popular é a que diz que super-ajuste é o efeito do ajuste de um modelo estatístico que tem um número excessivo de parâmetros. Esta definição consta, por exemplo, na Wikipedia [76]. Contudo, esta definição estabeleceria, em definitivo, a causa do problema que passaria a admitir uma única solução: a redução do número de parâmetros. Por isso, neste trabalho, adotamos a definição 2.1.

Quando um modelo probabilístico sofre de super-ajuste, ele resulta em probabilidades *pouco confiáveis*, no sentido de que seu desempenho, quando medido no conjunto de treinamento, é excelente, mas quando medido sobre dados novos (a medida que de fato interessa) é decepcionante.

2.1.2 Abordagem Bayesiana

Podemos estimar a probabilidade de que o resultado de uma nova execução de um experimento seja V_i dadas as observações de N execuções independentes do mesmo experimento de forma Bayesiana. Sabemos que $P(V_i|K) = \theta_i$, mas o vetor θ é desconhecido. Assim, devemos calcular a expectativa de θ_i dadas as observações do experimento sintetizadas pelo vetor K:

$$P(V_i|K) = E(\theta_i|K) = \int \theta_i P(\theta|K) d\theta. \tag{2.2}$$

Para calcular $E(\theta_i|K)$ precisamos de uma função que forneça a distribuição de probabilidade de θ dado K. Aplicando a regra de Bayes temos

$$P(\theta|K) = \frac{P(K|\theta)P(\theta)}{P(K)} = \frac{P(K|\theta)P(\theta)}{\int P(K|\theta)P(\theta)d\theta}. \tag{2.3}$$

Sabemos que $P(K|\theta)$ segue a distribuição multinomial, assim, para encontrar $E(\theta|K)$ temos apenas dois problemas: o primeiro é a escolha de uma distribuição $P(\theta)$, visto que esta é desconhecida, o segundo é a existência de solução analítica para $\int P(K|\theta)P(\theta)d\theta$ e $\int \theta_i P(\theta|K)d\theta$.

Sem que mais dados estejam disponíveis, o melhor que pode ser feito é escolher uma distribuição $P(\theta)$ que coincida aproximadamente com o conhecimento de especialistas a respeito da estrutura do problema e que ao mesmo tempo combine-se com a distribuição multinomial de forma a garantir a integrabilidade das funções de interesse. Antes de exibirmos uma família de distribuições de probabilidade conveniente vamos introduzir algum vocabulário.

Como $P(\theta|K)$ é a distribuição de probabilidade para θ após a observação do vetor K, ela é chamada de distribuição de probabilidade posterior para θ, enquanto que $P(\theta)$ é chamada de distribuição de probabilidade prévia para θ. Ao mesmo tempo, $P(V_i|K) = E(\theta_i|K)$ é chamada de probabilidade preditiva posterior para V_i e $P(V_i) = E(\theta_i)$ é chamada de probabilidade preditiva prévia para V_i.

As Distribuições de Dirichlet

Uma distribuição de probabilidade prévia muito conveniente para o problema que estamos abordando é uma Distribuição de Dirichlet [38, 57, 30]:

$$Dirichlet(\theta, \alpha) = \beta \prod_{i=1}^{M} \theta_i^{\alpha_i - 1},$$

onde o vetor $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_M$ é o vetor de parâmetros da distribuição e β é uma constante tal que

$$\beta \int \prod_{i=1}^{M} \theta_i^{\alpha_i - 1} d\theta = 1, \tag{2.4}$$

como é necessário a uma distribuição de probabilidade. Podemos encontrar o valor da constante β resolvendo a Equação (2.4). Com isso, temos que uma distribuição de Dirichlet é dada por

$$P(\theta) = Dirichlet(\theta, \alpha) = \frac{\Gamma\left(\sum_{i=1}^{M} \alpha_i\right)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} \prod_{i=1}^{M} \theta_i^{\alpha_i - 1}.$$

Uma distribuição de Dirichlet é conveniente em primeiro lugar porque se $P(\theta)$ segue a distribuição $Dirichlet(\theta, \alpha_1, \alpha_2, \ldots, \alpha_M)$ e $P(K|\theta)$ segue a distribuição $Multinomial(K_1, K_2, \ldots, K_M, \theta)$ então as integrais nas Equações (2.2) e (2.3) têm solução analítica e resultam em

$$E(\theta_j|\alpha) = \frac{\alpha_j}{\sum_{i=1}^{M} \alpha_i}, \quad \forall j, \tag{2.5}$$

e

$$P(\theta|K) = Dirichlet(\theta, \alpha_1 + K_1, \alpha_2 + K_2, \ldots, \alpha_M + K_M). \tag{2.6}$$

Note que a distribuição $P(\theta|K)$ ainda é uma distribuição de Dirichlet. Esse último fato evita que tenhamos que lidar com outra distribuição de probabilidade possivelmente mais complicada. Esta propriedade se torna mais importante se pretendermos atualizar uma distribuição de Dirichlet usando sucessivas observações de execuções de experimentos cujos resultados têm probabilidades que seguem a distribuição multinomial, pois sempre nos manteremos dentro da família das distribuições de Dirichlet.

Quando duas famílias de distribuições de probabilidade, f e F, são tais que se a distribuição de probabilidade dos dados (verossimilhança dos dados) pertence à família f e um membro de F é adotado como distribuição de probabilidade prévia, a distribuição

de probabilidade posterior é também um membro de F, diz-se que F é *conjugada* de f. A família das distribuições de Dirichlet é conjugada da família de distribuições multinomiais.

Um caso particular da distribuição multinomial é a distribuição binomial que corresponde ao caso onde o experimento tem apenas dois resultados possíveis. A distribuição de Dirichlet tem na distribuição Beta seu caso particular conjugado a distribuição binomial.

Uma distribuição de Dirichlet é bastante flexível. Variando os parâmetros α, nós podemos tornar a distribuição de Dirichlet consistente com quaisquer valores que acreditemos corresponder às probabilidades preditivas prévias de que o resultado da execução de um experimento seja V_1, V_2, ... ou V_M. Se acreditamos que as probabilidades preditivas prévias são $P(V_1), P(V_2), \ldots, P(V_M)$ podemos escolher os parâmetros de tal modo que $\alpha_i = S \cdot P(V_i)$ para todo i, onde S é uma constante positiva. Pela Equação (2.5), a esperança de θ corresponderá exatamente ao vetor de probabilidades preditivas prévias escolhidas.

Além disso, adotando uma distribuição de probabilidade prévia de Dirichlet, a partir de (2.5) e (2.6) temos que

$$P(V_j|K) = E(\theta_i|K) = \frac{K_j + \alpha_j}{\sum_{i=1}^{M} K_i + \sum_{i=1}^{M} \alpha_i} = \frac{K_j + \alpha_j}{N + N'}, \tag{2.7}$$

onde $N = \sum_{i=1}^{M} K_i$ e $N' = \sum_{i=1}^{M} \alpha_i$.

Os parâmetros da distribuição de Dirichlet podem ser interpretados como pseudo freqüências em uma seqüência de eventos hipoteticamente observada antes das execuções que levaram a obtenção do vetor de freqüências K. Esta interpretação tem um apelo intuitivo que facilita a construção das distribuições de Dirichlet por especialistas. Dentro desta interpretação, o valor $N' = \sum_{i=1}^{M} \alpha_i$ pode ser interpretado como o total de pseudo execuções do experimento e por isso esse valor é chamado de tamanho de amostra equivalente (TAE).

Se tivermos confiança em que a distribuição de probabilidade prévia está próxima da verdadeira distribuição de probabilidade, podemos usar uma amostra equivalente grande. Fazendo isso, obteremos probabilidades posteriores mais confiáveis, pois se N tender a infinito os dados reais dominarão a Equação (2.7) e as estimativas tenderão as probabilidades reais (como ocorre na abordagem de máxima verossimilhança), mas se N for pequeno o conhecimento prévio prevalecerá.

A adição de exemplos a amostras pequenas leva a variações abruptas em estimativas de máxima verossimilhança. As estimativas Bayesianas variam de forma mais suave como ocorreria com as estimativas de máxima verossimilhança caso a amostra tivesse tamanho igual a $N + N'$, ou seja um tamanho igual a soma dos tamanhos da amostra real e da

pseudo amostra. Por esse motivo, a adoção de distribuições de probabilidade prévia é considerada um mecanismo de suavização.

É freqüente a situação em que nada é sabido de antemão a respeito da distribuição de probabilidade. Nesse caso, costuma-se adotar distribuições de probabilidade prévias não informativas. Distribuições de probabilidade prévia não informativas são aquelas que tem *pouca* influencia sobre a distribuição de probabilidade posterior. No caso das distribuições de Dirichlet, faz sentido escolher um vetor de parâmetros α onde todos os parâmetros são iguais. Nesse caso, temos que

$$P(\theta) = Dirichlet(\theta, \lambda, \lambda, \ldots, \lambda) = \beta \prod_{i=1}^{M} \theta_i^{\lambda-1}$$

e

$$P(V_j|K) = \frac{K_j + \lambda}{N + \lambda \cdot M}. \tag{2.8}$$

A Equação (2.8) é conhecida como lei de sucessão de Lidstone [54]. Um caso particular importante ocorre quando $\lambda = 1$. Nesse caso,

$$P(\theta) = Dirichlet(\theta, 1, 1, \ldots, 1) = \beta \prod_{i=1}^{M} \theta_i^{0} = \beta.$$

Assim, $P(\theta)$ passa a ser constante e portanto temos que a distribuição uniforme é um caso particular de distribuição de Dirichlet. A adoção de distribuições de probabilidade prévia uniformes foi originalmente proposta por Laplace [52] e resulta na seguinte fórmula para a probabilidade preditiva posterior:

$$P(V_j|K) = \frac{K_j + 1}{N + M}. \tag{2.9}$$

Esta fórmula é conhecida como lei de sucessão de Laplace. Ela foi posteriormente generalizada por Lidstone, o que resultou na Equação (2.8).

Adotar uma distribuição de Dirichlet com todos os parâmetros iguais a uma constante não é o suficiente para considerá-la não informativa. Esta condição é suficiente para evitar desvios em direção a qualquer dos resultados, porém ela introduz uma tendência em direção a crença de que os resultados são todos igualmente prováveis. Se o tamanho da amostra equivalente for grande, isso impedirá que a distribuição de probabilidade posterior se ajuste adequadamente aos dados. Por isso, distribuições de Dirichlet não informativas sempre têm tamanhos de amostra equivalente pequenos.

O uso de distribuições de probabilidade prévia não informativas é adequado apenas quando a amostra real ainda é relativamente grande. Se a amostra for pequena ao ponto de tornar a distribuição de probabilidade prévia decisiva, então é melhor procurar uma distribuição de probabilidade prévia informativa [38]. Uma forma conseguir isso é usar um modelo Bayesiano hierárquico.

2.2 Estimando probabilidades usando uma amostra hierárquica

Como motivação para o uso de modelos hierárquicos, suponhamos que um certo fenômeno já tenha sido observado em L regiões distintas $R_1, R_2, \ldots, R_L$ e que o número de casos observados nestas regiões sejam respectivamente $N_1, N_2, \ldots, N_L$. Suponhamos ainda, que o fenômeno tenha M resultados possíveis $V_1, V_2, \ldots, V_M$ e que o número de observações de cada resultado V_i em cada região R_h seja K_{hi} e esteja disponível em uma matriz K. Respostas para duas questões são desejadas:

Questão 2.1 *Se o fenômeno é observado mais uma vez na região R_h, qual a probabilidade de que o resultado seja V_i?*

Questão 2.2 *Se o fenômeno é observado em uma nova região, qual a probabilidade de que o resultado seja V_i?*

A segunda pergunta pode ser vista como um caso particular da primeira. A nova região é apenas uma região onde o número de observações anteriores do fenômeno é zero.

Podemos representar a probabilidade de que o resultado da observação do fenômeno na região R_h seja V_i, por θ_{hi}. Assim, como ocorreu na Seção 2.1.2, tentaremos encontrar uma distribuição $P(\theta|K)$ (onde θ é uma matriz bidimensional e não mais um vetor) e calcular $E(\theta_{hi}|K)$ por integração.

2.2.1 Modelos planos

Uma forma simples de obter uma distribuição para $P(\theta|K)$ é aplicar o modelo descrito na Seção 2.1.2 tratando os fenômenos em regiões distintas como fenômenos distintos e adotando distribuições de probabilidade prévia independentes para cada região. A Equação (2.7) nos leva imediatamente a

$$P(V_j|K_h) = \frac{K_{hj} + \alpha_{hj}}{\sum_{i=1}^{M} K_{hi} + \sum_{i=1}^{M} \alpha_{hi}} = \frac{K_{hj} + \alpha_{hj}}{N_h + N'_h}, \tag{2.10}$$

onde K_h é o vetor de freqüências para região R_h. Note que as observações do fenômeno em uma região R_i em nada influência as probabilidades estimadas para uma região R_j, onde $j \neq i$. Por simplicidade, exibimos apenas a equação que calcula as probabilidades preditivas posteriores. A equação que exibe a distribuição de probabilidade posterior, $P(\theta_h|K_h)$, pode ser deduzida a partir da Equação (2.6).

Se para toda R_h, N_h for suficientemente grande, esta estratégia funcionará muito bem para responder a Questão 2.1. As distribuições de probabilidade prévia terão pouca relevância e as observações do fenômeno em uma região específica sob suas condições específicas dominarão as estimativas. Por outro lado, se tivermos amostras pequenas em cada região, teremos probabilidade pouco confiáveis. Uma situação extrema ocorre na tentativa de responder a Questão 2.2. Nesse caso, $N_h = 0$ e a probabilidade preditiva posterior torna-se igual a probabilidade preditiva prévia, um valor com pouca ou nenhuma conexão com a realidade caso tenhamos usado distribuições de probabilidade prévia não informativas.

É possível que, ainda que o número de observações do fenômeno em cada região seja pequeno, o número total de observações seja razoavelmente grande. Podemos ignorar a existência de regiões e aplicar Equação (2.7) à população global. Definindo que $N^g = \sum_{h=1}^{L} N_h$ e que $K_i^g = \sum_{h=1}^{L} K_{hi}$ temos

$$P(V_j|K^g) = \frac{K_j^g + \alpha_j}{\sum_{i=1}^{M} K_i^g + \sum_{i=1}^{M} \alpha_i} = \frac{K_j^g + \alpha_j}{N^g + N'^g}. \tag{2.11}$$

A Equação (2.11) reflete o fato de que não estamos falando de fenômenos distintos e que as observações em uma região devem afetar as demais. Outro efeito positivo é o fato de que novas regiões não são qualquer problema, já que a regiões são ignoradas. Por outro lado, perdemos a capacidade de fazer distinções entre as regiões e considerar suas condições específicas. Mesmo que inúmeras observações fenômeno em uma região R_h indiquem com clareza que um resultado V_i é muito mais provável em R_h que em outras regiões, nosso modelo não refletirá isso.

O uso da Equação (2.10), em conjunto com a adoção de distribuições de probabilidade prévia independentes, trata as manifestações do fenômeno em cada região como populações totalmente independentes. Já a Equação (2.11) coloca todas as manifestações do fenômeno em uma única população homogênea. Na verdade, temos uma grande população dividida em populações menores, formando uma hierarquia de populações. Assim, o uso das Equações (2.10) e (2.11) em conjunto com distribuições de probabilidade prévia não informativas ignoram a natureza hierárquica do problema. Elas correspondem a modelagens chamadas de não hierárquicas ou planas.

2.2.2 Modelos hierárquicos empíricos

Gelman et al. [38] afirmam que modelar dados hierárquicos de forma não hierárquica leva a resultados pobres. Com poucos parâmetros esses modelos não conseguem se ajustar

adequadamente aos dados. Com muitos parâmetros eles se ajustam bem aos dados existentes, mas fazem previsões inferiores para dados novos. Claramente o que está sendo dito é que os modelos planos tendem a sofrer problemas de super-ajuste.

Em conjunto com distribuições de probabilidade prévia não informativas, as Equações (2.10) e (2.11) correspondem respectivamente a um modelo com muitos e a outro com poucos parâmetros. Com a primeira, assumimos distribuições de probabilidades distintas (regidas por conjuntos de parâmetros distintos) em cada região. Com a segunda assumimos uma única distribuição de probabilidades para todas as regiões.

Podemos construir um modelo que reflita a hierarquia do problema considerando que em cada região a distribuição de probabilidade posterior é distinta, mas manter a conexão entre elas assumindo que todas as regiões tem a mesma distribuição de probabilidade prévia. Determinada uma distribuição de probabilidade prévia que reflita a população global, podemos seguir usando a Equação (2.10) para calcular probabilidades preditivas posteriores para cada região.

Uma forma de construir uma distribuição de probabilidade prévia que reflita a população global é usar a Equação (2.11) para, partindo de uma distribuição de probabilidade prévia não informativa, calcular probabilidades preditivas posteriores para a população global e adotá-las como probabilidades preditivas prévias para as subpopulações. Pode-se então adotar distribuições de probabilidade prévia informativas para as subpopulações que sejam consistentes com tais probabilidades preditivas prévias. Como vimos na Seção 2.1.2 é fácil construir uma distribuição de Dirichlet consistente com qualquer conjunto de probabilidades preditivas. No presente caso, basta definir que $\alpha_i = S \cdot P(V_i|K^g)$ para todo i, onde S é uma constante positiva. Isso nos leva imediatamente a

$$P(V_j|K_h) = \frac{K_{hj} + S \cdot P(V_j|K^g)}{N_h + S}. \tag{2.12}$$

Note que como $\sum_{i=1}^{n} P(V_i|K^g) = 1$, a constante S torna-se exatamente o tamanho da amostra equivalente. Esta estratégia é abordada em [21, 13, 35].

Observe que com a Equação (2.12), as grandes subpopulações têm o cálculo de suas probabilidades preditivas posteriores dominado por K_{hj}/N_h. Já nas subpopulações pequenas é $P(V_j|K^g)$ quem domina e quando $N_h = 0$ (novas regiões), as probabilidades preditivas posteriores tornam-se iguais as probabilidades preditivas prévias. Com isso, é possível capturar características específicas de uma região se houver dados suficientes para tal, sem produzir probabilidades tão pouco confiáveis no caso em que os dados são escassos ou mesmo ausentes.

Gelman et al. [38] apresentam outro modelo hierárquico. As proporções $\hat{P}(V_{hi}) = K_{hi}/N_h$ são calculadas para todas as subpopulações, assim como a média e o desvio padrão para estas proporções. A seguir é construída uma distribuição de probabilidade

prévia de Dirichlet (que no caso apresentado é uma distribuição Beta) coerente com a média e o desvio padrão das proporções.

Ambos os modelos hierárquicos referidos nesta seção operam em duas etapas. Em primeiro lugar definem distribuições de probabilidade prévia para as subpopulações que refletem a população global e então utilizam os dados a respeito de cada uma delas para obter distribuições de probabilidade posteriores. Esses modelos são chamado em [38] de modelos empíricos e são conhecidos na literatura pelo nome de *empirical bayes* [12]. Eles partem da intuição de que os resultados obtidos para o população global são bons pontos de partida para a análise das subpopulações sem usar um modelo matemático completo (embora em [21] encontremos justificativas matemáticas para a adoção das probabilidades incondicionais como probabilidades prévias para as probabilidades condicionais).

2.2.3 Um modelo Bayesiano completo

Além do modelo empírico, em [38], é apresentado um modelo hierárquico que permite a determinação das distribuições de probabilidade posteriores considerando todas as subpopulações e a população global de uma só vez. Esse modelo é chamado em [38] de modelo Bayesiano hierárquico completo. Ele é utilizado em vários trabalhos [4, 74, 3] e é mais conhecido como *hierarchical Bayes*.

O modelo apresentado em [38] é genérico, podendo ser aplicado a variáveis contínuas ou discretas. Aqui nos concentraremos no caso em que os dados são representados por variáveis discretas.

Com o modelo Bayesiano completo, ao invés de fixar os parâmetros α, assumimos uma distribuição de probabilidade para eles. A distribuição de probabilidade conjunta para todas as variáveis no modelo pode ser escrita como

$$P(K, \theta, \alpha) = P(K|\theta) \cdot P(\theta|\alpha) \cdot P(\alpha), \tag{2.13}$$

onde θ corresponde a uma matriz onde θ_h é o vetor de probabilidades para os resultados $V_1, V_2, \ldots, V_M$ na região R_h, K é uma matriz onde K_h é o vetor que contém número de observações dos resultados $V_1, V_2, \ldots, V_M$ na região R_h e α é o vetor de parâmetros da distribuição de probabilidade prévia de Dirichlet comum a todas as subpopulações.

Nosso objetivo é calcular $P(V_i|K, h)$ para cada região R_h e cada resultado possível V_i. Isso equivale a calcular $E(\theta_{hi}|K)$ para todo h e todo i. Podemos fazer isso por integração se soubermos a distribuição para $P(\theta, \alpha|K)$. Aplicando a regra de Bayes à Equação (2.13) temos

$$P(\theta, \alpha|K) \propto P(K|\theta) \cdot P(\theta|\alpha) \cdot P(\alpha).$$

Observando que o vetor K_h depende apenas do vetor θ_h quando esse é conhecido e que

cada vetor θ_h independe dos demais vetores θ_x quando α é conhecido, temos

$$P(\theta, \alpha|K) \propto P(\alpha) \prod_{h=1}^{L} P(K_h|\theta_h) \cdot P(\theta_h|\alpha).$$

Assim,

$$E(\theta_{hi}|K) \propto \int \int \theta_{hi} \cdot P(\alpha) \prod_{h=1}^{L} P(K_h|\theta_h) \cdot P(\theta_h|\alpha) d\theta d\alpha.$$

Em geral, é mais fácil não resolver a integral acima diretamente. Em [38] consta uma estratégia para determinação da expressão analítica de $P(\alpha|K)$ que pode depois ser usada para calcular $P(\theta|\alpha, K)$. A estratégia nem sempre funciona, pois ainda restam integrais cuja solução analítica pode não ser conhecida. Nesta situação são utilizados métodos numéricos que estão fora do escopo desta tese.

Em [38], $P(\alpha)$, a distribuição de probabilidade prévia para o nível mais alto na hierarquia, é descrita como *pouco importante*. Isso decorre da suposição de que a população global é sempre grande o suficiente para dominá-la. No entanto, se tentarmos construir uma hierarquia com múltiplos níveis, podemos facilmente encontrar populações que, embora tenham subdivisões, sejam pequenas. Nesse caso, $P(\alpha)$ ganha relevância.

2.3 O problema de classificação de padrões

Antes de descrever um problema de classificação de padrões, precisamos de algumas definições:

Definição 2.2 *Um classificador é uma função que recebe como entrada uma instância e devolve o rótulo da classe a qual a instância pertence.*

Definição 2.3 *Um problema de classificação*[1] *[27, 59] é o problema de dado um conjunto de treinamento, D, de pares, (y_t, c_t), onde y_t é a t-ésima instância no conjunto de treinamento e c_t é o rótulo indicando a classe a qual y_t pertence, aprender um classificador.*

As instâncias podem ser registros, documentos, imagens, etc. Porém, freqüentemente as instâncias são representadas por vetores de valores de atributos que são chamados de padrões. Esses vetores podem corresponder a representação original das instâncias ou ter sido obtidos por um extrator de atributos aplicado em uma fase anterior a fase de classificação [28].

[1] *Em toda esta tese, nos referimos apenas a problemas de classificação supervisionada. Para ler sobre classificação não supervisionada veja por exemplo [28].*

Em alguns problemas, os valores para todos os atributos considerados estão disponíveis para todas as instâncias enquanto em outros algumas instâncias podem ter valores faltantes. Nesta tese, lidamos apenas com problemas onde todos os atributos estão definidos para todas as instâncias, mas pela forma com que abordamos o problema no capítulo 3 adotamos a definição para padrões utilizada em problemas onde pode haver valores faltantes.

Definição 2.4 *Um padrão é um conjunto de pares da forma* $(Atributo = Valor)$*, onde cada atributo pode aparecer no máximo uma vez.*

Definição 2.5 *Um problema de classificação de padrões é um problema de classificação onde as instâncias são representadas por padrões.*

Como a classificação perfeita é freqüentemente impossível, convém reinterpretar um classificador como sendo uma função que determina a probabilidade de pertinência de uma nova instância a cada uma das classes possíveis [28]. Esta reinterpretação corresponde justamente à maneira Bayesiana de lidar com a classificação de padrões.

Determinadas as probabilidades de pertinência de uma nova instância, x, a cada classe, a forma mais simples de fazer a classificação propriamente dita é atribuir a instância a classe que corresponder a maior probabilidade calculada:

$$C = \arg\max_{c_r} P(c_r|x).$$

No entanto, é possível que errar a classificação em um determinado sentido implique em custos maiores que em outro sentido. Por exemplo, classificar um paciente portador de uma doença altamente contagiosa como não doente tem um custo mais alto que classificar um paciente são como doente. Assim, é comum a existência de uma matriz de custos de dimensões $M_c \times M_c$, onde M_c é o número de classes no problema, informando o custo de classificar uma instância pertencente a classe c_i como pertencente a classe c_j. A partir desta matriz, a instância pode ser atribuída à classe que minimize a expectativa de custo:

$$C = \arg\min_{c_i} \sum_{\forall j \neq i} Custo(j,i) \cdot P(c_j|x),$$

onde $Custo(j,i)$ é o custo incorrido ao atribuir uma instância pertencente a classe c_j a classe c_i.

Em outras situações, pode ter sido pré-determinada a quantidade de instâncias atribuídas a uma certa classe. Por exemplo, pode-se estar escolhendo quem deve ser visitado por um vendedor que só é capaz de visitar 10% dos potenciais clientes. Nesse caso, as probabilidades são utilizadas para formação de um ranking entre clientes. Os

10% com maior probabilidade de adquirir o produto são então classificados como *a visitar*, enquanto que os demais são classificados com *a não visitar*.

É fácil ver que se calcularmos $P(c_r|x)$ para todo r de forma correta, seremos capazes de fazer as melhores classificações possíveis em cada um dos contextos descritos acima. Entretanto, mesmo que nossas estimativas para $P(c_r|x)$ não sejam muito próximas dos verdadeiros valores para estas probabilidades, sob certas condições, as classificações serão ótimas.

Condição de Otimalidade 2.1 *Se atribuímos uma instância a classe de maior probabilidade estimada, nossa atribuição será ótima se houver uma coincidência entre classe correspondente a maior estimativa e a classe com maior probabilidade real de ser a classe correta.*

Isso pode ocorrer ainda que as probabilidades reais e estimadas sejam bastante distintas.

Condição de Otimalidade 2.2 *Se atribuímos uma instância a classe de menor custo esperado, nossa atribuição será ótima se houver uma coincidência entre classe correspondente a menor expectativa de custo estimada e a classe com menor expectativa de custo real.*

Condição de Otimalidade 2.3 *Se a classificação depende essencialmente de um ranking ela será ótima se*

$$P'(c_r|y) > P'(c_r|z) \Rightarrow P(c_r|y) > P(c_r|z),$$

onde $P'(c_r|y)$ e $P'(c_r|z)$ representam estimativas enquanto $P(c_r|y)$ e $P(c_r|z)$ representam as probabilidades reais[2].

Esta implicação pode ser válida ainda que as estimativas sejam bastante diferentes dos valores reais.

É possível aplicar a técnica de aprendizado de probabilidades a partir de exemplos descrita na Seção 2.1 ao problema de classificação de padrões de forma direta. Desejamos calcular $P(c_r|x)$ para todo r, onde c_r é uma classe e x é uma instância representada por um padrão w.

Podemos assumir que, dentro do conjunto de instâncias de treinamento representadas por w, a variável C segue uma distribuição multinomial que independe das instâncias

[2] *Em geral não distinguimos a representação das estimativas da representação das probabilidades reais sendo ambas representadas por $P(.)$, pois, quase sempre, o contexto deixa claro o que está sendo representado. Nas situações onde poderia haver confusão, representamos as estimativas por $P'(.)$.*

representadas por padrões diferentes de w. Adotando uma distribuição de probabilidade prévia de Dirichlet não informativa podemos aplicar diretamente a Equação (2.8):

$$P(c_r|w) = \frac{N_{rw} + \lambda}{N_w + \lambda M}, \tag{2.14}$$

onde N_w é o número de instâncias de treinamento representadas exatamente pelo padrão w, N_{rw} é o número de instâncias de treinamento representadas pelo padrão w cuja classe correta é c_r, λ é uma constante de suavização e M é o número de classes.

Se a quantidade de instâncias de treinamento representadas por w fosse suficientemente grande, esta estratégia funcionaria bem. No entanto, o número total de padrões possíveis, considerando apenas os padrões onde todos os atributos estão definidos, é dado por

$$\prod_{j=1}^{L} ||X_j||,$$

onde L é o número de atributos no problema e $||X_j||$ é o número de valores possíveis para o j-ésimo atributo. Se houver, por exemplo, sete atributos no problema, cada um deles com vinte valores possíveis, teremos mais de 10^9 padrões. É impossível esperar que todos os padrões estejam bem representados na maior parte dos conjuntos de treinamento reais. Na verdade, é mais provável que a maior parte dos padrões possíveis não corresponda exatamente a nenhuma instância no conjunto de treinamento. Assim, estaríamos lidando com amostras muito pequenas ou mesmo vazias, o que como já vimos comprometeria a confiabilidade das probabilidades estimadas. A situação torna-se ainda mais grave quando os atributos tem alta cardinalidade.

Aplicar um modelo hierárquico de dois níveis como os apresentados na Seção 2.2 diretamente a um problema de classificação com vários atributos seria apenas ligeiramente melhor que usar um modelo plano, e fazer isso não é comum na literatura. Por outro lado, como veremos na Seção 2.6.1, um modelo hierárquico empírico apresentado na Seção 2.2 costuma ser usado na tarefa de preencher tabelas de probabilidades condicionais de redes Bayesianas levando ao que chamamos de *estimativa quase direta*.

2.4 Recorrendo a suposições de independência

Para classificar uma instância x, precisamos calcular $P(c_r|x)$ para todo r. Pelo teorema de Bayes,

$$P(c_r|x) = \frac{P(x|c_r)P(c_r)}{P(x)} \propto P(x|c_r)P(c_r).$$

Quebrando a instância x em seus atributos,

$$P(c_r|x) \propto P(X_1 = x_1, X_2 = x_2, \ldots, X_l = x_L|c_r)P(c_r),$$

onde L é o número total de atributos, X_j é o j-ésimo atributo no problema e x_j é o valor assumido pelo j-ésimo atributo na instância x. Se supusermos que cada atributo é independente dos demais dada a classe teremos que

$$P(c_r|x) \propto P(c_r) \prod_{j=1}^{L} P(x_j|c_r), \tag{2.15}$$

onde por simplicidade representamos $X_j = x_j$ apenas por x_j.

Esta suposição é a característica essencial do mais simples e mais difundido entre os métodos de classificação Bayesianos: o *naïve Bayes* [27]. Usando o *naïve Bayes* não precisamos mais estimar a probabilidade da classe considerando todos os atributos ao mesmo tempo. Com isso, não precisamos nos preocupar com o grande número de padrões possíveis e o pequeno número de exemplos de treinamento representando cada padrão. Precisamos apenas estimar $P(x_j|c_r)$, para todo j e todo r. A partir destas estimativas, a Equação (2.15) fornecerá $P(c_r|x)$ que é nosso objetivo .

Podemos estimar $P(x_j|c_r)$ usando o mecanismo Bayesiano descrito na Seção 2.1. Ainda que tenhamos atributos com alta cardinalidade, podendo assumir, por exemplo, 1000 valores distintos, um conjunto de treinamento de tamanho razoável, contendo, por exemplo, 100000 instâncias de treinamento, é perfeitamente suficiente para a obtenção de boas estimativas.

Entretanto, seria um erro assumir que o tamanho de um conjunto de treinamento típico é suficiente para que possamos dispensar o uso de distribuições de probabilidade prévia e estimar $P(x_j|c_r)$ usando máxima verossimilhança. Para percebermos isso, basta observar o que ocorreria se o valor de um dos atributos na instância sendo classificada (digamos x_j) não tiver ocorrido no conjunto de treinamento nem uma vez.

Usando máxima verossimilhança, estimaríamos que $P(x_j|c_r) = 0$. Isso levaria todo o lado direito da Equação (2.15) a tornar-se igual a zero. A influência de todos os demais atributos seria perdida e não seria possível fazer a classificação, visto que $P(c_r|x)$ seria igual a zero para todas as classes. Ao mesmo tempo, se x_j ocorresse em uma única instância de treinamento, a nova instância, x, seria sempre classificada na classe onde x_j tivesse sido observado em sua única aparição no conjunto de treinamento. Os demais atributos perderiam qualquer influência, o que não é desejável. Em geral, valores raros levariam a variações abruptas e indevidas nos resultados do *naïve Bayes*.

Para evitar esses problemas, o *naïve Bayes* é sempre acompanhado de algum mecanismo de suavização. No caso mais comum, ao invés de

$$P(x_j|c_r) = \frac{N_{jr}}{N_r}$$

onde N_r representa o número de instâncias de treinamento cuja classificação é c_r e N_{jr} representa o número de instâncias treinamento cuja classificação é c_r e cujo valor do

j-ésimo atributo é igual a x_j, temos

$$P(x_j|c_r) = \frac{N_{jr} + 1}{N_r + M}$$

onde M é o número de valores possíveis para o atributo X_j. Esse mecanismo de suavização é chamado de Laplace smoothing e corresponde diretamente a lei de sucessão de Laplace (Equação (2.9)), que, por sua vez, decorre da adoção de uma distribuição de probabilidade prévia uniforme.

Esse mecanismo de suavização é tão popular que é o único disponível para o *naïve Bayes* em uma ferramenta de aprendizado de máquina tão difundida quanto o Weka [77] em sua versão 3.4.2. O Lidstone smoothing, a simples generalização apresentada na Equação (2.8), é também comum e com a seleção experimental do tamanho da amostra equivalente pode apresentar um desempenho superior.

O *naïve Bayes* em conjunto com os mecanismos de suavização tem poucos problemas com super-ajuste, mas uma deficiência óbvia: a suposição de independência entre todos os atributos dada a classe é falsa em praticamente todos os problemas de interesse prático. Esta falsa suposição é a razão para o nome pejorativo do *naïve Bayes*.

Em uma primeira análise, poderíamos esperar que o desempenho prático do *naïve Bayes* em problemas de classificação fosse bastante insatisfatório, mas não é o que se verifica. O *naïve Bayes* tem desempenho competitivo com os melhores métodos disponíveis em vários domínios de classificação de padrões e de texto [26, 66, 59, 34].

Domingos e Pazzani [26] demonstram o fato de que a Condição de Otimalidade 2.1 é sempre verificada para as estimativas produzidas pelo *naïve Bayes* em alguns contextos (aprendizado de conjunções e disjunções) onde a suposição de independência é violada. Assim, sabemos que, quando a tarefa em questão é a de atribuir instâncias a classes minimizando a quantidade absoluta de erros, o *naïve Bayes* pode até ser ótimo em problemas de classificação onde sua suposição básica é falsa.

Isso não significa que suas estimativas de probabilidades sejam próximas das probabilidades reais.

As estimativas produzidas pelo *naïve Bayes* são descritas na literatura como sendo extremas ou excessivamente confiantes [26, 79, 6]. Isso quer dizer que quando a probabilidade real está próxima de um a estimativa produzida pelo *naïve Bayes* está ainda mais próxima de um e quando a probabilidade real está próxima de zero a estimativa produzida pelo *naïve Bayes* está ainda mais próxima de zero.

Um classificador que produz probabilidades extremas é um caso particular de classificador mal balanceado. Um classificador bem balanceado é aquele cujas estimativas de probabilidade tendem as probabilidades empíricas quando o número de previsões tende a infinito [60].

Esse problema pode ser atenuado através de mecanismos de balanceamento. Esses mecanismos tentar encontrar uma função monotônica que mapeie as probabilidades não balanceadas (normalmente a saída de algum método de classificação), em estimativas de probabilidade aproximadamente balanceadas [6, 79, 78, 65].

Se a deficiência essencial de uma estimativa feita pelo *naïve Bayes* é a de ser extrema é possível que ele também tenha um bom desempenho quando a classificação depende apenas de um ranking e não das estimativas em si. Esse bom desempenho é verificado experimentalmente por Zhang e Suem [80] que também apresentam condições dentro das quais a Condição de Otimalidade 2.3 sempre se verifica, apesar da violação da suposição de independência.

Rish et al. [66], demonstram experimentalmente que o desempenho do *naïve Bayes* é, em geral, melhor quando as dependências entre os atributos são fracas, o que é esperado dada sua suposição essencial. No mesmo trabalho também mostram que quando as dependências são muito fortes o desempenho do *naïve Bayes* também é muito bom, o que é inesperado. O *naïve Bayes* apresenta um desempenho pior quando o grau de dependência entre os atributos está em níveis intermediários.

Tudo isso não significa que não seja possível obter ganhos de desempenho relaxando a forte suposição de independência do *naïve Bayes*. Esse relaxamento é geralmente conseguido pela aplicação de redes Bayesianas, assunto da Seção 2.6.

Antes de discutirmos quais são nossas perdas ao assumir independência entre os atributos, vamos descrever o conceito mais genérico de modelo de interação linear ou de não interação.

2.5 Modelos de interação linear ou de não interação

Até o momento, associamos atributos a classes de forma estatística sem discutir a natureza das relações causais entre eles. Para apresentar alguns outros modelos probabilísticos, convém interpretar os atributos como causas da classe ou como efeitos desta. Nosso objetivo aqui é avaliar a forma com que a presença de uma causa influencia ou deixa de influenciar o efeito de outras causas sobre a variável alvo. Modelos causais, ou modelos de interação são discutidos por exemplo em [64, 63, 75].

Se um conjunto de variáveis $X_1, X_2, \ldots, X_l$ representa um conjunto de causas e uma variável E representa o efeito destas causas, diz-se que as causas não interagem se o efeito de cada uma delas for separável, ou seja, se uma variação na variável X_i provocar uma mesma variação na variável E independentemente do valor de X_j, para todo $j \neq i$. Esta

condição se verifica se e somente se for possível escrever

$$E = F(X_1, X_2, \ldots, X_l) = B_0 + \sum_{j=1}^{l} B_j \cdot F_j(X_j), \tag{2.16}$$

onde B é um vetor de coeficientes e F é um vetor de funções que podem ser ou não lineares. Note que na Equação (2.16) o efeito das causas é combinado de forma linear. Assim, um modelo chamado de modelo de não interação é na verdade um modelo de interação linear.

Podemos ver a Equação (2.15), a equação central do *naïve Bayes*, como um caso particular de modelo de não interação. Basta definir que

1. $E = \log P(c_r|x)$;
2. $B_0 = 0$;
3. $B_j = 1, \forall j \in 1, \ldots, L+1$, onde L é o número de atributos;
4. $F_j(x_j) = \log P(x_j|c_r), \quad \forall j \in 1, \ldots, L$;
5. $F_j(x_j) = \log P(c_r), \quad j = L+1$;

Assim, ao utilizar o *naïve Bayes*, é possível interpretar os atributos e a probabilidade prévia da classe, $P(c_r)$, como causas que combinadas linearmente determinam o logaritmo da probabilidade posterior da classe $P(c_r|x)$. No entanto, como $F_j(X_j)$ foi definida como uma função de $P(x_j|c_r), \forall j$ e não de $P(c_r|x_j), \forall j$, a interpretação mais comum na literatura é a de que os atributos são efeitos da classe.

Um outro modelo de não interação causal importante é o *noisy-OR* [39, 63]. Esse modelo é inspirado em uma porta *OR* da lógica booleana. Se qualquer das causas está presente, então o efeito deverá, a princípio, ocorrer. É, no entanto, considerada a possibilidade de que uma dada causa seja inibida ou falhe na tarefa de produzir o efeito.

Para cada variável causa, X_j, considera-se que existe um fator de inibição I_j que pode estar ativo com probabilidade q_j e que os fatores de inibição agem de forma (incondicionalmente) independente. O *noisy-OR* define o efeito E, como uma função das variáveis causa e das variáveis inibidoras:

$$E = \bigvee_{\forall j} (X_j \wedge I_j), \tag{2.17}$$

onde o valor da variável I_j é considerado *falso* se I_j está ativa. Esta função possui algumas propriedades atraentes como *accountability* [63], *exception independence* [63], *associativity* [81], *reverse independence* [1], *explaining away* [19] e *cumulativity* [19]. Em [20] é

demonstrado que as funções do tipo *noisy-OR* são as únicas funções determinísticas que satisfazem a todas estas propriedades.

Como conseqüência da Equação (2.17) e do fato de que a probabilidade de que a causa X_j não seja inibida é $p_j = 1 - q_j$:

$$P(E = Presente|X_1, \ldots, X_L) = 1 - \prod_{\forall j|X_j=Presente} (1 - p_j). \tag{2.18}$$

As definições que exibem o *noisy-OR* como um caso particular de modelo de não interação podem ser vistas por exemplo em [75]. Nesse mesmo trabalho é descrito outro modelo de não interação, o modelo logístico.

O *noisy-OR* foi definido originalmente para variáveis binárias, mas existem extensões para variáveis multinomiais por exemplo em [69, 23]. Ferreira [32] empregou outra extensão do *noisy-OR* na solução de um problema de classificação de padrões, onde a variável que representa a classe é binária (podendo assumir os valores *positivo* ou *negativo*), mas os atributos são multinomiais.

Ao invés de uma causa, que pode estar ausente ou presente, cada atributo passa a representar um conjunto de causas mutuamente exclusivas, cada uma delas associada a um valor possível para o atributo. Cada uma destas causas tem probabilidade de não inibição distinta definida por

$$p_{ji} = \frac{N_{ji,positivo}}{N_{ji}},$$

onde N_{ji} é o número de instâncias de treinamento onde o atributo X_j assume seu i-ésimo valor possível e $N_{ji,positivo}$ é o número de instâncias de treinamento onde o atributo X_j assume seu i-ésimo valor possível e a classe correta é *positivo*. Temos então que

$$P(positivo|x) = 1 - \prod_{\forall j} \left(1 - \frac{N_{jy,positivo}}{N_{jy}}\right),$$

onde N_{jy} é o número de instâncias de treinamento onde o atributo X_j assume o mesmo valor que o próprio atributo X_j assume na instância x sendo classificada e $N_{jy,positivo}$ é o número de instâncias de treinamento onde o atributo X_j assume o mesmo valor que o próprio atributo X_j assume na instância x e a classe correta é *positivo*.

No problema de classificação de padrões, a grande vantagem da utilização de um modelo de não interação é que a probabilidade de que uma classe C_r seja a classe correta para uma nova instância deixa de ser inferida a partir do conjunto de treinamento em $\prod_{j=1}^{L} ||X_j||$ contextos diferentes, um para cada combinação de valores de atributos, onde L é o número de atributos no problema e $||X_j||$ é o número de valores possíveis para o j-ésimo atributo e passa a ser estimada em $\sum_{j=1}^{L} ||X_j||$ contextos diferentes, um para cada valor individual de um atributo.

Não discutimos ainda o preço que pagamos ao utilizar um modelo de interação linear, mas ele não reserva qualquer surpresa: quando há interações não lineares entre os atributos, elas não são refletidas pelos classificadores construídos a partir dos modelos de interação linear. Sempre que estas interações são decisivas, os classificadores cometem erros.

A despeito do sucesso do *naïve Bayes* e do *noisy-OR*, existem contextos onde a representação não linear de interações entre atributos é fundamental. Um exemplo clássico é a situação onde existem dois atributos, o primeiro representando a presença ou ausência de um ácido forte em uma composição química. O segundo representa a presença ou ausência de uma base forte na composição e o efeito sendo avaliado é a corrosão causada pela composição em um objeto de testes. Se a composição não contém nem ácido nem base ela é inócua. Se contém ou ácido ou base ela mostra um forte poder de corrosão. Um modelo de interação linear esperaria que a composição contendo as duas substâncias juntas fosse ainda mais corrosiva, quando na realidade ela corresponde a um sal inócuo.

O exemplo acima exibe uma interação onde ocorre neutralização, mas as interações podem produzir uma intensificação inesperada dos efeitos, produzir um efeito na direção esperada porém com uma força reduzida ou até mesmo causar um efeito inverso ao que se esperaria pela análise dos efeitos individuais.

2.6 Redes Bayesianas

Suponhamos que queremos representar a distribuição de probabilidade conjunta das L variáveis $X_1, X_2, \ldots, X_L$, que compõem um dado problema. Pela regra da cadeia a distribuição de probabilidade conjunta pode ser escrita como um produto de distribuições de probabilidade condicionais onde a distribuição de probabilidade de cada variável depende de todas as suas antecessoras em uma ordem pré-especificada qualquer:

$$P(X = x_t) = P(X_1 = x_{1t}, X_2 = x_{2t}, \ldots, X_L = x_{Lt}) = \prod_{j=1}^{L} P(X_j = x_{jt} | Ant_j = ant_{jt}),$$

onde X é o conjunto de todas as variáveis no problema, x_t é uma instanciação para X, Ant_j é o conjunto de todas as variáveis que antecedem X_j na ordem pré-especificada e ant_{jt} é o conjunto de valores assumidos na instanciação x_t pelas variáveis presentes em Ant_j.

Suponhamos agora dispormos de um conjunto de declarações de independência condicional I que nos permita afirmar que para todo j,

$$P(X_j = x_{jt} | Ant_j = ant_{jt}) = P(X_j = x_{jt} | \Pi_j = \pi_{jt}),$$

onde Π_j é um subconjunto de Ant_j. Temos que

$$P(x_{1t}, x_{2t}, \ldots, x_{Lt}) = \prod_{j=1}^{L} P(x_{jt}|\pi_{jt}), \tag{2.19}$$

onde passamos a representar $X_j = x_{jt}$ apenas por x_{jt} e $\Pi_j = \pi_{jt}$ apenas por π_{jt}.

Uma rede Bayesiana (BN) [63] é um grafo direcionado acíclico (DAG) anotado que representa a distribuição de probabilidade conjunta expressa pela Equação (2.19). No grafo, cada nó corresponde a uma variável e existe uma aresta partindo do nó X_z e chegando ao nó X_w se e somente se $X_z \in \Pi_w$. Ao mesmo tempo, se tal aresta existe diz-se que X_z é pai de X_w.

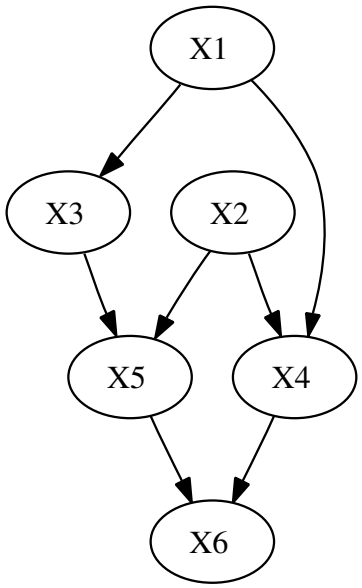

Figura 2.1: um exemplo de rede Bayesiana

Uma BN tenta representar o conjunto de declarações de independência I, mas nem sempre isso é possível. A rede representa um conjunto de declarações de independência $I_{bn} \subseteq I$.

O conjunto I_{bn} pode ser lido como: para todo j, a variável X_j independe de todas as variáveis que não são suas descendentes dados os seus pais. A partir destas, outras relações de independência podem ser deduzidas.

Além das declarações de independência, uma rede Bayesiana, também representa uma ordem parcial entre a variáveis do problema. Um mesmo conjunto de declarações de independência pode ser representado por diferentes redes Bayesianas, cada uma delas

correspondendo a uma ordem distinta. Na Figura 2.2, ambas as redes Bayesianas representam o conjunto de declarações de independência $I = \{\}$.

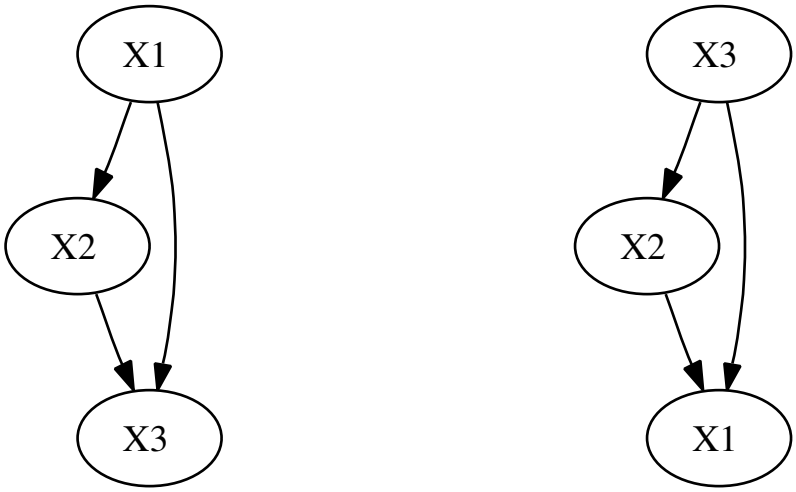

Figura 2.2: duas redes Bayesianas representando o mesmo conjunto de independências

Por outro lado, a depender da ordem escolhida pode ser ou não possível representar certas declarações de independência presentes no conjunto I.

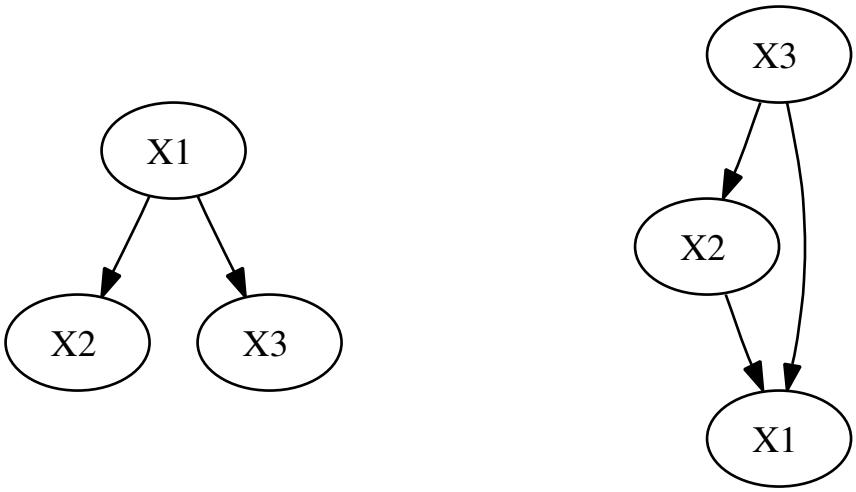

Figura 2.3: duas redes Bayesianas tentando representar o mesmo conjunto de independências

Ambas as redes Bayesianas na Figura 2.3 tentam representar o conjunto de declarações de independência: $I = \{P(X_3|X_1, X_2) = P(X_3|X_1)\}$. No primeiro caso, usando a ordem

X_1, X_2, X_3 o conjunto I é plenamente representado. No segundo caso, a adoção da ordem X_3, X_2, X_1 impediu a representação da declaração de independência. Note que não é possível remover qualquer aresta em qualquer dos grafos apresentados na Figura 2.3, sob pena de introduzir declarações de independência que não estão presentes em I.

Mesmo que as mesmas declarações de independência possam ser representadas por redes Bayesianas construídas a partir de diferentes ordenações entre variáveis, estas redes Bayesianas não serão intuitivamente equivalentes. Em geral, interpreta-se que duas variáveis conectadas por uma aresta em uma rede Bayesiana mantém uma relação de causalidade, onde a origem da aresta é a causa e o destino é o efeito.

Até o momento mostramos a relação entre a estrutura da rede Bayesiana e a independência entre variáveis, contudo isso não é suficiente para representar a distribuição de probabilidade conjunta das últimas. Como já afirmamos uma rede Bayesiana é um grafo acíclico anotado. A cada nó X_j está associada uma função F_j que fornece a distribuição de probabilidade condicional (CPD) para a variável X_j dada qualquer instanciação para todas as variáveis pertencentes a Π_j: $F_j(X_j, \Pi_j) = P(X_j|\Pi_j)$.

Uma rede Bayesiana pré-construída pode ser usada na solução do problema de classificação de padrões. Nesse caso, um dos nós da rede representa a classe enquanto que os demais representam atributos explanatórios. Para fazer a classificação precisamos obter a probabilidade da classe condicionada a todos os atributos. Aplicando a regra de Bayes,

$$P(X_c|X_1, X_2, \ldots, X_L) = \frac{P(X_c, X_1, X_2, \ldots, X_L)}{P(X_1, X_2, \ldots, X_L)} \propto P(X_c, X_1, X_2, \ldots, X_L),$$

assim, pela Equação (2.19),

$$P(X_c|X_1, X_2, \ldots, X_L) \propto \prod_{j=0}^{L} P(X_j|\Pi_j), \tag{2.20}$$

onde consideramos que $P(X_c|\Pi_c)$ está representada por $P(X_j|\Pi_j)$ quando $j = 0$.

Como dispomos dos conjuntos Π_j para todo j e de funções F_j que fornecem $P(X_j|\Pi_j)$ para todo j a aplicação da Equação (2.20) é trivial. Naturalmente o problema é a construção da rede Bayesiana.

Supondo que todas as variáveis sejam discretas, como é sempre o caso nesta tese, as funções F_j passam a ser tabelas. Estas tabelas são chamadas de tabelas de probabilidades condicionais (CPTs). A CPT do nó X_j possui uma coluna para cada instanciação de valor possível para X_j e uma linha para cada instanciação completa de valores possível para Π_j. As células são preenchidas de tal modo que a célula Cel_{jki} contém $P(X_j = x_{ji}|\Pi_j = \pi_{jk})$, onde x_{ji} é a i-ésima instanciação possível para variável X_j e π_{jk} é a k-ésima instanciação completa possível para o conjunto Π_j.

Se temos um número pequeno de variáveis, se cada variável X_j pode assumir um número pequeno de valores e para todo j o número de elementos em Π_j é pequeno, então a rede Bayesiana poderá ser construída por um especialista humano. Caso contrário, isso precisa ser feito de forma automática.

2.6.1 Preenchendo tabelas de probabilidades condicionais a partir dos dados

Suponhamos que a estrutura, S, de uma rede Bayesiana seja dada e que disponhamos de um conjunto de treinamento D de instâncias da forma $y_t = y_{1t}, y_{2t}, \ldots, y_{nt}$, onde t é o índice da instância, para as quais conhecemos os valores de todas as variáveis. Desejamos obter estimativas que preencham as tabelas de probabilidades condicionais de todos os nós.

Podemos considerar que os valores, $x_{j1}, x_{j2}, \ldots, x_{jm}$, que podem ser assumidos por um nó, X_j, dada uma atribuição de valores, π_{jk}^S, para os pais de X_j, Π_j^S, na estrutura S, seguem uma distribuição multinomial que independe de qualquer outra parte de S e independe das distribuições de X_j dada qualquer outra atribuição de valores para Π_j^S. Em geral, as tabelas de probabilidades condicionais são estimadas adotando-se uma distribuição de probabilidade prévia não informativa de Dirichlet para os parâmetros de cada multinomial. Aplicando a estratégia Bayesiana apresentada na Seção 2.1 temos

$$P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + \lambda}{N_{jk} + \lambda M_j}, \tag{2.21}$$

onde N_{jki} é o número de observações simultâneas de x_{ji} e π_{jk} no conjunto de treinamento, N_{jk}, é o número de observações de π_{jk}, λ é uma constante de suavização e M_j é o número de valores possíveis para X_j.

A Equação (2.21) é a estratégia mais comum para estimar tabelas de probabilidades condicionais sendo descrita por exemplo em [61, 68, 8]. Chamamos esta estratégia de *estimativa direta*, estendendo ligeiramente o escopo da denominação usada em [8], onde apenas o caso em que $\lambda = 1$ é chamado desta forma.

Uma pequena alteração na Equação (2.21) apresentada em [35] utiliza as probabilidades preditivas marginais $P(X_j = x_{ji})$, como probabilidades preditivas prévias para a probabilidades condicionais $P(X_j = x_{ji}|\Pi_j = \pi_{jk})$ e a partir delas constrói uma distribuição de probabilidade prévia de Dirichlet como foi descrito na Seção (2.2):

$$P(X_j = x_{ji}|\Pi_j = \pi_{jk}) = \frac{N_{jki} + S \cdot P(X_j = x_{ji})}{N_{jk} + S}, \tag{2.22}$$

onde $P(X_j = x_{ji})$ é estimada pela equação

$$P(X_j = x_{ji}) = \frac{N_{ji}}{N},$$

onde N é o número total de instâncias no conjunto de treinamento. Chamamos esse mecanismo modificado de *estimativa quase direta* (ADE).

Se o número de instanciações possíveis para X_j, para algum j, for muito grande, ao usar a Equação (2.21) ou a Equação (2.22), teremos estimativas pouco confiáveis para as probabilidades. De fato, se a estrutura da nossa rede for análoga a da Figura 2.4, teremos caído exatamente na mesma situação em que estávamos quando aplicamos a estratégia para estimar probabilidades a partir de uma amostra apresentada na Seção 2.1 diretamente ao problema de classificação de padrões (Equação (2.14)).

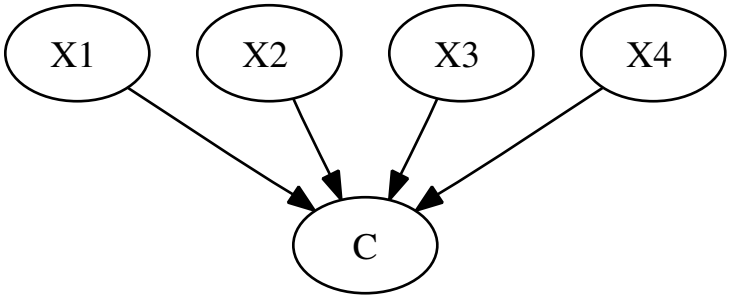

Figura 2.4: rede causal direta

De uma forma geral, as estimativas de probabilidade tornam-se menos confiáveis a medida em que pais adicionais são acrescidos [47]. Por isso, é comum limitar o número de pais por nó.

Uma abordagem extrema é a de limitar o número de pais de um nó a um, como ocorre na Figura 2.5, que nada mais é que uma representação do *naïve Bayes* como uma rede Bayesiana.

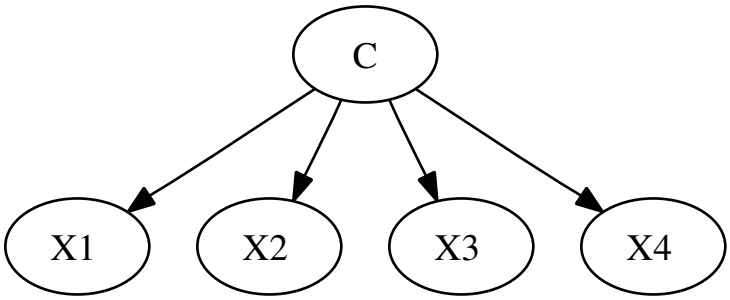

Figura 2.5: *naïve Bayes*

É fácil ver, pelas Equações (2.19) e (2.15), que uma rede Bayesiana com a estrutura apresentada na Figura (2.5) é de fato equivalente ao *naïve Bayes* e que esse é portanto um

caso particular de Rede Bayesiana. Uma opção menos extrema é a limitação do número de pais a dois [35, 47, 40].

Infelizmente limitar o número de pais para um nó pode nos obrigar a codificar na rede declarações de independência que na verdade não existem. Estas declarações falsas reduzem nossa capacidade de capturar interações não lineares relevantes entre atributos. Ling e Zhang [56] demonstram limitações da capacidade de representação de redes Bayesianas quando o número máximo de nós pais é limitado. Na prática, a estrutura da rede Bayesiana deve buscar um equilíbrio entre a confiabilidade das estimativas e o tratamento de todas as dependências desejadas.

Para reduzir os problemas com as estimativas pouco confiáveis resultantes de CPTs muito grandes podemos representar a CPD de um nó dados seus pais usando um modelo que empregue menos parâmetros. Modelos de não interação (Seção 2.5) são a recomendação de Pearl [63]. Outras alternativas são tabelas default [37], árvores de decisão [37, 10] e grafos de decisão [16, 44]. Uma breve descrição destas três últimas técnicas e uma análise de seu efeito sobre o problema de motivação desta tese pode se encontrado em [45].

2.6.2 Construindo uma estrutura de rede a partir dos dados

Suponhamos agora que a estrutura da rede Bayesiana também precise ser inferida a partir do conjunto de treinamento. Podemos definir algum critério que avalie a qualidade de uma possível estrutura de rede frente ao conjunto de dados e procurar no espaço (ou em um subespaço) das possíveis estruturas de rede aquela que o maximiza. Precisamos portanto:

- escolher um critério de comparação entre estruturas candidatas;
- escolher o espaço de busca (que pode ser o espaço de todas as estruturas de rede possíveis ou não);
- escolher um algoritmo de busca.

Critério de máxima verossimilhança

Um critério simples para comparar estruturas de rede é o critério de máxima verossimilhança [33]. Esse critério determina que o modelo a ser adotado é aquele que maximiza a verossimilhança dos exemplos de treinamento. Se o modelo em questão é uma rede Bayesiana, então a BN a ser adotada, incluídos aí sua estrutura e seus parâmetros, é aquela que que maximiza a verossimilhança dos exemplos de treinamento.

Assim, a estrutura S a ser adotada é aquela que, com a parametrização de máxima verossimilhança, maximiza a verossimilhança dos exemplos de treinamento. Como esses exemplos são considerados independentes uns dos outros dada a rede, temos que a verossimilhança dos dados é dada por

$$L^{ml}(S|D) = P(D|S, \theta^S) = \prod_{t=1}^{N} P(y_t|S, \theta^S) = \prod_{t=1}^{N} \prod_{j=1}^{L} P(y_{jt}|\pi_{jt}, S, \theta^S),$$

onde S é a estrutura da rede Bayesiana, D é o conjunto de treinamento, N é o número de instâncias nesse conjunto, θ^S é o conjunto de tabelas de probabilidades condicionais estimadas a partir de D assumindo que a estrutura da rede está fixada em S. $P(y_t|S, \theta^S)$ é a estimativa para a probabilidade de que as variáveis do problema assumam exatamente os valores correspondentes a t-ésima instância de treinamento, y_t, obtida pelo emprego da rede Bayesiana definida pela estrutura S e pelo conjunto de tabelas de probabilidades condicionais, θ^S, e $P(y_{jt}|\pi_{jt}, S, \theta^S)$ é a probabilidade de que o j-ésimo atributo assuma o valor que assumiu na instância y_t dados os valores assumidos por seus pais em y_t.

Em um processo de maximização, é comum que um produto seja transformado em uma soma de logaritmos. Esta transformação dá origem a verossimilhança logarítmica:

$$LL^{ml}(S|D) = \log L^{ml}(S|D) = \sum_{t=1}^{N} \sum_{j=1}^{L} \log P(y_{jt}|\pi_{jt}, S, \theta^S).$$

O conceito de verossimilhança logarítmica é também importante por sua relação com o mínimo comprimento de descrição (Seção 2.6.2).

A estrutura de máxima verossimilhança S^{ml} é dada por

$$S^{ml} = \arg\max_S L^{ml}(S|D) = \arg\max_S LL^{ml}(S|D).$$

O critério de máxima verossimilhança captura bem o ajuste do modelo aos dados de treinamento, mas carece de qualquer preocupação com a confiabilidade das estimativas. Redes Bayesianas com estruturas complexas, com vários pais por nó e CPTs muito grandes, podem atribuir uma verossimilhança muita alta aos dados de treinamento, mas ter desempenho pobre diante de dados novos.

Mínimo comprimento de descrição

O critério denominado mínimo comprimento de descrição (MDL) [67, 49] é motivado por uma analogia com a teoria da informação e incorpora ao critério de máxima verossimilhança um custo para cada parâmetro requerido pela estrutura S.

Suponhamos que queremos armazenar o conjunto de dados de treinamento usando uma representação compacta. Podemos fazer isso usando mais bits para codificar valores mais

raros e menos para representar valores freqüentes. Para tal, precisamos de um modelo que forneça a probabilidade de cada valor no conjunto de dados.

Se usarmos uma rede Bayesiana como modelo, o total de bits necessário para armazenar o conjunto de dados em formato compacto será

$$DL_S^{data} = -\sum_{t=1}^{N}\sum_{j=1}^{L} \log P(y_{jt}|\pi_{jt}, S, \theta^S),$$

o que corresponde exatamente a negação da verossimilhança logarítmica.

Além de armazenar os dados, para que a descompressão seja possível, é preciso armazenar o modelo empregado na compressão. Esse modelo é compreendido pela estrutura da rede e por seus parâmetros.

Para codificar o grafo direcionado acíclico que corresponde a estrutura da rede Bayesiana, precisamos, para cada nó, armazenar o seu número de pais e listar esses pais. Como tanto o número de pais de um nó, quanto um índice indicando um pai são limitados superiormente pelo número de nós na rede o número total de bits requerido para codificar o grafo direcionado acíclico é

$$DL_S^{dag} = \sum_{j=1}^{L}(1 + |\Pi_j)|) \log L.$$

Para codificar os parâmetros presentes na CPT associada a um nó X_j, precisamos, para cada combinação de valores possíveis para Π_j, armazenar a probabilidade de que X_j assuma cada um de seus possíveis valores com exceção do último que é dado por $1 - \sum_{i=1}^{||X_j||-1} P(x_{ji})$. Com isso temos que

$$DL_S^{paracpt} = \sum_{j=1}^{L} d \cdot ||\Pi_j||(||\Pi_j|| - 1),$$

onde d é o número de bits usado para representar cada probabilidade. Em [7, 42, 36, 34, 37] adota-se $d = 1/2 \log N$. Esse valor decorre do relação entre o MDL e o *Bayesian information criterion* (BIC) [70]. Um paralelo entre esses critérios está disponível em [51]. No presente trabalho também adotamos esse valor, o que nos dá

$$DL_S^{paracpt} = \sum_{j=1}^{L} \frac{1}{2}||\Pi_j||(||X_j|| - 1) \log N.$$

A codificação completa dos dados e do modelo requer um número de bits dados por

$$DL_S^{totalcpt} = DL_S^{data} + DL_S^{dag} + DL_S^{paracpt}.$$

Redes Bayesianas complexas, com vários pais por nó, ajustam-se bem aos dados levando a um valor pequeno para DL^{data}, porém elas precisam de CPTs grandes, cada uma delas com muitos parâmetros. Esses parâmetros levam a um valor alto para $DL^{paracpt}$.

O mínimo comprimento de descrição propõe a seleção da estrutura que leve ao menor valor para $DL^{totalcpt}$:

$$S^{mdl} = \arg\min_{S} DL_S^{totalcpt}$$

Com isso, o MDL representa um ponto de equilíbrio entre o ajuste aos dados e a complexidade do modelo empregado. Ele equivale a negação da verossimilhança logarítmica acrescida de uma componente que pune estruturas complexas.

Bayesian score

O *Bayesian score* (BS) [43] é um critério popular para seleção de estruturas de redes Bayesianas. O valor do *Bayesian score* de uma estrutura S é definido como sendo a probabilidade posterior $P(S|D)$. Pela regra de Bayes temos

$$BS_S = P(S|D) \propto P(S)P(D|S).$$

Assim, precisamos empregar conhecimento do domínio para definir uma distribuição probabilidade prévia para as estruturas candidatas. Como esse conhecimento, freqüentemente, não está disponível, é comum a adoção de um valor constante para $P(S)$.

$P(D|S)$ é a verossimilhança dos dados dada a estrutura da rede Bayesiana, o que não é o mesmo que a verossimilhança dos dados dada a própria rede Bayesiana, posto que esta inclui a estrutura e os parâmetros armazenados nas tabelas de probabilidades condicionais.

É correto assumir que os exemplos de treinamento são independentes entre si dada uma rede Bayesiana, mas não podemos assumir o mesmo dada apenas a estrutura da rede, pois os exemplos afetam nossas estimativas para os parâmetros, ainda desconhecidos, da BN. Pela regra da cadeia uma instância y_t depende de todas as instâncias y_r onde $r < t$. Assim,

$$P(D|S) = \prod_{t=1}^{N} P(y_t|S, D_t) = \prod_{t=1}^{N} P(y_t|S, \theta_t^S), \tag{2.23}$$

onde D_t é o conjunto que contém as instâncias $y_1, \ldots, y_{t-1}$ e θ_t^S é o conjunto de parâmetros estimados a partir de D_t e de uma distribuição de probabilidade prévia para θ^S assumindo que a estrutura da rede está fixa em S.

Como já dissemos na Seção 2.6.1, geralmente, a estimativa de θ_{jt}^S é feita supondo uma distribuição de probabilidade prévia de Dirichlet. Esta suposição dá origem a um importante caso particular de BS, o *Bayesian Dirichlet score* (BDS). O K2 [18] é um

caso particular de BDS onde todos os parâmetros de todas as distribuições de Dirichlet são fixados em um, ou seja, todas as distribuições de Dirichlet são na verdade uniformes. Como já dissemos, redes Bayesianas distintas podem codificar as mesmas suposições de independência. Heckerman et al. [43], demonstram que, se os valores dos parâmetros das distribuições de Dirichlet satisfizerem a restrição,

$$\alpha_{jki} = N' \cdot P(x_{ji}, \pi_{jk}|B_c), \tag{2.24}$$

onde N' é um tamanho de amostra equivalente e B_c é qualquer rede Bayesiana com estrutura completa (nenhuma aresta faltando), então para quaisquer duas estruturas S_1,S_2, que sejam equivalentes (codifiquem as mesmas suposições de independência) $P(D|S_1) = P(D|S_2)$. O caso particular de BDS onde os parâmetros α satisfazem a Equação (2.24) é chamado de *Bayesian Dirichlet likelihood equivalent score* (BDES). Note que o K2 não é um BDES. O caso particular de BDES que atribui a mesma probabilidade a todos os estados possíveis para a rede é chamado de *Bayesian Dirichlet likelihood equivalent uniform score* (BDEUS) [11]. O BDEUS é obtido submetendo os parâmetros α do BDS a restrição:

$$\alpha_{jki} = \frac{N'}{||X_j|| \cdot ||\Pi_j||} \tag{2.25}$$

Critério de máxima verossimilhança em validação cruzada do tipo *leave one out*

O critério de máxima verossimilhança em validação cruzada do tipo *leave one out* (LLOO), seleciona a estrutura S que maximiza a verossimilhança do conjunto de treinamento avaliada em um processo de validação cruzada do tipo *leave one out*(LOO). Isso significa que a verossimilhança de um exemplo de treinamento y_t é calculada usando a estrutura candidata S e um conjunto de parâmetros que é estimado a partir do conjunto $D - \{y_t\}$, onde D é o conjunto de treinamento:

$$LLOO(S|D) = \prod_t P(y_t|S, D - \{y_t\}) = \prod_t \prod_j P(x_{jt}|\pi^S_{jt}, D - \{y_t\}).$$

Para estimar o conjunto de parâmetros (as probabilidades contidas pelas tabelas de probabilidades condicionais), usa-se a técnica descrita na Seção 2.6.1, com a diferença de que o conjunto de treinamento muda para cada instância y_t a ser avaliada:

$$P_t(x_{ji}|\pi_{jk}) = \frac{N^{D-\{y_t\}}_{jki} + \lambda}{N^{D-\{y_t\}}_{jk} + \lambda M_j}, \tag{2.26}$$

onde $N^{D-\{y_t\}}_{jki}$ é o número de observações simultâneas de x_{ji} e π_{jk} no conjunto $D - \{y_t\}$ e $N^{D-\{y_t\}}_{jk}$ é o número de observações de π_{jk} no conjunto $D - \{y_t\}$.

Não é realmente preciso executar um processo de treinamento completo para cada instância. Podemos obter as freqüências no conjunto de treinamento original D e usar a relações:

$$N_{jk}^{D-\{y_t\}} = \begin{cases} N_{jk}^{D} - 1 & \text{se } \pi_{jk}^{S} = \pi_{jt}^{S}; \\ N_{jk}^{D} & \text{caso contrário;} \end{cases}$$

$$N_{jki}^{D-\{y_t\}} = \begin{cases} N_{jki}^{D} - 1 & \text{se } \pi_{jk}^{S} = \pi_{jt}^{S} \wedge x_{ji} = x_{jt}; \\ N_{jki}^{D} & \text{caso contrário;} \end{cases}$$

onde π_{jt}^{S} é o conjunto de valores assumidos por Π_j na instância y_t e x_{jt} é o valor assumido por X_j na instância y_t.

Dois outros critérios difundidos para seleção de estruturas de rede não tratados neste trabalho são o *Bayesian information criterion* (BIC) [70] e o *Akaike's Information Criterion* (AIC) [2].

Escolha do espaço e do algoritmo de busca

Se dispuséssemos de uma quantidade ilimitada de tempo de processamento, dado um critério de comparação entre estruturas, a seleção da melhor delas seria trivial. Simplesmente testaríamos todas as estruturas possíveis. Em uma situação real, como o número de estruturas possíveis é exponencial no número de variáveis, geralmente é preciso adotar alguma medida que reduza o número de estruturas efetivamente testadas.

O *naïve Bayes* usa uma estrutura fixa e logo dispensa qualquer mecanismo de busca. O *tree augmented naïve Bayes* (TAN) limita seu espaço de busca a um conjunto de estruturas que correspondem a estrutura do *naïve Bayes* acrescida de arestas que sozinhas formariam uma árvore cujos nós seriam os atributos explanatórios.

Dentro desse espaço de busca, Friedman et al. [34] exibem um algoritmo que identifica a estrutura TAN que atribui máxima verossimilhança aos dados em tempo quadrático no número de atributos no problema e linear no número de instâncias de treinamento. Esse algoritmo é uma extensão do algoritmo apresentado por Chow e Liu [17] para aproximação de distribuições de probabilidades discretas usando árvores de dependência.

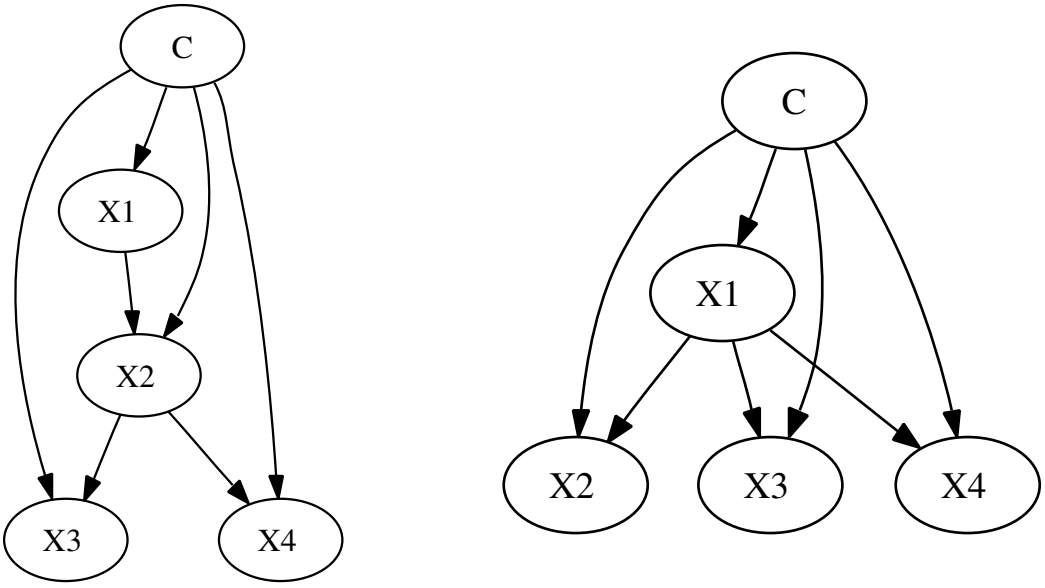

Figura 2.6: duas estruturas do tipo *tree augmented naïve Bayes*

Hamine e Helman, mostram como encontrar uma estrutura ótima do tipo *forest augmented naïve Bayes* (FAN) [40], uma generalização do TAN e posteriormente, mostram o mesmo com relação estruturas do tipo *selected forest augmented naïve Bayes*(SFAN) [41], uma generalização do FAN.

Em espaços de busca mais amplos, selecionar estruturas ótimas não é computacionalmente viável e em geral, métodos heurísticos são empregados. Partindo de uma estrutura inicial, que pode ser vazia (sem nenhuma aresta), ou de uma estrutura simples como a do *naïve Bayes* os métodos realizam transformações como acrescentar uma aresta, remover uma aresta e reverter a direção de uma aresta, e constantemente avaliam as estruturas frente a um critério de qualidade como os já citados. Estas avaliações permitem a escolha da operação que produz o maior ganho imediato e conduzem naturalmente a algoritmos gulosos como o escalada de montanha, o K2[3] [18] e o B [11].

O maior problema com o uso de algoritmos gulosos é a possibilidade de que o processo de otimização fique preso em um ótimo local. Existem várias estratégias para reduzir os problemas com mínimos locais como escalada de montanha repetida, busca genética [8], busca tabu [8] e arrefecimento simulado [8].

Além de estratégias baseadas em busca e critérios de comparação (*search and scoring*) como as descritas anteriormente, estruturas de rede também podem ser escolhidas através de testes de dependência como ocorre, por exemplo, em [15, 73].

[3]O nome K2 designa tanto um critério de comparação entre estruturas quanto um algoritmo de busca.

As redes Bayesianas são amplamente usadas na solução de problemas de classificação de padrões, mas como vimos CPTs grandes podem ser um problema. No próximo capítulo, apresentaremos um modelo que tanto pode ser uma alternativa a BNs quanto ser usado dentro delas como um substituto para CPTs.

Capítulo 3

Hierarchical Pattern Bayes

Neste Capítulo, apresentamos o *hierarchical pattern Bayes* (HPB), um novo método de classificação baseado em uma hierarquia de padrões. Descrevemos a construção desta hierarquia e a forma como o HPB a utiliza para fazer estimativas de probabilidade confiáveis.

Na Seção 3.1, introduzimos as equações que definem o HPB. Na Seção 3.2, discutimos a tendência de uma de suas equações a produzir estimativas excessivamente confiantes e descrevemos o mecanismo de balanceamento que utilizamos na atenuação desse efeito. Na Seção 3.3 analisamos o HPB, descrevendo de forma intuitiva o efeito esperado para cada uma de suas características. Na Seção 3.4 mostramos como o HPB pode substituir tabelas de probabilidades condicionais em redes Bayesianas. Na Seção 3.5, exibimos estratégias para seleção de parâmetros para o HPB e na Seção 3.6 analisamos a complexidade computacional do HPB.

O HPB é um método de classificação de padrões que funciona em problemas onde todos os atributos são categóricos. Dados um padrão w e um conjunto de treinamento, D, de pares (y_t, c_t), onde y_t é a t-ésima instância em D e c_t é o rótulo de classe da t-ésima instância em D, o HPB calcula $P(c_r|w)$ para toda classe c_r, onde um padrão segue a definição 2.4, abaixo repetida:

Definição 2.4 *Um padrão é um conjunto de pares da forma* $(Atributo = Valor)$*, onde cada atributo pode aparecer no máximo uma vez.*

Um atributo que não está presente no padrão é chamado de faltante. Antes de mostrar os detalhes do HPB, precisamos de mais algumas definições:

Definição 3.1 *Um padrão* y *é mais genérico que um padrão* w *se e somente se* $y \subseteq w$*.*

Se y é mais genérico que w, dizemos que w satisfaz y. Se uma instância y_t é representada por um padrão y e y satisfaz w, também dizemos que y_t satisfaz w.

Definição 3.2 *Um padrão y é estritamente mais genérico que um padrão w se e somente se $y \subset w$.*

Definição 3.3 *O nível de um padrão w, $level(w)$ é o número de atributos definidos em w.*

Definição 3.4 *$G(w)$ é o conjunto de todos os padrões estritamente mais genéricos que w.*

3.1 Modelo hierárquico

O HPB calcula a probabilidade posterior $P(c_r|w)$, usando uma estratégia que é similar a estimativa quase direta (ADE), apresentada na Equação (2.22), porém, as probabilidades preditivas previas são consideradas iguais a $P(c_r|G(w))$.

Os parâmetros da distribuição de probabilidade prévia de Dirichlet são dados por $\alpha_r = S \cdot P(c_r|G(w))$, onde S é um coeficiente de suavização. Conseqüentemente,

$$P(c_r|w) = \frac{N_{wr} + S \cdot P(c_r|G(w))}{N_w + S}, \tag{3.1}$$

onde N_w é o número de instâncias no conjunto de treinamento satisfazendo o padrão w e N_{wr} é o número de instâncias no conjunto de treinamento satisfazendo o padrão w cujo rótulo de classe é c_r.

Dada a Equação (3.1), o problema torna-se calcular $P(c_r|G(w))$. Nossa idéia básica é escrever $P(c_r|G(w))$ como uma função das várias probabilidades $P(c_r|w_j)$, onde os w_j são padrões pertencentes a $G(w)$ e calcular cada $P(c_r|w_j)$ recursivamente, usando a Equação (3.1). Para tornar isso possível, precisamos de mais uma definição:

Definição 3.5 *$g(w)$ é o subconjunto de $G(w)$ que contém todos os elementos de nível igual a $level(w) - 1$.*

Por exemplo, se w é $\{A = a, B = b, C = c\}$, $g(w)$ é

$$\{\ \{B = b, C = c\}, \{A = a, C = c\}, \{A = a, B = b\}\ \}.$$

Consideramos que apenas $g(w)$ influencia $P(c_r|G(w))$ diretamente, de modo que $P(c_r|G(w)) = P(c_r|g(w))$. A influência dos demais padrões em $G(w)$ é capturada pelo processo recursivo.

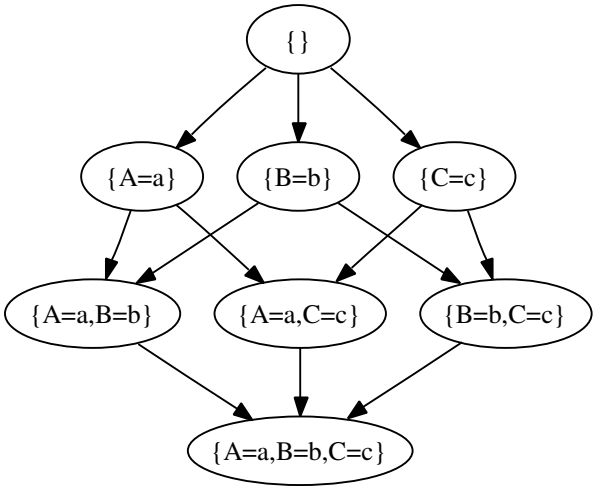

Figura 3.1: exemplo de estrutura usada pelo HPB

A Figura 3.1 mostra uma hierarquia de padrões, onde A, B e C são os atributos. Cada padrão é representado por um nó e o conjunto de pais de w no grafo apresentado na Figura 3.1 é $g(w)$. O HPB combina as distribuições de probabilidade preditivas posteriores, $P(c_r|w_j)$, da classe dado cada pai, w_j, de um padrão w para construir a distribuição de probabilidade preditiva prévia para a classe dado w, $P(c_r|g(w))$.

O primeiro passo para escrever $P(c_r|g(w))$ como uma função de todas as $P(c_r|w_j)$ é aplicar a regra de Bayes:

$$\begin{aligned} P(c_r|g(w)) &= \frac{P(g(w)|c_r)P(c_r)}{P(g(w))} \\ &\propto P(w_1, w_2, \ldots, w_L|c_r)P(c_r), \end{aligned}$$

onde w_1,w_2,...,w_L são os elementos de $g(w)$. Então, aproximamos a probabilidade conjunta $P(w_1, w_2, \ldots, w_L|c_r)$ pelo produto das probabilidades marginais:

$$P'(c_r|g(w)) \propto P(c_r) \prod_{j=1}^{L} P(w_j|c_r), \tag{3.2}$$

mas aplicamos um mecanismo de balanceamento (explicado na Seção 3.2):

$$P(c_r|g(w)) \propto P'(c_r|g(w)) + B \cdot P(c_r), \tag{3.3}$$

onde B é um coeficiente de balanceamento.

Dadas as Equações (3.2) e (3.3) precisamos calcular $P(w_j|c_r)$. Aplicando a regra de Bayes novamente,

$$P(w_j|c_r) = \frac{P(c_r|w_j)P(w_j)}{P(c_r)}. \tag{3.4}$$

Podemos estimar $P(c_r)$ usando a abordagem de máxima verossimilhança: $P(c_r) = N_r/N$, onde N_r é o número de exemplos no conjunto de treinamento cujo rótulo de classe é c_r, e N é o número total de exemplos no conjunto de treinamento.

Se variável classe for binária, podemos supor que ambas as classes possíveis serão relativamente bem representadas e que esta estratégia funcionará bem, mas se a variável classe não tem cardinalidade tão baixa, é melhor empregar uma distribuição de probabilidade prévia não informativa:

$$P(c_r) = \frac{N_r + S^{NI}/M_c}{N + S^{NI}},$$

onde M_c é o número de classes e S^{NI} é uma constante de suavização.

Quando substituimos $P(w_j|c_r)$ pelo lado direito da Equação (3.4) na Equação (3.2) podemos eliminar o fator $P(w_j)$, pois ele é idêntico para todas as classes:

$$\begin{aligned} P'(c_r|g(w)) &\propto P(c_r)\prod_{j=1}^{L} P(w_j|c_r) \\ &\propto P(c_r)\prod_{j=1}^{L} \frac{P(c_r|w_j)P(w_j)}{P(c_r)} \\ &\propto P(c_r)\prod_{j=1}^{L} \frac{P(c_r|w_j)}{P(c_r)}, \end{aligned}$$

então não precisamos nos preocupar com ele.

Como w_j é um padrão, a estimativa de $P(c_r|w_j)$ pode ser feita recursivamente, usando a Equação (3.1). A recursão acaba quando $g(w)$ contém apenas o padrão vazio. Nesse caso, $P(c_r|g(w)) = P(c_r|\{\{\}\}) = P(c_r)$.

3.2 Mecanismo de balanceamento

Apesar de suas fortes suposições de independência, o *naïve Bayes* (Seção 2.4) é conhecido por ter bom desempenho em vários domínios quando apenas a taxa de erros de classificação é considerada. Contudo, o NB tem uma tendência a produzir probabilidades extremas que pode ser atenuada através de um mecanismo de balanceamento (Seção 2.4).

Usando a Equação (3.2), estamos fazendo suposições de independência mais fortes que o NB. O *naïve Bayes* assume que os atributos são independentes dada a classe, o que é pelo menos possível. A Equação (3.2) assume que algumas agregações de atributos são independentes dada a classe. Como muitas destas agregações têm atributos em comum sabemos que estas suposições são falsas. A maior conseqüência de nossas suposições fortes e irrealistas são estimativas ainda mais extremas que as feitas pelo NB. Isso é parcialmente

compensado pelo mecanismo de balanceamento na Equação (3.3). Ele é mais simples que os apresentados em [6, 79] e é não supervisionado. Isso o torna muito rápido e fácil de aplicar a cada passo do HPB.

Apenas combinamos linearmente o resultado da Equação (3.2) e $P(c_r)$. Fazemos isso considerando que se as estimativas são mais extremas que as probabilidades reais tanto perto de 0 quanto perto de 1, elas têm que coincidir com as probabilidades reais em algum ponto intermediário. Acreditamos que esse ponto seja próximo a $P(c_r)$.

Probabilidades extremas são produzidas quando a evidência a favor ou contra uma classe é considerada duas vezes. $P(c_r)$ é um ponto onde ou não há evidência alguma, ou a há evidência em direções conflitantes de modo que o efeito é nulo. Assim, tal ponto não pode ser considerado extremo. Nosso mecanismo de balanceamento atenua as probabilidades quando são extremas sem afetá-las no ponto em que acreditamos que já esteja correta.

Com alguma manipulação algébrica a Equação (3.3) pode ser escrita como

$$P''(c_r|g(w)) = (1 - A) \cdot P'(c_r|g(w)) + A \cdot P(c_r),$$

onde $A = B/(1 + B)$. No lado esquerdo da equação, colocamos $P''(c_r|g(w))$ ao invés de $P(c_r|g(w))$, apenas para tornar explícito o fato de que o resultado da Equação (3.3) é na verdade uma segunda estimativa, embora o HPB o utilize como se fosse a probabilidade real.

Na Figura 3.2 mostramos o efeito do balanceamento.

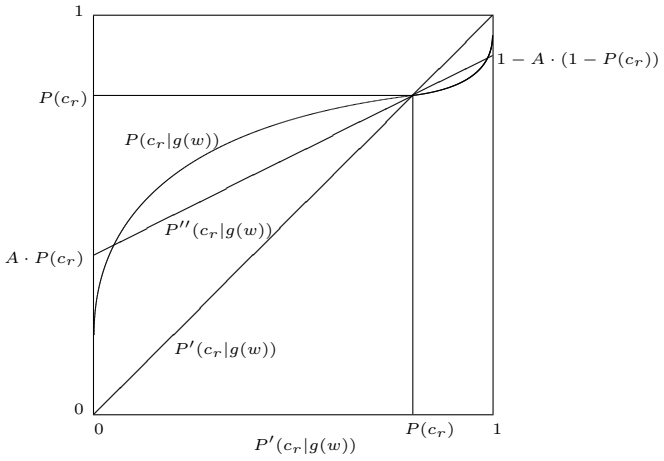

Figura 3.2: efeito do balanceamento linear sobre probabilidades extremas

No eixo horizontal a estimativa não calibrada $P'(c_r|g(w))$ é representada. A linha curva representa a probabilidade real, $P(c_r|g(w))$, como uma função de $P'(c_r|g(w))$. Como toda

a informação a respeito de $P'(c_r|g(w))$ vem de um conjunto finito de dados, tal função nunca atinge 1 nem 0. Quando $P'(c_r|g(w))$ está próxima de 0, $P(c_r|g(w))$ não está tão próxima. O mesmo ocorre quando $P(c_r|g(w))$ está próxima de 1.

A linha reta de 45^o representa o que seria nossa estimativa final caso não empregássemos nenhum balanceamento, i.e., a prória $P'(c_r|g(w))$. A outra linha reta oblíqua é o resultado do nosso mecanismo de balanceamento, $P''(c_r|g(w))$. Ainda é uma aproximação linear, mas está bem mais próxima de $P(c_r|g(w))$ que $P'(c_r|g(w))$.

3.3 Análise do HPB

O HPB tenta explorar o conjunto de treinamento tanto quanto possível. Se existem L atributos o HPB começa seu trabalho capturando a influência dos padrões de nível L. Nesse nível, todas as interações entre atributos podem ser capturadas desde que existam instâncias de treinamento suficientes. Contudo, nenhum conjunto de treinamento é tão grande que possamos esperar que todos os padrões no nível L estejam bem representados. Na verdade, se existem atributos de alta cardinalidade, é mais provável que apenas uma minoria deles esteja. Para esta minoria, o nível L domina a Equação (3.1) e as probabilidades prévias não são muito importantes. Por outro lado, elas são críticas para a grande maioria dos casos onde os padrões de nível L não estão bem representados no conjunto de treinamento.

Então, o HPB passa ao nível $L-1$. Nesse nível, uma fração maior de padrões está bem representado e ainda é possível capturar a maior parte das interações entre atributos. Muitos padrões de nível $L-1$ ainda não estão bem representados e é preciso recorrer a padrões de nível mais baixo. Quanto menor é o nível de um padrão, mais fraca é a capacidade de capturar interações, mas menos comuns são os problemas com amostras pequenas.

A Equação (3.1) combina a influencia de padrões de níveis diferentes de modo a que os padrões mais específicos sempre dominem se estiverem bem representados. A Equação (3.2) combina padrões de mesmo nível fazendo fortes suposições de independência. Tais suposições são o preço a pagar para capturar a influência de todos os padrões em todos os níveis. Esse preço é atenuado pelo mecanismo de balanceamento na Equação (3.3).

Como a população de instâncias (de treinamento ou de testes) satisfazendo um padrão w, é uma subpopulação contida pela população de instâncias satisfazendo $w_j, \forall w_j \in g(w)$, podemos dizer que o HPB usa resultados pré-calculados a partir populações mais amplas para construir probabilidade prévias para as inferências relativas a populações mais estritas. Esta é a estratégia geral dos modelos Bayesianos hierárquicos empíricos. O HPB é portanto um modelo empírico e não um modelo Bayesiano completo.

Em [38, 4, 74] modelos hierárquicos completos são apresentados, mas eles contém apenas dois níveis. O HPB lida com uma hierarquia de múltiplos níveis recursivamente e também com a fato de cada subpopulação está contida não em uma, mas em várias superpopulações que se sobrepõem. Esse último fato torna mais difícil o desenvolvimento de um modelo completo que calcule todas as distribuições de probabilidade envolvidas de um só vez considerando toda a evidência disponível.

3.4 O HPB como substituto para tabelas de probabilidades condicionais

O HPB foi projetado para ser um classificador independente bem adaptado a nosso domínio alvo e similares. Ele usa tempo e espaço exponenciais no número de atributos e portanto não pode ser aplicado diretamente a domínios onde existam muitos atributos.

Contudo, o HPB pode ser usado no lugar de tabelas de probabilidades condicionais de redes Bayesianas. O número de pais de qualquer nó em uma rede Bayesiana tem que ser pequeno porque o tamanho da tabela é exponencial no número de atributos. Além disso, tempo e espaço freqüentemente não são o fator limitante para o número de pais. Mais pais normalmente significam probabilidades menos confiáveis [47] e não é incomum limitar seu número a dois [35, 47, 40]. Assim, se o HPB produzir estimativas melhores, ele, na verdade, permitirá a adição de mais pais a cada nó na rede.

Se a estrutura da rede Bayesiana for dada, o uso do HPB em lugar da tabela de probabilidades condicionais de um nó, X_j, qualquer é trivial. Para calcular, $P(x_{jk}|\pi_{ji})$ basta agir como se $c_r = x_{ji}$ e $w = \pi_{jk}$, ignorar todos os demais atributos e usar o HPB para calcular $P(c_r|w)$.

Se a estrutura da BN precisar ser aprendida a partir dos dados, é preciso escolher um critério de seleção de estrutura que possa funcionar bem com o HPB. Nós propomos o uso do Critério de máxima verossimilhança logarítmica em validação cruzada do tipo *leave one out*.

$$LLLOO = \sum_t \log P(y_t|S, D - \{y_t\}) = \sum_t \sum_j \log P(x_{jt}|\pi_{jt}^S, D - \{y_t\}),$$

onde D é o conjunto de treinamento, y_t é a t-ésima instância de D, S é a estrutura sendo avaliada, x_{jt} é o valor assumido pelo atributo X_j na instância y_t, π_{jt}^S é o conjunto de valores assumido, em y_t, pelos pais de X_j em S e $P(x_{jt}|\pi_{jt}^S, D - \{y_t\})$ é o valor calculado pelo HPB para $P(x_{jt}|\pi_{jt}^S)$ usando $D - \{y_t\}$ como conjunto de treinamento.

O HPB usa o conjunto de treinamento apenas através das freqüências N_{wr} e N_w na Equação (3.1). Para computação rápida de $LLLOO$, podemos medir estas freqüências

em D e usar as relações:

$$N_w^{D-\{y_t\}} = \begin{cases} N_w^D - 1 & \text{if } w \subset \pi_{jt}^S; \\ N_w^D & \text{caso contrário;} \end{cases}$$
$$N_{wr}^{D-\{y_t\}} = \begin{cases} N_{wr}^D - 1 & \text{if } w \subset \pi_{jt}^S \wedge x_{jr} = x_{jt}; \\ N_{wr}^D & \text{caso contrário.} \end{cases}$$

3.5 Seleção dos coeficientes empregados pelo HPB

As Equações (3.1) e (3.3) requerem respectivamente as especificações dos coeficientes S e B. Na classificação de uma única instância, estas equações são aplicadas várias vezes no cálculo de $P(c_r|w)$ para diferentes padrões, w. Os valores ótimos de S e B podem ser diferentes para cada padrão.

No caso dos coeficientes B, usamos uma heurística motivada pelo fato de que o nível de todos os padrões em $g(w)$ é $level(w) - 1$. Quanto maior é esse nível, mais atributos em comum têm as agregações de atributos, mais extremas são as estimativas de probabilidade e mais forte tem que ser o efeito do mecanismo de balanceamento. Assim, fizemos o coeficiente B na Equação (3.3) igual a $b(level(w)-1)$, onde b é um constante experimental.

No caso dos coeficientes S podemos empregar um mecanismo de otimização, ou, para treinamento mais rápido, definir S como uma constante.

Propomos o uso da área sob a curva de acerto [82] como critério de otimização. A curva de acerto de um classificador C sobre o conjunto de dados D é a função, $h_{C,D}(r)$, onde r é uma taxa de seleção (um número real no intervalo $[0, 1]$). O classificador é usado para atribuir a cada exemplo, y_t em D a probabilidade de que y_t seja uma instância positiva. O valor de $h_{C,D}(r)$ é o número de instâncias positivas dentre as $r \cdot |D|$ instâncias que foram consideradas as mais prováveis de ser positivas pelo classificador.

Empregamos curvas de acerto ao invés das mais populares *Receiver Operating Characteristic Curves* (ROC) [29], porque elas refletem o interesse de um usuário de um sistema de detecção de fraudes diretamente. Dada uma taxa de seleção que reflete os recursos humanos disponíveis, deseja-se maximizar a quantidade de fraudes detectadas.

Como o conceito de instância positiva só faz sentido quando a variável classe é binária, o processo aplica-se apenas a problemas que têm esta característica. Quando aplicado, o processo começa pela família de padrões mais genérica e caminha em direção às mais específicas, onde uma família de padrões é o conjunto contendo todos os padrões que definem os mesmos atributos (possivelmente com valores distintos).

Assumindo que os coeficientes S já tenham sido fixados para todas as famílias de padrões mais genéricas que a família F, apenas um coeficiente S precisa ser especificado para permitir o uso da Equação (3.1) no cálculo de $P(c_r|w)$, onde w é qualquer padrão pertencente a F.

Esse coeficiente é selecionado de forma a maximizar a área sob a curva de acerto que é induzida quando, usando validação cruzada do tipo *leave one out* calculamos $P(c_0|w)$ para todos os padrões de treinamento, w, em F, onde c_0 é a classe definida como *positiva*.

Calcular $P(c_0|w)$ usando validação cruzada do tipo *leave one out*, significa, como explicado na Seção 3.4, simplesmente subtrair 1 de algumas freqüências usadas pela Equação (3.1).

3.6 Complexidade computacional

A fase de treinamento da versão do HPB que emprega coeficientes S constantes, consiste apenas em contabilizar as freqüências usadas pela Equação (3.1). É fácil ver que cada instância, u, representada por um padrão, w, no conjunto de treinamento, D, requer o incremento de exatamente 2^L freqüências (uma para cada padrão em $G(w) \cup \{w\}$). Assim, o tempo de treinamento do HPB é

$$O(N_{tr} \cdot 2^L),$$

onde N_{tr} é o número de instâncias de treinamento.

A fase de teste (ou de aplicação) do HPB requer que, para cada instância, u, representada por um padrão, w, a distribuição de probabilidade para classe seja computada dados 2^L padrões (todos os padrões em $G(w) \cup \{w\}$. Como cada computação é proporcional ao número de classes, o tempo de teste do HPB é

$$O(N_{ts} \cdot M_c \cdot 2^L),$$

onde N_{ts} é o número de instâncias de teste e M_c é o número de classes.

Note que, em ambos os casos, o tempo de execução do HPB é exponencial no número de atributos, linear no número de instâncias e independente da cardinalidade dos atributos (exceto pelo atributo classe).

Quando os coeficientes S são escolhidos pelo processo de otimização descrito na Seção 3.5, o tempo de teste do HPB não muda, mais o treinamento requer que, para cada família de padrões, vários coeficientes candidatos sejam testados. Existem 2^L famílias de padrões e cada teste requer a aplicação do HPB a todas as instâncias de treinamento. Assim, o tempo de treinamento do HPB torna-se

$$O(N_{tr} \cdot 2^L + N_{cand} \cdot 2^L \cdot N_{tr} \cdot M_c \cdot 2^L) = O(N_{cand} \cdot N_{tr} \cdot M_c \cdot 2^{2L}),$$

onde N_{cand} é o número de coeficientes candidatos na escolha de um coeficiente S, o que depende do algoritmo de busca.

O HPB precisa guardar, para cada instância de treinamento, menos de 2^L freqüências. Assim, o uso de espaço do HPB é

$$O(N_{tr} \cdot 2^L).$$

Capítulo 4

Resultados Experimentais

Neste capítulo, avaliamos o HPB em três contextos diferentes:

- *detecção de erros de classificação fiscal*: O problema de motivação para o HPB. Um problema importante para a RFB, onde quatro atributos de alta cardinalidade, que supõe-se interagir de forma relevante, são usados para prever uma variável classe binária;
- *previsão de comportamento conjunto*: outro problema originado da RFB, onde dois atributos de alta cardinalidade são usados para prever um terceiro atributo de alta cardinalidade;
- *o HPB como um substituto geral para tabelas de probabilidades condicionais*: testes sobre vários conjuntos de dados da UCI [5] comparando o HPB à CPTs e outras representações para a distribuição de probabilidade condicional de um nó de uma BN dados os seus pais.

Em todos os casos, os métodos de classificação foram testados pela ferramenta *Weka Experimenter* [77] usando validação cruzada em cinco subconjuntos. Para usar o Weka, fizemos as seguintes implementações dentro de sua estrutura de classes:

- classificadores: HPB, Noisy-OR seguindo [32], *naïve Bayes* com coeficiente de suavização variável, grafo de decisão seguindo [16], árvore de decisão seguindo [37], tabela default seguindo [37];
- extrator de atributos: *agglomerative information bottleneck* seguindo [72];
- estimador de CPD para uso dentro de uma rede Bayesiana: adaptador genérico de classificadores produtores de probabilidade para estimadores de CPD;

- critério de seleção de estrutura de rede Bayesiana: máxima verossimilhança avaliada em validação cruzada do tipo *leave one out*;
- métricas de avaliação para classificadores: entropia cruzada média, curva de acerto.

A máquina de testes, em todos os casos, foi um Intel Core 2 Duo 6300, com 2 GB de memória primária.

4.1 Detecção de erros de classificação fiscal

A detecção de erros de classificação fiscal é o problema de motivação para o HPB. Considerando quatro atributos explanatórios: *classificação fiscal declarada* (NCMD), *importador* (IMP), *país de origem* (PAIS) e *unidade de entrada* (URF), precisamos estimar, para cada novo exemplo, a probabilidade de que ele envolva um erro de classificação, i.e., a probabilidade de que NCMD não seja o código correto para a mercadoria sendo importada.

Nosso conjunto de dados, fornecido pela RFB, contém 682226 exemplos de classificação correta (que chamamos de exemplos negativos) e 6460 exemplos de erro de classificação (exemplos positivos). Nesse conjunto de dados, o primeiro atributo assumiu 7608 valores distintos, o segundo, 18846 valores, o terceiro, 161 valores e o quarto, 80 valores.

Comparamos classificadores construídos usando os seguintes métodos:

- *HPB-OPT*: HPB com seleção de coeficientes S pelo processo de otimização descrito na Seção 3.5, estimativa direta, estimativa quase direta;
- *HPB*: HPB com coeficientes S fixos;
- *NB*: *naïve Bayes*;
- noisy-OR: BN com a estrutura apresentada na Figura 1.1 usando o *noisy-OR* como descrito em [32] em lugar de uma CPT;
- *TAN*: versão suavizada do *tree augmented naïve Bayes* como descrito em [34];
- *DE*: estimativa direta. BN com a estrutura apresentada na Figura 1.1 e CPTs tradicionais;
- *ADE*: estimativa quase direta. BN com a estrutura apresentada na Figura 1.1, CPTs tradicionais e o mecanismo de suavização descrito em [34];
- *DG*: grafo de decisão construído seguindo [16]. Porém, desviando do que foi feito em [16] usamos o DG como método de classificação independente, ao invés de dentro de BNs, substituindo CPTs.

- *BN-HC-DT*: BN com árvores de decisão em lugar de CPTs, usando escalada de montanha como algoritmo de busca e MDL como critério de comparação para escolha da estrutura da rede como descrito em [37];
- *PRIOR*: classificador trivial que atribui a probabilidade prévia a todas as instâncias.

Não fomos capazes de construir BNs com DGs em lugar de CPTs seguindo [16] porque o processo é demasiadamente longo, tomando todo um dia sem complementar um único subconjunto, com apenas a primeira parametrização a ser testada. Descobrimos que a construção de um DG torna-se muito lenta quando o nó da rede em questão tem alta cardinalidade e seus pais também têm alta cardinalidade. Pais de alta cardinalidade implicam em muitas operações de divisão/junção para comparar a cada passo do algoritmo e um filho de alta cardinalidade significa que cada comparação requer muitos cálculos.

Em muitas de nossas experiências anteriores com BNs com DGs aplicadas sobre conjuntos de dados menores [45, 46] nas estruturas globais escolhidas pelo algoritmo de busca descrito em [16], todos os quatro atributos explanatórios eram pais do atributo classe. Isso significa que se tivéssemos usado um grafo de decisão como classificador independente, teríamos exatamente os mesmos resultados. Assim, concluímos que valeria a pena testar um grafo de decisão independente no conjunto de dados ampliado que usamos agora. Como nossa variável classe é binária, o tempo de execução torna-se aceitável.

Tentamos diferentes parametrizações para cada método e ficamos com o conjunto de parâmetros que propiciou os melhores resultados, onde os melhores resultados são entendidos com sendo a maior área sob a curva de acerto, medida até a taxa de seleção de 20%. Ignoramos a área para taxas de seleção acima de 20%, porque todas as taxas de interesse prático estão abaixo desse limiar.

Além de usar a curva de acerto, comparamos as distribuições de probabilidade estimadas pelos modelos com a distribuição observada no conjunto de testes usando duas medidas: raiz quadrada do erro quadrático médio (RMSE) e entropia cruzada média (MCE):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}\sum_{r=1}^{M}(P'(c_{rt}) - P(c_{rt}))^2}{MN}}, \quad MCE = \frac{\sum_{i=1}^{N}\sum_{t=1}^{M} -P(c_{rt})\log_2 P'(c_{rt})}{MN},$$

onde N é o número de instâncias no conjunto de testes, M é o número de classes, $P'(c_{rt})$ é a probabilidade estimada de que a t-ésima instância pertença a classe c_r e $P(c_{rt})$ é a probabilidade real de que a t-ésima instância pertença a classe c_r. $P(c_{rt})$ é sempre 0 ou 1.

Vários dos métodos testados requerem a especificação de parâmetros e muitos desses são constantes reais. Usamos uma estratégia comum para escolher tais parâmetros:

1. com base na experiência, escolher um intervalo de busca, $SI = [beg, end]$, dentro do qual acreditamos que a constante ideal esteja;

2. construir a enumeração SE contendo todas as potências de 10, todas as metades de potências de 10 e todos os quartos de potências de 10 em SI;

3. tentar todas as constantes em SE. Se o método precisar de mais de um parâmetro, tentar todas as combinações exaustivamente;

4. se a constante ótima, C, estiver no meio de SE, ficamos com C;

5. se a constante ótima, C, for um dos valores extremos de SE expandimos SE adicionando mais um valor e tentamos novamente. O valor a ser adicionado é o número real mais próximo de C que não pertença a SI e que seja uma potência de 10, uma metade de potência de 10 ou um quarto de potência de 10.

Nos restringindo a potências de 10, a metades de potências de 10 e quartos de potências de 10 evitamos ajuste fino e testamos diferentes ordens de grandeza para cada constante.

Os coeficientes de suavização empregados pelo HPB-OPT são todos automaticamente selecionados. Tal seleção envolve uma validação cruzada do tipo *leave one out* que ocorre totalmente dentro do conjunto de treinamento corrente (a validação cruzada de 5 subconjuntos varia o conjunto de treinamento corrente). Os coeficientes B foram escolhidos pela heurística descrita na Seção 3.5 e pela constante b. A escolha de b foi feita começando com $SI = [0.5, 2.5]$.

O HPB requer a especificação da constante S, que é usada diretamente e da constante b que define os coeficientes B através da heurística na Seção 3.5. A escolha de b foi feita começando com $SI = [0.5, 2.5]$. Para escolher S, definimos, $s = S/NumClasses = S/2$, e escolhemos s começando com $SI = [1.0, 10.0]$. A razão para a introdução da constante s é manter o padrão da ferramenta Weka no que tange a constantes de suavização.

Em ambas as versões do HPB o valor de S^{NI} foi definido como 0. Não há razão para acreditar que este seja o valor ideal, mas como o conjunto de dados é grande e temos apenas duas classes podemos concluir que sua relevância será muito pequena e evitar o esforço de otimizá-lo em conjunto com os demais parâmetros.

Grafos de decisão têm quatro parâmetros, a constante de suavização e três valores booleanos definindo o estado de ativação de cada uma das possíveis operações, que são *divisões completas*, *divisões binárias* e *junções*. A constante de suavização foi escolhida começando com $SI = [0.01, 1.0]$. Sempre mantivemos as divisões completas ativadas e tentamos exaustivamente todas as variações resultantes do processo de habilitar e desabilitar divisões binárias e junções, para cada constante de suavização.

O *noisy-OR* e o PRIOR não tem parâmetros. A otimização de todos os demais métodos envolve apenas a constante de suavização, que, em todos os casos, foi escolhida começando de $SI = [0.01, 2.5]$.

Abaixo reportamos os valores ótimos para os parâmetros de cada método:

- *HPB-OPT*: $b = 1.0$;
- *HPB*: $s = 5.0$ e $b = 1.0$;
- *NB*: $s = 0.1$;
- *TAN*: $s = 0.25$;
- *ADE*: $s = 0.01$;
- *DE*: $s = 2.5$;
- *DG CBM*: $s = 0.05$, divisões completas, divisões binárias e junções habilitadas;
- *BN-HC-DT*: $s = 0.01$;
- *BN-HC-DF*: $s = 0.025$;

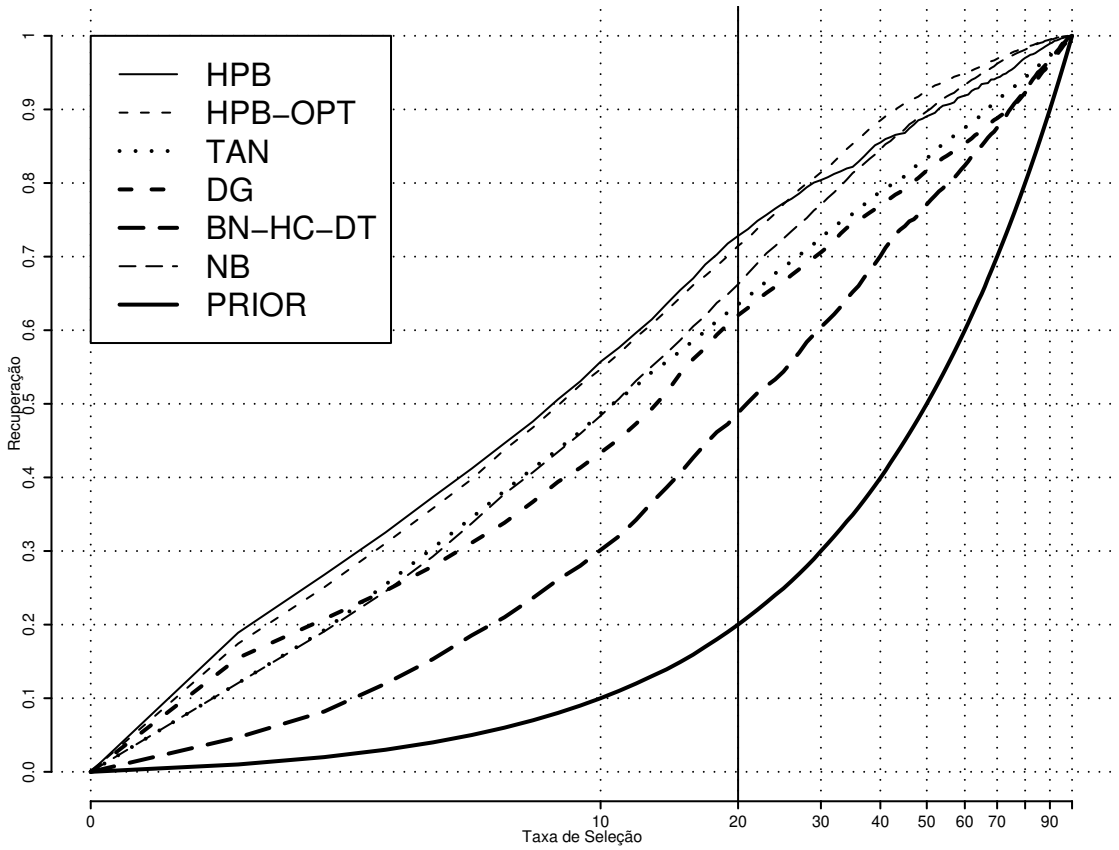

Figura 4.1: detecção de erros de classificação - curvas de acerto (para evitar poluição apresentamos apenas um subconjunto dos métodos testados)

	1%	2%	5%	10%	20%
HPB	18.89±0.77	26.77±0.57	41.20±1.10	55.72±1.82	72.81±1.69
HPB-OPT	17.41±1.55	25.08±1.10	39.76±0.61	54.70±1.44	71.45±1.74
TAN	12.06±0.59	19.26±0.70	34.52±1.32	48.70±1.82	63.52±1.06
ADE	13.32±1.37	15.06±1.46	20.70±1.65	30.61±1.18	49.39±1.06
DE	8.32±0.69	10.42±0.73	16.49±0.73	26.58±0.56	45.54±0.58
DG	15.47±1.29	20.76±0.61	31.12±1.61	43.36±2.19	62.03±1.41
BN-HC-DT	4.68±0.23	8.20±0.62	18.54±0.51	30.14±1.13	48.78±1.32
BN-HC-DF	4.44±0.39	8.22±0.49	18.45±0.44	30.06±0.30	47.45±0.98
NB	12.06±0.35	19.07±0.87	33.76±0.68	48.37±1.70	66.24±1.56
Noisy-Or	12.86±0.46	20.36±1.13	33.45±0.73	47.36±1.69	63.26±1.52
PRIOR	1.00±0.00	2.00±0.00	5.00±0.00	10.00±0.00	20.00±0.00

Tabela 4.1: detecção de erros de classificação - recuperação a diferentes taxas de seleção

	AUC	AUC20	RMSE	MCE	TR	TS
HPB	83.17±0.73	53.34±1.37	9.86±0.03	3.47±0.04	9.84±0.55	7.79±1.03
HPB-OPT	84.47±0.70	52.21±1.21	!10.06±0.05	!3.67±0.05	!517.66±4.76	!11.43±1.50
TAN	!78.10±0.72	!45.78±1.17	!11.55±0.05	!4.84±0.07	!43.67±0.12	1.34±0.01
ADE	!74.96±0.19	!31.43±1.25	!10.05±0.06	!4.59±0.04	4.04±0.12	0.34±0.09
DE	!72.33±0.57	!27.37±0.40	!34.62±0.02	!28.25±0.03	4.35±0.11	0.28±0.00
DG	!76.12±0.90	!42.89±1.55	!10.07±0.06	!5.19±0.30	!577.78±29.29	4.47±0.48
BN-HC-DT	!70.47±0.76	!29.95±0.85	9.60±0.00	!3.64±0.01	!125.01±1.21	!2446.17±113.19
BN-HC-DF	!69.79±0.76	!29.63±0.43	9.60±0.00	!3.65±0.01	!2433.02±20.20	!265.02±3.41
NB	!81.73±0.79	!46.33±1.08	!11.20±0.06	!4.19±0.06	4.79±0.06	0.28±0.00
Noisy-Or	!79.13±0.64	!45.07±1.09	!10.16±0.05	$!inf$±0.00	4.73±0.07	0.28±0.00
PRIOR	!50.48±0.01	!10.48±0.01	9.63±0.00	!3.83±0.00	4.87±0.46	0.28±0.00

Tabela 4.2: detecção de erros de classificação - outras medidas

Na Figura 4.1, escolhemos representar a $Recuperacao = N_{VerdadeirosPositivos}/N_{Positivos}$, no eixo vertical, ao invés do número absoluto de acertos, porque isso não altera a forma da curva e torna a interpretação mais fácil. Representamos a taxa de seleção em escala logarítmica para enfatizar o início das curvas. Na Tabela 4.1 representamos a recuperação para diferentes taxas de seleção.

Na Tabela 4.2 mostramos a área sob a curva de acerto (AUC), a área sob a curva de acerto até a taxa de seleção de 20% (AUC20), a raiz quadrada do erro quadrático médio (RMSE), a entropia cruzada média (MCE), o tempo de treinamento (TR) e o tempo de teste (TS) de cada método. A presença do símbolo *!* antes de um resultado significa que, de acordo com um teste T com significância de 5%, ele é significativamente pior que seu equivalente na primeira linha da tabela[1]. Como o HPB está na primeira linha, podemos ver que ele é significativamente melhor que todos os outros classificadores no que diz respeito à AUC, à AUC20 e à MCE. No que diz respeito à RMSE, o HPB não foi melhor que o BN-HC-DT, o BN-HC-DF e o PRIOR.

O método PRIOR é muito conservador, atribuindo a probabilidade prévia a todas as instâncias. Nesse conjunto de dados, tal estratégia resulta em bons MCE e RMSE. Por outro lado, o PRIOR não possui absolutamente nenhum poder de discriminação, considerando que todas as instâncias têm a mesma probabilidade de ser positivas. Na Figura 4.1 e na Tabela 4.1 podemos ver que isso resulta em seleção aleatória apenas verificando que a recuperação é sempre aproximadamente igual a a taxa de seleção.

O BN-HC-DT e o BN-HC-DF produziram curvas similares, o que pode ser visto na Tabela 4.1. Na Figura 4.1, a curva de acerto do BN-HC-DT foi desenhada e exceto pelo método PRIOR foi claramente a pior. A razão para isso é que a construção de DTs e DFs apresentada em [37] se mostrou muito conservadora, tendendo a preferir estruturas simples: DFs com poucas linhas e DTs como poucas divisões. Observando os resultados do método PRIOR não é surpreendente que esta estratégia conservadora resulte em uma

[1]Existem mecanismos mais sofisticados para comparação do desempenho de classificadores [22, 24]. Porém, como nosso conjunto de dados de estudo é bem grande, possuindo mais de 600000 instâncias, podemos usar o simples e popular teste T.

boa MCE, em uma boa RMSE e em uma curva de acerto insatisfatória em comparação com outros métodos.

A uma taxa de seleção de 1%, o ADE tem desempenho melhor que o NB, que o *noisy-OR* e que o TAN, mas para taxas de seleção mais altas ele é pior por uma margem bastante significativa. A razão para isso é que padrões críticos envolvendo todos os atributos são decisivos bem no início das curvas. O ADE trata todos os atributos em conjunto e se beneficia dos padrões críticos, mas logo o ADE se vê forçado a escolher entre padrões de teste para os quais não há padrões de treino exatamente iguais. Nesse ponto, o ADE começa a fazer escolhas aleatórias.

Usando grafos de decisão, os padrões mais críticos foram separados dos demais, o que resultou em uma melhoria significativa no início da curva de acerto em comparação com métodos como o NB, o *noisy-OR* e o TAN que não são capazes de capturar a influência de vários atributos de uma vez. Contudo, os demais padrões foram agrupados em poucas folhas. Dentro de uma folha, todos os padrões são considerados como tendo a mesma probabilidade de ser positivos. Isso resultou na queda do poder de discriminação para taxas superiores a 5%.

O HPB (em ambas as versões) se beneficia de padrões críticos envolvendo muitos ou mesmo todos os atributos, mas também considera a influência de padrões menos específicos. Como conseqüência, ele tem bom desempenho para qualquer taxa de seleção. A versão do HPB que usa um valor fixo para os coeficientes S é pior que o NB para taxas de seleção acima de 45%, mas, nesse ponto, a Recuperação já é de 87% para ambos os métodos e a diferença entre eles nunca é significativa. Exceto por sua versão simplificada, o HPB-OPT é melhor que qualquer outro método para todas as taxas de seleção, mas o processo de otimização o torna cinquenta vezes mais lento que o HPB com S fixo.

Como a cardinalidade dos atributos é um problema neste domínio, decidimos também testar todos os métodos de classificação sobre um conjunto de dados transformado onde a cardinalidade de todos os atributos é reduzida pelo *Agglomerative Information Bottleneck Method* (AIBN) [72]. Para evitar que o AIBN use informação proveniente dos conjuntos de testes, nós implementamos um meta classificador no Weka que aplica o AIBN imediatamente antes de treinar o classificador real e após cada conjunto de treinamento ter sido separado de seu conjunto de testes associado no processo de validação cruzada de cinco subconjuntos.

O AIBN reduz a cardinalidade de um atributo sucessivamente executando a junção de dois valores que resulte na menor perda de informação mútua. O processo pode continuar até que reste apenas um valor, mas pode ser parado em qualquer ponto conveniente. Escolhemos limitar a perda de informação mútua a $1e-4$, um valor bastante pequeno. Apesar disso, a redução de cardinalidade foi acentuada. A Tabela 4.3 mostra as cardinalidades antes e depois da aplicação do AIBN.

Atributo	Cardinalidade Original	Cardinalidade Final
NCMD	7608	101
IMP	18846	84
PAIS	161	50
URF	80	28

Tabela 4.3: redução de cardinalidade usando AIBN

Por causa da cardinalidade mais baixa dos atributos resultantes, foi possível testar BNs com DGs em lugar de CPTs ao invés de DGs independentes.

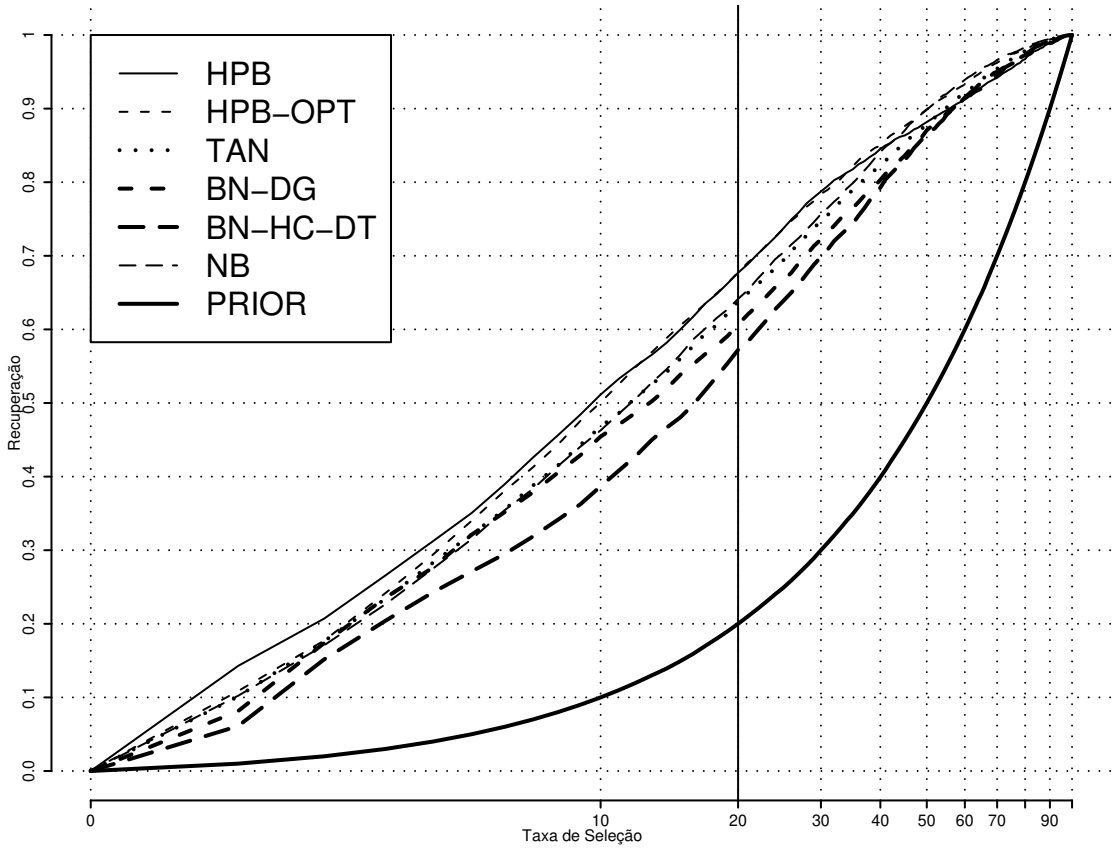

Figura 4.2: detecção de erros de classificação com redução de cardinalidade - curvas de acerto (para evitar poluição apresentamos apenas um subconjunto dos métodos testados)

	1%	2%	5%	10%	20%
HPB	14.28±0.40	20.72±0.47	35.05±0.92	51.14±2.00	67.70±2.04
HPB-OPT	10.86±0.51	17.74±0.73	34.00±1.06	50.08±2.09	67.76±2.12
TAN	10.11±0.67	17.66±0.90	32.15±1.54	46.78±1.70	63.78±0.76
ADE	13.10±0.53	16.36±1.20	20.42±1.34	33.82±1.44	55.72±1.06
DE	8.28±0.59	11.17±0.64	19.40±0.64	32.82±0.47	56.66±0.63
BN-DG	8.14±0.46	17.40±0.66	32.12±1.38	45.44±1.12	60.66±1.48
BN-HC-DT	6.10±0.53	15.18±0.19	27.12±1.66	38.68±2.26	57.21±1.99
BN-HC-DF	6.22±0.45	14.94±0.15	26.33±0.56	37.92±1.53	55.05±1.21
NB	10.22±0.55	17.09±0.83	31.50±0.84	46.28±1.73	64.14±1.85
Noisy-Or	4.84±0.26	14.80±0.52	29.79±0.87	44.70±1.72	62.78±1.97
PRIOR	1.00±0.00	2.00±0.00	5.00±0.00	10.00±0.00	20.00±0.00

Tabela 4.4: detecção de erros de classificação com redução de cardinalidade - recuperação a diferentes taxas de seleção

	AUC	AUC20	RMSE	MCE	TR	TS
HPB	81.51±0.72	48.07±1.43	10.37±0.03	3.85±0.04	8.30±0.07	5.74±0.03
HPB-OPT	82.16±0.85	47.28±1.44	9.56±0.01	3.50±0.01	!148.66±2.73	!6.32±0.02
TAN	!80.27±0.61	!44.21±1.15	!11.03±0.05	!4.19±0.05	!18.40±0.53	1.40±0.02
ADE	!75.90±0.52	!35.05±1.20	9.53±0.01	3.54±0.02	2.96±0.02	0.69±0.01
DE	!75.85±0.48	!34.45±0.47	!19.67±0.07	!9.14±0.06	!17.90±0.14	0.70±0.03
BN-DG	!78.98±0.84	!42.59±1.18	!10.64±0.07	!3.93±0.08	!33.35±1.11	!7.84±0.09
BN-HC-DT	!77.56±0.87	!37.72±1.73	!10.65±0.06	!3.93±0.01	!154.79±9.83	!21.37±0.80
BN-HC-DF	!77.09±0.68	!36.75±0.99	!10.58±0.03	3.89±0.04	!234.64±56.99	!7.80±0.11
NB	81.11±0.84	!44.24±1.53	!11.42±0.06	!4.29±0.06	!16.78±0.11	0.45±0.01
Noisy-Or	!80.11±0.84	!42.15±1.48	!11.22±0.05	$!inf$±0.00	!16.65±0.11	0.44±0.02
PRIOR	!50.48±0.01	!10.48±0.01	9.63±0.00	3.83±0.00	!17.42±1.75	0.44±0.01

Tabela 4.5: detecção de erros de classificação com redução de cardinalidade - outras medidas

O HPB e o HPB-OPT ainda são os melhores métodos, mas perderam muito de sua habilidade de explorar padrões críticos, e, a uma taxa de seleção de 1%, eles nem sequer se aproximam de seu resultado no conjunto de dados original. A razão para isso é que o AIBN une valores de atributo olhando cada atributo em separado e portanto ignorando quaisquer interações entre eles. Nesse caso, interações relevantes foram perdidas.

Redes Bayesianas com grafos de decisão em lugar de tabelas de probabilidades condicionais também perderam muito da capacidade de explorar padrões críticos, o que também fica nítido através seu desempenho à taxa de seleção de 1%.

4.2 Previsão de comportamento conjunto

Na Seção 4.1 testamos o HPB em um problema onde existem apenas quatro atributos explanatórios todos afetando diretamente o atributo classe. Na maior parte dos problemas de classificação existem muito mais atributos e o HPB não pode ser aplicado diretamente. Podemos, no entanto, modelar o problema usando uma rede Bayesiana e procurar por

um ou mais nós onde o HPB possa substituir uma CPT tradicional com vantagem. Nesta seção damos um exemplo de tal nó.

Em alguns problemas de interesse da RFB que atualmente estão sendo modelados como redes Bayesianas existem nós cuja idéia central é responder a seguinte questão: o que dois ou mais atores tendem a fazer quando agem em conjunto? Em geral, esses nós determinam uma distribuição de probabilidade prévia para o que os atores fazem. Enquanto isso, outros nós avaliam a verossimilhança de quaisquer variáveis observáveis dada uma hipótese a respeito do que eles de fato fizeram.

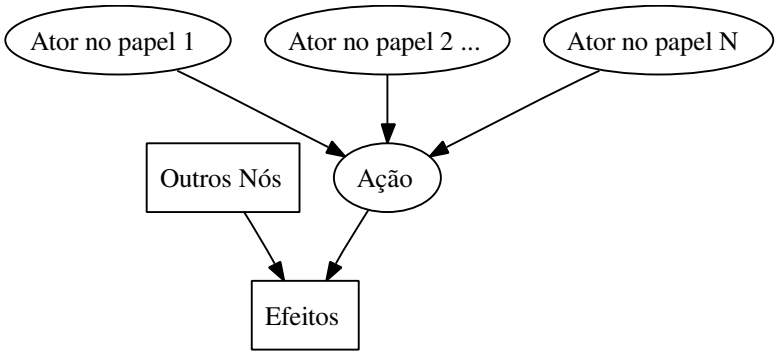

Figura 4.3: rede Bayesiana de atores

Como o número de atores possíveis pode ser muito grande, mas o número de papeis é normalmente pequeno, parece razoável substituir a CPT do nó *Ação* na Figura 4.3 pelo HPB. Contudo, na Seção 4.1, o HPB foi usado para prever uma classe binária. O número de ações possíveis pode ser alto, logo temos um desafio diferente para o HPB.

Nesta seção, apresentamos o desempenho do HPB em um problema de classificação independente que foi construído para se assemelhar ao problema de calcular a distribuição de probabilidade prévia do nó *Ação* na Figura 4.3. Usamos dois atributos de alta cardinalidade: o *importador* (IMP) e o *exportador* (EXP)[2] para prever um terceiro atributo de alta cardinalidade, a *Classificação Fiscal Declarada* (NCMD). Note que não estamos prevendo se há ou não um erro de classificação, mas a própria NCMD.

O atributo IMP pode assumir 18846 valores distintos, o atributo exportador pode assumir 43880 valores distintos e o atributo classe, NCMD, pode assumir 7608 valores distintos.

Os métodos testados foram:

- *HPB*: HPB com coeficientes S fixos;

[2]O atributo EXP não estava disponível quando fizemos os testes da Seção 4.1, por isso não o utilizamos lá.

- *NB*: *naïve Bayes*;
- *ADE*: estimativa quase direta. BN com a estrutura apresentada na Figura 1.1 e o mecanismo de suavização descrito em [34];
- *DE*: estimativa direta. BN com a estrutura apresentada na figura 1.1 e CPTs tradicionais;
- *DE Imp*: estimativa direta, ignorando o atributo EXP.
- *DE Exp*: estimativa direta, ignorando o atributo IMP.

O HPB-OPT não foi testado porque seu processo de otimização requer uma variável classe binária. Não testamos DGs, DFs e DTs porque a combinação de pais de alta cardinalidade com um nó filho também de alta cardinalidade os tornou muito lentos.

Os parâmetros de cada método foram escolhidos como na Seção 4.1, mas a MCE foi usada como critério de seleção. Abaixo apresentamos os intervalos de busca iniciais e as constantes ótimas ($s = S/NumClasses = S/7608$):

- *HPB*: o SI para a constante s foi $[1e-4, 1e-03]$ e o valor ótimo para s foi igual a $1e-3$. O SI para a constante b foi $[0.5, 2.5]$ e o valor ótimo para b foi igual a 1.0. Ao coeficiente S^{NI} sempre atribuímos um valor igual a S;
- *NB*: $SI = [1e-3, 2.5]$, $s = 0.05$;
- *ADE*: $SI = [1e-3, 2.5]$, $s = 1e-3$;
- *DE*: $SI = [1e-3, 2.5]$, $s = 1e-3$;
- *DE Imp*: $SI = [1e-3, 2.5]$, $s = 1e-3$;
- *DE Exp*: $SI = [1e-3, 2.5]$, $s = 2.5e-3$;

A Tabela 4.6 mostra que o HPB é o melhor método no que diz respeito a RMSE, MCE e a número de classificações corretas (NC). Testes de hipótese mostram que a diferença é significativa com a exceção de que o HPB não foi significativamente melhor que o NB no que diz respeito ao número de classificações corretas.

	RMSE	MCE	NC	TR	TS
HPB	10.83±0.00	8.31±0.00	26882.40±89.76	35.48±0.17	1800.73±2.99
DE	!10.88±0.00	!9.37±0.01	!25796.20±63.73	1.87±0.05	39.82±0.03
ADE	!10.88±0.00	!8.78±0.01	!26039.20±58.11	2.13±0.01	46.17±0.06
DE Exp	!10.89±0.00	!9.07±0.01	!25257.60±64.14	2.95±0.15	77.03±7.94
DE Imp	!11.04±0.00	!8.95±0.01	!22077.60±90.97	3.24±0.20	73.50±0.51
NB	!11.18±0.00	!9.23±0.01	26803.00±118.12	4.01±0.19	357.24±1.18

Tabela 4.6: previsão de comportamento conjunto

4.3 O HPB como um substituto geral para tabelas de probabilidades condicionais

Nesta seção testamos o HPB sobre conjuntos de dados da UCI. Nosso objetivo é observar seu desempenho em domínios cujas características divergem das que o inspiraram. Porque seu uso de recursos é exponencial no número de atributos, não podemos aplicar o HPB a conjuntos dados da UCI diretamente. Portanto, avaliamos o desempenho de redes Bayesianas, onde as CPTs usuais foram substituídas por instâncias do HPB. Para comparação, também avaliamos o desempenho de redes Bayesianas onde a distribuição de probabilidade condicional de um nó dados os seus pais (CPD) foi representada por outros modelos. Abaixo listamos todas as representações testadas para as CPDs:

- *HPB*: HPB como descrito na Seção 3.4;
- *DE*: estimativa direta, i.e., CPTs tradicionais;
- *ADE*: estimativa quase direta. Também CPTs, porém usando o mecanismo de suavização apresentado em [34];
- *DG*: grafos de decisão como apresentados em [16];
- *DT*: árvores de decisão como apresentadas em [34];
- *DF*: tabelas default com apresentadas em [34].

Em todos os casos aprendemos estrutura global da BN usando o algoritmo de escalada de montanha implementado na Weka 3.4.2 e o NB como ponto de partida. Para garantir que não teríamos tempos de execução excessivamente longos, limitamos o máximo número de pais a 10 e porque o HPB não lida com atributos contínuos nós os removemos. Também removemos todas as instâncias com valores de atributos faltantes.

Dependendo da representação escolhida para as CPDs, empregamos diferentes critérios de seleção na busca da estrutura da rede Bayesiana. Abaixo listamos nossas escolhas:

- HPB: verossimilhança logarítmica avaliada usando validação cruzada do tipo *leave one out*;
- DE: MDL;
- ADE: MDL;
- DGs: *Bayesian Dirichlet score* como apresentado em [16];
- DTs: MDL como apresentado em [37];

- DFs: MDL como apresentado em [37].

Os conjuntos de dados testados foram: *anneal*, *audiology*, *autos*, *breast-cancer*, *horse-colic*, *credit-rating*, *german-credit*, *cleveland-14-heart-disease*, *hungarian-14-heart-disease*, *hepatitis*, *hypothyroid*, *kr-vs-kp*, *labor*, *lymphography*, *mushroom*, *primary-tumor*, *sick*, *soybean*, *vote* e *zoo*.

Antes de construir a rede Bayesiana escolhemos um tamanho de amostra equivalente para as distribuições de probabilidade prévia (o que significa um S fixo) e o usamos para todas as instâncias do HPB dentro da rede. Felizmente, os valores ótimos para os tamanhos de amostra equivalente tendem a ser similares.

Escolhemos S começando com $SI = [1.0, 25.0]$ e, forçamos S^{NI} a ser sempre igual a S. A constante b foi escolhida começando com $SI = [0.5, 2.5]$.

Escolhemos a constante s ($s = S/NumClasses$) para grafos de decisão começando com $SI = [0.01, 2.5]$. Sempre mantivemos divisões completas ativadas e variamos exaustivamente o estado de ativação de divisões binárias e junções. Escolhemos as constantes s para outros métodos começando com $SI = [0.01, 2.5]$.

Contrariamente ao que fizemos nas Seções 4.1 e 4.2 não expandimos os intervalos de busca iniciais caso o valor ótimo de uma constante seja encontrado em um de seus pontos extremos.

Comparamos os resultados usando três critérios diferentes: número de classificações corretas (NC), entropia cruzada média (MCE) e raiz quadrada do erro quadrático médio (RMSE). Na Tabela 4.7, mostramos o números de conjuntos de dados sobre os quais cada método apresentou o melhor desempenho. Como selecionar a melhor parametrização usando um determinado critério e comparar os métodos usando apenas o mesmo critério poderia fornecer pouca informação ao leitor, nós selecionamos as parametrizações usando todos os três critérios e comparamos os métodos também usando todos os três critérios em combinações exaustivas. Em alguns casos houve empate quanto ao número de classificações corretas. Neste situação, a MCE foi usada para definir o vencedor.

Crit.Sel.	Crit.Comp.	HPB	DG	DF	DT	ADE	DE
NC	NC	9	6	2	1	1	1
NC	MCE	9	5	2	1	2	1
NC	RMSE	7	6	4	1	1	1
MCE	NC	9	5	2	0	3	1
MCE	MCE	10	5	2	1	0	2
MCE	RMSE	8	6	4	0	1	1
RMSE	NC	7	7	4	0	1	1
RMSE	MCE	9	5	2	1	1	2
RMSE	RMSE	8	6	4	0	1	1

Tabela 4.7: número de resultados vencedores em conjuntos da UCI

HPB	BDG	DF	DT	ADE	DE
3.18	3.79	1.49	1.56	1.0	1.0

Tabela 4.8: proporções entre o número de arcos em estruturas de rede

Na Tabela 4.8 mostramos as proporções médias entre o número de arcos nas estruturas aprendidas usando cada representação para as CPDs e as estruturas aprendidas usando a Estimativa Direta (CPTs tradicionais). Como previsto por Friedman e Goldszmidt [37], o uso de estruturas como DFs, DTs e DGs leva a estruturas de rede com um número maior de arcos. O uso do HPB também tem esse efeito.

A vasta maioria das variáveis nos conjuntos testados tem baixa cardinalidade (a cardinalidade mais alta entre todas as variáveis em todos os conjuntos foi a da variável classe do conjunto *audiology* com 24 valores possíveis) e muitas delas são binárias. Apesar disso, os resultados do HPB são os melhores na Tabela 4.7, mostrando que boas distribuições de probabilidade prévia, podem, em muitos casos melhorar a qualidade das estimativas.

Com mostrado na Seção 4.1, o HPB é muito mais rápido que DGs, DTs e DFs na tarefa de tratar um pequeno conjunto de atributos de alta cardinalidade. Contudo, nos conjuntos da UCI, muitos conjuntos de nós pais de baixa cardinalidade tornam o HPB comparativamente lento. Em todos os casos, o HPB foi o método mais lento e em alguns consumiu mais de 10 vezes o tempo consumido pelo segundo método mais lento.

Além disso, a vantagem do HPB nos três critérios raramente foi estatisticamente significante. Portanto, não podemos recomendar o HPB como um substituto geral para CPTs.

Podemos, no entanto, afirmar que BNs onde o HPB representa a CPD tem uma probabilidade relativamente alta de produzir resultados melhores que BNs empregando outras representações. A única explicação que encontramos para isso é que o HPB representa algumas CPDs melhor que as alternativas e que estas representações superiores resultam em previsões de classe melhores.

Isso sugere que não deve ser difícil encontrar problemas onde, se uma BN for empregada, haja um ou mais nós onde seja vantajoso o uso do HPB.

Capítulo 5

Conclusão

No domínio da pré-seleção de mercadorias em processo de importação para verificação humana, as interações entre atributos têm forte influência sobre a probabilidade de que uma dada instância seja positiva. Devido a alta cardinalidade dos atributos nesse domínio, explorar tais interações é um desafio. Tratamos o problema usando o HPB, um novo método de classificação de padrões, baseado em modelo Bayesiano empírico hierárquico de múltiplos níveis.

Apresentamos o HPB em duas versões. A primeira envolve um processo de otimização para escolha dos melhores coeficientes de suavização para cada família de padrões, enquanto a segunda e mais simples emprega um coeficiente de suavização fixo.

Avaliamos o HPB usando curvas de acerto, RMSE e MCE. Mesmo a versão mais simples do HPB se mostrou capaz de capturar a influência das interações entre atributos de alta cardinalidade e obteve ganhos de desempenho com relação a:

- redes Bayesianas com estruturas tradicionais como o *naïve Bayes* e o *tree augmented naïve Bayes*;
- redes Bayesianas onde tabelas de probabilidades condicionais tradicionais foram substituídas pelo *noisy-OR*, por tabelas default, por árvores de decisão e por grafos de decisão;
- redes Bayesianas construídas após uma fase de pré-processamento para redução de cardinalidade usando o *agglomerative information bottleneck*.

O tempo de execução do HPB é exponencial no número de atributos, mas independe de sua cardinalidade. Assim, em domínios onde os atributos são poucos, mas possuem alta cardinalidade, ele é muito mais rápido que BNs onde tabelas default, árvores de decisão ou grafos de decisão são empregados para representar a distribuição de probabilidade condicional de um nó dados os seus pais.

Mostramos que além de funcionar como um classificador independente, o HPB pode representar a distribuição de probabilidade condicional de um nó dados os seus pais e, portanto, substituir todas ou algumas tabelas de probabilidades condicionais em uma rede Bayesiana. Esse uso alternativo elimina a limitação no número de atributos que decorre da complexidade computacional do HPB.

Testamos o HPB em outro problema de classificação: a previsão do comportamento de dois atores quando eles agem juntos. Como um subproblema, esta predição é relevante em vários domínios de detecção de fraude e, se o problema geral for modelado como uma BN, geralmente aparece na forma da tarefa de representar a CPD de um determinado nó dados os seus pais. Os resultados, novamente, favoreceram o HPB.

Também exibimos resultados experimentais em conjuntos de dados da UCI, onde classificadores construídos usando redes Bayesianas com diferentes representações para as CPDs são comparados. A despeito de que esses conjuntos de dados não possuem atributos de alta cardinalidade o HPB apresentou mais resultados vencedores que qualquer outra representação em três critérios de comparação. Os tempos de execução comparativamente altos e o fato de que a maior parte das diferenças nos critérios de comparação não foi significativa, não nos permite propor o HPB como um substituto geral para CPTs. No entanto, ainda podemos concluir que nós de BNs cujas CPDs dados os seus pais são melhor representados pelo HPB que por outros métodos não são raros. Esse fato indica que o HPB pode ter uma aplicação bastante ampla.

Modelos Bayesianos hierárquicos têm sido amplamente empregados na comunidade de marketing com o nome de *hierarchical Bayes* [3, 53]. Esses modelos também têm sido usados em domínios médicos [4] e em robótica [74]. Contudo, desconhecemos que qualquer desses modelos possa ser empregado no tratamento de atributos de alta cardinalidade com interações relevantes em um problema de classificação de padrões. Além disso, o HPB difere de outros modelos por lidar com uma hierarquia de múltiplos níveis recursivamente e também por lidar com o fato de que a população de instâncias associada a cada padrão está contida em várias e não em apenas uma superpopulação.

Como trabalho futuro deixamos o desenvolvimento de mecanismos melhores para a escolha de coeficientes para o HPB. Tanto processos de otimização quanto estratégias heurísticas devem ser considerados, os primeiros para obtenção de previsões de melhor qualidade e os últimos para previsões rápidas e aceitáveis.

A hierarquia de padrões empregada pelo HPB é fixa, simétrica (todos os atributos são tratados da mesma forma), e completa (todos os subconjuntos de cada padrão de interesse são considerados no cálculo da probabilidade da classe dado o padrão). É possível que exista uma hierarquia incompleta e possivelmente assimétrica que produza resultados melhores. Desenvolver um algoritmo que procure por tal estrutura é também deixado como trabalho futuro.

Por fim, também deixamos para o futuro a exploração de uma hierarquia expandida que inclua a hierarquia criada pelo HPB e as hierarquias naturais dos atributos, como por exemplo o agrupamento de países por continente, e de importadores por atividade econômica.

Referências Bibliográficas

[1] Agosta, John Mark: *"Conditional inter-causally independent" node distributions, a property of "Noisy-OR" models.* In *Proceedings of the seventh conference (1991) on Uncertainty in Artificial Intelligence*, páginas 9–16, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc., ISBN 1-55860-203-8.

[2] Akaike, Hirotsugu: *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[3] Allenby, Greg M., Robert P. Leone, e Lichung Jen: *A dynamic model of purchase timing with application to direct marketing.* Journal of the American Statistical Association, 94(446):365–374, 1999.

[4] Andreassen, Steen, Brian Kristensen, Alina Zalounina, Leonard Leibovici, Uwe Frank, e Henrik C. Schonheyder: *Hierarchical Dirichlet Learning - Filling in the Thin Spots in a Database.* In Dojat, Michel, Elpida T. Keravnou, e Pedro Barahona (editores): *Proceedings of the 9th Conference on Artificial Intelligence in Medicine (AIME)*, volume 2780 de *Lecture Notes in Computer Science*, páginas 204–283. Springer, 2003.

[5] Asuncion, A. e D.J. Newman: *UCI Machine Learning Repository*, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[6] Bennett, Paul N.: *Assessing the calibration of Naive Bayes' posterior estimates.* Relatório Técnico CMU-CS-00-155, School of Computer Science, Carnegie Mellon University, 2000.

[7] Bouckaert, R. R.: *Properties of Bayesian network learning algorithms.* In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, páginas 102–109. Morgan Kaufmann, 1994.

[8] Bouckaert, Remco. R.: *Bayesian Belief Networks: from Construction to Inference.* Tese de Doutoramento, University of Utrecht, 1995.

[9] Boullé, Marc: *A Bayes Optimal Approach for Partitioning the Values of Categorical Attributes*. Journal of Machine Learning Research, 6:1431–1452, 2005.

[10] Boutilier, Craig, Nir Friedman, Moises Goldszmidt, e Daphne Koller: *Context-Specific Independence in Bayesian Networks*. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Inteligence (UAI)*, páginas 115 – 123, San Francisco, CA, 1996. Morgan Kaufmann Publishers.

[11] Buntine, Wray: *Theory refinement on Bayesian networks*. In *Proceedings of the seventh conference (1991) on Uncertainty in Artificial Intelligence*, páginas 52–60, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc., ISBN 1-55860-203-8.

[12] Carlin, Bradley. P. e Thomas A. Louis: *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, 2. edição, 2000.

[13] Cestnik, B.: *Estimating probabilities: a crucial task in machine learning*. In *Proceedings of the European Conference on Artificial Intelligence*, páginas 147–149, 1990.

[14] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, e W. Philip Kegelmeyer: *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence and Research, 16:321–357, 2002.

[15] Cheng, J., D. Bell, e W. Liu: *An algorithm for Bayesian belief network construction from data*. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics (AI & STAT)*, páginas 83–90, 1997.

[16] Chickering, David Maxwell, David Heckerman, e Christopher Meek: *A Bayesian Approach to Learning Bayesian Networks with Local Structure*. Relatório Técnico MSR-TR-97-07, Microsoft Research, Redmond, WA 98052, 1997.

[17] Chow, C. e C. Liu: *Approximating discrete probability distributions with dependence trees*. Information Theory, IEEE Transactions on, 14(3):462–467, 1968.

[18] Cooper, Greg F. e Ed Herskovits: *A Bayesian Method for the Induction of Probabilistic Networks from Data*. Machine Learning, 9:309–347, 1992.

[19] Cozman, Fabio Gagliardi: *Axiomatizing Noisy-OR*. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence (ECAI)*, páginas 979–980. IOS Press, 2004, ISBN 1-58603-452-9.

[20] Cozman, Fabio Gagliardi: *Axiomatizing Noisy-OR*. Relatório Técnico BT/PMR/0409, Escola Politécnica da Universidade de São Paulo (USP), 2004.

[21] Cussens, James: *Bayes and Pseudo-Bayes Estimates of Conditional Probabilities and Their Reliability*. In *Proceedings of the European Conference on Machine Learning (ECML)*, páginas 136–152, London, UK, 1993. Springer-Verlag, ISBN 3-540-56602-3.

[22] Demšar, Janez: *Statistical Comparisons of Classifiers over Multiple Data Sets*. Journal of Machine Learning Research, 7:1–30, 2006, ISSN 1533-7928.

[23] Díez, Francisco Javier: *Parameter Adjustment in Bayes Networks. The generalized Noisy OR-gate*. In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, páginas 99–105, San Francisco, CA, 1993. Morgan Kaufmann Publishers.

[24] Dietterich, Thomas G.: *Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms*. Neural Computation, 10(7):1895–1923, 1998.

[25] Domingos, Pedro: *Bayesian Averaging of Classifiers and the Overfitting Problem*. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, páginas 223–230, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc., ISBN 1-55860-707-2.

[26] Domingos, Pedro e Michael J. Pazzani: *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss*. Machine Learning, 29(2-3):103–130, 1997.

[27] Duda, Richard O. e Peter E. Hart: *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[28] Duda, Richard O., Peter E. Hart, e David G. Stork: *Pattern Classification*. John Wiley and Sons Inc., New York, 2. edição, 2003.

[29] Egan, James P.: *Signal Detection Theory and Roc Analysis*. Academic Press, New York, 1975.

[30] Evans, Merran, Nicholas Hastings, e Brian Peacock: *Statistical Distributions*. John Wiley & Sons, 3. edição, 2000.

[31] Fenoeux, T. e M.S. Bjanger: *Induction of decision trees from partially classified data using belief functions*. In *Proceedings of the Systems, Man, and Cybernetics, IEEE International Conference on*, volume 4, páginas 2923–2928, 2000.

[32] Ferreira, Marcos Antônio Cardoso: *Uso de Redes de Crença para seleção de Declarações de Importação*. Tese de Mestrado, Instituto Tecnológico de Aeronáutica, 2003.

[33] Fisher, R. A: *The goodness of fit of regression formulae, and the distribution of regression coefficients*. Journal of the Royal Statistical Society, 85:597–612, 1922.

[34] Friedman, Nir, Dan Geiger, e Moises Goldszmidt: *Bayesian Network Classifiers*. Machine Learning, 29(2-3):131–163, 1997.

[35] Friedman, Nir e Moises Goldszmidt: *Building Classifiers Using Bayesian Networks*. In *Proceedings of the American Association for Artificial Intelligence (AAAI)/Innovative Applications of Artificial Intelligence (IAAI)*, volume 2, páginas 1277–1284, 1996.

[36] Friedman, Nir e Moises Goldszmidt: *Discretization of continuous attributes while learning Bayesian networks*. In *Proceedings of the Internatinal Conference of Machine learning (ICML)*. Morgan Kaufmann, July 1996.

[37] Friedman, Nir e Moises Goldszmidt: *Learning Bayesian Networks with Local Structure*. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Inteligence (UAI)*, páginas 252–262, San Francisco, CA, 1996. Morgan Kaufmann Publishers.

[38] Gelman, Andrew B., John S. Carlin, Hal S. Stern, e Donald B. Rubin: *Bayesian Data Analysis*. Chapman and Hall, 2. edição, 2003.

[39] Good, I.: *A causal calculus (i)*. British Journal for the Philosophy of Science II, páginas 305–318, 1961.

[40] Hamine, V. e P. Helman: *Learning Optimal Augmented Bayes Networks*. Relatório Técnico TR-CS-2004-11, Computer Science Department, University of New Mexico, 2004.

[41] Hamine, V. e P. Helman: *A Theoretical and Experimental Evaluation of Augmented Bayesian Classifiers*. Relatório Técnico TR-CS-2006-03, Computer Science Department, University of New Mexico, 2006.

[42] Heckerman, David: *A tutorial on learning with Bayesian networks*. In Jordan, Michael (editor): *Learning in graphical models*, páginas 301–354. MIT Press, Cambridge, MA, USA, 1999, ISBN 0-262-60032-3.

[43] Heckerman, D., D. Geiger, e D. M. Chickering: *Learning Bayesian networks: The combination of knowledge and statistical data*. Machine Learning, 20:197–243, 1995.

[44] Jaeger, Manfred: *Probabilistic decision graphs-combining verification and AI techniques for probabilistic inference*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 12(SUPPLEMENT):19–42, 2004.

[45] Jambeiro Filho, Jorge e Jacques Wainer: *Analyzing Bayesian Networks with Local Structure and Cardinality Reduction over a Practical Case*. In *Proceedings of the Workshop on Computational Intelligence (WCI)*, 2006.

[46] Jambeiro Filho, Jorge e Jacques Wainer: *Using a Hierarchical Bayesian Model to Handle High Cardinality Attributes with Relevant Interactions in a Classification Problem*. In *Proceedings of the International Joint Conference of Artificial Intelligence (IJCAI)*. AAAI Press, 2007.

[47] Keogh, Eamonn J. e Michael J. Pazzani: *Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches*. In *Proceeeding of the Seventh International Workshop on Artificial Intelligence and Statistics*, páginas 225–230, Ft. Lauderdale, FL, 1999.

[48] Kókai, Gabriella: *Development of methods how to avoid the overfitting-effect within the GeLog-system*. In Palade, Vasile, Robert J. Howlett, e Lakhmi C. Jain (editores): *Proc. Seventh International Conference on Knowledge-Based Intelligent Information Engineering Systems*, Lecture Notes in Computer Science, páginas 958–966, Heidelberg, 2003. Springer Verlag.

[49] Lam, W. e F. Bacchus: *Learning Bayesian Belief Networks: An Approach Based on the MDL Principle*. Computational Intelligence, 10:269–293, 1994.

[50] Lanna, Antonella Saraiva: *Brazilian Customs Risk Management Model*. In *Proceedings of Best Practices Exchange Program for Customs Administrations of the Caribbean, Latin America, and East Asian countries*, Japan, 2001.

[51] Lanterman, Aaron D.: *Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Order Estimation*. International Statistical Review, 69(2):185–212, 8 2001.

[52] Laplace, Pierre Simon: *Théorie analytique des probabilités*. Paris: Veuve Courcier, 1812.

[53] Lenk, P., W. DeSarbo, P. Green, e M. Young: *Hierarchical Bayes conjoint analysis: recovery of part worth heterogeneity from reduced experimental designs*. Marketing Science, 15:173–191, 1996.

[54] LidStone, G. J.: *Note on the Bayes-Laplace formula for inductive or a posteriori probabilities*. Transactions of the Faculty of Actuaries, 8:182 – 192, 1920.

[55] Liere, Ray: *Active Learning with Committees: An Approach to Efficient Learning in Text Categorization Using Linear Threshold Algorithms*. Tese de Doutoramento, Computer Science Department of Oregon State University, 1999.

[56] Ling, Charles X. e Huajie Zhang: *The representational power of discrete Bayesian networks*. Journal of Machine Learning Research, 3:709–721, 2002.

[57] Little, Roderick J. A. e Donald B. Rubin: *Statistical Analysis with Missing Data*. Wiley-Interscience, 2. edição, 2002.

[58] Micci-Barreca, Daniele: *A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems*. SIGKDD Explor. Newsl., 3(1):27–32, 2001.

[59] Mitchell, Tom M.: *Machine Learning*. McGraw-Hill, New York, 1997.

[60] Murphy, Allan H. e Robert L. Winkler: *Reliability of Subjective Probability Forecasts of Precipitation and Temperature*. Applied Statistics, 26(1):41–47, 1977.

[61] Neapolitan, Richard E.: *Learning Bayesian Networks*. Prentice Hall, 1. edição, 2003.

[62] Paris, Gregory, Denis Robilliard, e Cyril Fonlupt: *Exploring Overfitting in Genetic Programming*. In Liardet, Pierre, Pierre Collet, Cyril Fonlupt, Evelyne Lutton, e Marc Schoenauer (editores): *Evolution Artificielle, 6th International Conference*, volume 2936 de *Lecture Notes in Computer Science*, páginas 267–277, Marseilles, France, 27-30 outubro 2003. Springer, ISBN 3-540-21523-9. Revised Selected Papers.

[63] Pearl, Judea: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988, ISBN 1558604790.

[64] Pearl, Judea: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[65] Platt, J.: *Probabilistic outputs for support vector machines and comparison to regularized likelihood methods*. In Smola, A.J., P. Bartlett, B. Schoelkopf, e D. Schuurmans (editores): *Proceedings of Advances in Large Margin Classiers*, páginas 61–74. MIT Press, 1999.

[66] Rish, Irina, Joseph Hellerstein, e Jayram Thathachar: *An analysis of data characteristics that affect Naive Bayes performance*. Relatório Técnico RC21993, Watson Research Center, 2001.

[67] Rissanen, J.: *Modeling by shortest data description*. Automatica, 14:465–471, 1978.

[68] Sacha, Jaroslaw P.: *New Synthesis of Bayesian Network Classifiers and Cardiac SPECT Image Interpretation*. Tese de Doutoramento, University of Toledo, 1999.

[69] Sampath, Srinivas: *A Generalization of the Noisy-OR Model*. In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, páginas 208–215, San Francisco, CA, 1993. Morgan Kaufmann Publishers.

[70] Schwarz, G.: *Estimating the dimension of a model*. Annals of Statistics, 6:461–464, 1978.

[71] Sharma, Rita e David Poole: *Probabilistic Reasoning with Hierarchically Structured Variables*. In Kaelbling, Leslie Pack e Alessandro Saffiotti (editores): *IJCAI*, páginas 1391–1397. Professional Book Center, 2005, ISBN 0938075934.

[72] Slonim, Noam e Naftali Tishby: *Agglomerative Information Bottleneck.* In *Advances in Neural Information Processing Systems 12 (NIPS)*, páginas 617–623, Denver, Colorado, USA, 1999. The MIT Press, ISBN 0-262-19450-3.

[73] Spirtes, P., C. Glymour, e R. Scheines: *An algorithm for fast recovery of sparse causal graphs*. Social Science Computer Review, 9:62–72, 1991.

[74] Stewart, Benjamin, Jonathan Ko, Dieter Fox, e Kurt Konolige: *The Revisiting Problem in Mobile Robot Map Building: A Hierarchical Bayesian Approach.* In Meek, Christopher e Uffe Kjærulff (editores): *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence (UAI)*, páginas 551–558, Acapulco, Mexico, 2003. Morgan Kaufmann, ISBN 0-127-05664-5.

[75] Twardy, Charles e Kevin Korb: *Causal Interaction*. Relatório Técnico CSSE 118, Monash University, 2002.

[76] Wikipedia: *Overfitting*, 2007. http://en.wikipedia.org/wiki/Overfitting, [acessada em 26-July-2007].

[77] Witten, Ian H. e Eibe Frank: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers Inc., 1999.

[78] Zadrozny, Bianca e Charles Elkan: *Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers*. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, páginas 609–616, MA, USA, 2001. Morgan Kaufmann, ISBN 1-55860-778-1.

[79] Zadrozny, Bianca e Charles Elkan: *Transforming Classifier Scores into Accurate Multiclass Probability Estimates*. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 694–699. ACM Press, 2002.

[80] Zhang, Harry e Jiang Su: *Naive Bayesian Classifiers for Ranking*. Lecture Notes in Computer Science, 3201:501–512, 2004.

[81] Zhang, N. L. e D. Poole: *Exploiting causal independence in Bayesian network inference*. Journal of Artificial Inteligence Research, 5:301–328, 1996.

[82] Zhu, Mu: *Recall, Precision and Average Precision*. Relatório Técnico 09, Department of Statistics & Actuarial Science, University of Waterloo, 2004.

Apêndice A

Tabela de siglas

Sigla	Seção	Significado	Comentário
ADE	2.6.1	estimativa quase direta	É uma versão suavizada da DE. A sigla vem do inglês: *almost direct estimation*.
AIBN	4.1	*agglomerative information bottleneck*	É um método de agrupamento de valores para redução da cardinalidade de atributos.
AIC	2.6.2	*Akaike's Information Criterion*	É um critério para seleção de estruturas de BNs ou outros tipos de modelo.
B	2.6.2	B	É um algoritmo de busca para estruturas de BNs .
BDES	2.6.2	*Bayesian Dirichlet likelihood equivalent score*	É um caso particular de BDS.
BDEUS	2.6.2	*Bayesian Dirichlet likelihood equivalent uniform score*	É um caso particular de BDES.
BDS	2.6.2	*Bayesian Dirichlet score*	É um caso particular de BS.
BIC	2.6.2	*Bayesian information criterion*	É um critério para seleção de estruturas de BNs ou outros tipos de modelo.
BN	2.6	rede Bayesiana	É uma representação para a distribuição de probabilidade conjunta de um grupo de variáveis e pode ser usada como método de classificação
BS	2.6.2	*Bayesian score*	É um critério para seleção de estruturas de BNs ou outros tipos de modelo.
CPD	2.6	distribuição de probabilidade condicional	A sigla vem do inglês: *conditional probability distribution*.
CPT	2.6.1	tabela de probabilidades condicionais	A sigla vem do inglês: *conditional probability table*.
DE	2.6.1	estimativa direta	É um método popular para estimar probabilidades a partir de proporções. A sigla vem do inglês: *direct estimation*.
DF	2.6.1	tabela default	É uma alternativa a CPT como forma de representação de uma CPD. A sigla vem do inglês: *default table*.
DG	2.6.1	grafo de decisão	É uma alternativa a CPT como forma de representação de uma CPD ou um método de classificação independente. A sigla vem do inglês: *decision graph*.
DT	2.6.1	árvore de decisão	É uma alternativa a CPT como forma de representação de uma CPD ou um método de classificação independente. A sigla vem do inglês: *decision tree*.
FAN	2.6.2	*forest augmented naïve Bayes*	É uma generalização do TAN.
HPB	3	*hierarchical pattern Bayes*	É novo método de classificação independente ou uma alternativa a CPT como forma de representação de uma CPD.
IMP	1	Importador	
K2	2.6.2	K2	É um caso particular de BDS e também um algoritmo de busca para estruturas de BNs .
LLLOO	3.4	verossimilhança logarítmica medida em LOO	É um critério para seleção de estruturas de BNs ou outros tipos de modelo e é equivalente ao LLOO. A sigla vem do inglês: *log likelihood evaluated under leave one out cross-validation*
LLOO	2.6.2	verossimilhança medida em LOO	É um critério para seleção de estruturas de BNs ou outros tipos de modelo. A sigla vem do inglês: *likelihood evaluated under leave one out cross-validation*.
LOO	2.6.2	validação cruzada do tipo *leave one out*	É um método para estimar os valores de vários indicadores para dados novos a partir de medições sobre exemplos conhecidos.
MCE	4.1	entropia cruzada média	É uma medida de divergência para estimativas de probabilidade. A sigla vem do inglês: *mean cross entropy*.

MDL	2.6.2	mínimo comprimento de descrição	É um critério para seleção de estruturas de BNs ou outros tipos de modelo. A sigla vem do inglês: *minimum description length*.
NB	2.4	*naïve Bayes*	É um método de classificação que é um caso particular de BN
NCM	1	nomenclatura comum do mercosul	É uma Tabela de quase 10000 posições na qual todas as mercadorias importadas têm que ser enquadradas.
NCMD	1	classificação fiscal declarada na NCM	
PAIS	1	país	É o país origem da mercadoria.
RFB	1	Receita Federal do Brasil	
RMSE	4.1	raiz quadrada do erro quadrático médio	É uma medida de divergência para estimativas de probabilidade. A sigla vem do inglês: *root mean squared error*.
SFAN	2.6.2	*selected forest augmented naïve Bayes*	É uma generalização do FAN.
TAN	2.6.2	*tree augmented naïve Bayes*	É um método de classificação que é um caso particular de BN.
URF	1	unidade da RFB	É a unidade da RFB pela qual a mercadoria ingressou na Brasil

Tabela A.1: siglas utilizadas nesta tese