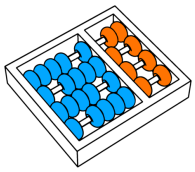
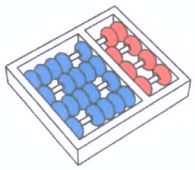


Priscila Tiemi Maeda Saito

"Active learning with applications to the diagnosis of parasites"

"*Aprendizado ativo com aplicações ao diagnóstico de parasitos*"

CAMPINAS
2014



University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

Priscila Tiemi Maeda Saito

"Active learning with applications to the diagnosis of parasites"

Supervisor: **Prof. Dr. Alexandre Xavier Falcão**
Orientador(a):

Co-Supervisor: **Prof. Dr. Pedro Jussieu de Rezende**
Co-orientador(a):

"Aprendizado ativo com aplicações ao diagnóstico de parasitos"

PhD Thesis presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a PhD degree in Computer Science.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Ciência da Computação.

THIS VOLUME CORRESPONDS TO THE FINAL VERSION OF THE THESIS DEFENDED BY PRISCILA TIEMI MAEDA SAITO, UNDER THE SUPERVISION OF PROF. DR. ALEXANDRE XAVIER FALCÃO.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA POR PRISCILA TIEMI MAEDA SAITO, SOB ORIENTAÇÃO DE PROF. DR. ALEXANDRE XAVIER FALCÃO.

Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

Sa28a

Saito, Priscila Tiemi Maeda, 1985-
Active learning with applications to the diagnosis of parasites / Priscila Tiemi Maeda Saito. – Campinas, SP : [s.n.], 2014.

Orientador: Alexandre Xavier Falcão.
Coorientador: Pedro Jussieu de Rezende.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizado de máquina. 2. Inteligência artificial. 3. Reconhecimento de padrões. 4. Análise de imagem. 5. Parasitos - Diagnóstico. I. Falcão, Alexandre Xavier,1966-. II. Rezende, Pedro Jussieu de,1955-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

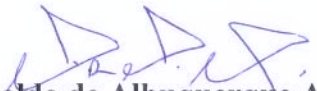
Informações para Biblioteca Digital

Título em outro idioma: Aprendizado ativo com aplicações ao diagnóstico de parasitos
Palavras-chave em inglês:
Machine learning
Artificial intelligence
Pattern recognition
Image analysis
Parasites - Diagnosis
Área de concentração: Ciência da Computação
Titulação: Doutora em Ciência da Computação
Banca examinadora:
Alexandre Xavier Falcão [Orientador]
Aparecido Nilceu Marana
Arnaldo de Albuquerque Araújo
Hélio Pedrini
Anderson de Rezende Rocha
Data de defesa: 28-04-2014
Programa de Pós-Graduação: Ciência da Computação

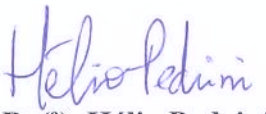
TERMO DE APROVAÇÃO

Defesa de Tese de Doutorado em Ciência da Computação, apresentada pelo(a) Doutorando(a) **Priscila Tiemi Maeda Saito**, aprovado(a) em **28 de abril de 2014**, pela Banca examinadora composta pelos professores doutores:

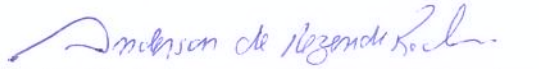
Prof(ª). Dr(ª). Aparecido Nilceu Marana
Titular



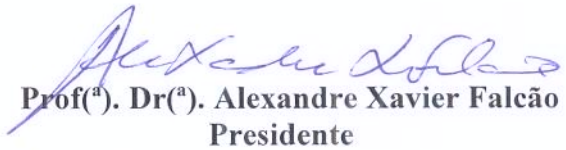
Prof(ª). Dr(ª). Arnaldo de Albuquerque Araújo
Titular



Prof(ª). Dr(ª). Hélio Pedrini
Titular



Prof(ª). Dr(ª). Anderson de Rezende Rocha
Titular



Prof(ª). Dr(ª). Alexandre Xavier Falcão
Presidente

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

Active learning with applications to the diagnosis of parasites

Priscila Tiemi Maeda Saito[1]

April 28, 2014

Examiner Board/*Banca Examinadora*:

- Prof. Dr. Alexandre Xavier Falcão (Supervisor/*Orientador*)
- Prof. Dr. Anderson de Rezende Rocha
  IC-UNICAMP
- Prof. Dr. Hélio Pedrini
  IC-UNICAMP
- Prof. Dr. Aparecido Nilceu Marana
  FC-UNESP
- Prof. Dr. Arnaldo de Albuquerque Araújo
  DCC-UFMG
- Prof. Dr. Jefersson Alex dos Santos
  DCC-UFMG (Substitute/*Suplente*)
- Prof. Dr. Léo Pini Magalhães
  FEEC-UNICAMP (Substitute/*Suplente*)
- Prof. Dr. Ricardo da Silva Torres
  IC-UNICAMP (Substitute/*Suplente*)

[1]Financial support: CNPq scholarship (process 141795/2010-7) — 2010–2012 and CAPES scholarship (process 01-P-01965/2012) — 2012

Abstract

Image datasets have grown large with the fast advances and varieties of the imaging technologies, demanding urgent solutions for information processing, organization, and retrieval. Processing here aims to annotate the image by assigning to it a label that represents its semantic content. Annotation is crucial for the effective organization and retrieval of the information related to the images. However, manual annotation is unfeasible in large datasets and successful automatic annotation by a pattern classifier strongly depends on the quality of a much smaller training set. Active learning techniques have been proposed to select those representative training samples from the large dataset with a label suggestion, which can be either confirmed or corrected by the expert. Nevertheless, these techniques very often ignore the need for interactive response times during the active learning process. Therefore, this PhD thesis presents active learning methods that can reduce and/or organize the large dataset such that sample selection does not require to reprocess it entirely at every learning iteration. Moreover, it can be interrupted as soon as a desired number of samples from the reduced and organized dataset is identified. These methods show an increasing progress, first with data reduction only, and then with subsequent organization of the reduced dataset. However, the thesis also addresses a real problem — the diagnosis of parasites — in which the existence of a diverse class (i.e., the impurity class), with much larger size and samples that are similar to some types of parasites, makes data reduction considerably less effective. The problem is finally circumvented with a different type of data organization, which still allows interactive response times and yields a better and robust active learning approach for the diagnosis of parasites. The methods have been extensively assessed with different types of unsupervised and supervised classifiers using datasets from distinct applications and baseline approaches that rely on random sample selection and/or reprocess the entire dataset at each learning iteration. Finally, the thesis demonstrates that further improvements are obtained with semi-supervised learning.

Resumo

Conjuntos de imagens têm crescido consideravelmente com o rápido avanço de inúmeras tecnologias de imagens, demandando soluções urgentes para o processamento, organização e recuperação da informação. O processamento, neste caso, objetiva anotar uma dada imagem atribuindo-na um rótulo que representa seu conteúdo semântico. A anotação é crucial para a organizaçao e recuperação efetiva da informação relacionada às imagens. No entanto, a anotação manual é inviável em grandes conjuntos de dados. Além disso, a anotação automática bem sucedida por um classificador de padrões depende fortemente da qualidade de um conjunto de treinamento reduzido. Técnicas de aprendizado ativo têm sido propostas para selecionar, a partir de um grande conjunto, amostras de treinamento representativas, com uma sugestão de rótulo que pode ser confirmado ou corrigido pelo especialista. Apesar disso, essas técnicas muitas vezes ignoram a necessidade de tempos de resposta interativos durante o processo de aprendizado ativo. Portanto, esta tese de doutorado apresenta métodos de aprendizado ativo que podem reduzir e/ou organizar um grande conjunto de dados, tal que a fase de seleção não requer reprocessá-lo inteiramente a cada iteração do aprendizado. Além disso, tal seleção pode ser interrompida quando o número de amostras desejadas, a partir do conjunto de dados reduzido e organizado, é identificado. Os métodos propostos mostram um progresso cada vez maior, primeiro apenas com a redução de dados, e em seguida com a subsequente organização do conjunto reduzido. Esta tese também aborda um problema real — o diagnóstico de parasitos — em que a existência de uma classe diversa (isto é, uma classe de impureza), com tamanho muito maior e amostras que são similares a alguns tipos de parasitos, torna a redução de dados consideravelmente menos eficaz. Este problema é finalmente contornado com um tipo de organização de dados diferente, que ainda permite tempos de resposta interativos e produz uma abordagem de aprendizado ativo melhor e robusta para o diagnóstico de parasitos. Os métodos desenvolvidos foram extensivamente avaliados com diferentes tipos de classificadores supervisionados e não-supervisionados utilizando conjunto de dados a partir de aplicações distintas e abordagens baselines que baseiam-se em seleção aleatória de amostras e/ou reprocessamento de todo o conjunto de dados a cada iteração do aprendizado. Por fim, esta tese demonstra que outras melhorias são obtidas com o aprendizado semi-supervisionado.

Acknowledgements

First and foremost I thank God, my guardian angel and my spiritual protector for the health, wisdom and perseverance that has been bestowed upon me during this research project and indeed, for all my life. I am deeply grateful to my beloved family: parents, sister, aunts, cousins and others for their unwavering love and support throughout my life. I am eternally thankful to my dearest, encouraging, and patient, Pedro, whose immeasurable and faithful support is so appreciated. I also thank his parents and family for their warm help. I would like to express my deep gratitude to all of them mentioned above for all the inexpressible. Thank you for giving so much love, and taking such good care of me!

I would like to express my deepest gratitute to Professor Falcão and Professor Rezende for accepting to be my supervisors in this project. I thank them for their guidance, directives and precious time spent offering me a lot of their expertise and research insight. I admire their inspiring enthusiasm for science and their bright wisdom. It would be hard to overstate how much I benefited from all of this. Their constant support, thoughtful advices and invaluable suggestions made this work successful and my Ph.D. experience productive and stimulating. I cannot finish without acknowledge how eternally grateful and thankful I am to them for giving me the opportunity to enjoy all the experiences I have had during my time at UNICAMP. The oportunity of studying in the Institute really has broadened my horizon and opened new chances and research directions for me.

My great thanks also go to professors, staff, labmates and others from the Institute. I acknowledge all the collaborators that I have had the pleasure to work. I am tempted to individually thank all of them but as the list might be long and for fear I might omit someone, I will simply say: Thank you to you all for your great assistance. Very special thanks are to Celso and Jancarlo, especially for their extensive help in the project of parasites. I am extremely grateful as well, to my colleagues and friends for their inspirational and helpful discussions, studies in group which kept us together and afforded a mutual learning, as well as by the ride offered.

I would like to express my sincere gratitude to the Federal Technological University of Parana and its directors, coordinators and professors. I would like to thank all people who have helped me.

I am also very grateful to CNPq (process 141795/2010-7), CAPES (process 01-P-01965/2012), FAPESP, FAEPEX and UNICAMP for the financial support awarded.

Last but not least, I would like to whole-heartedly thank all people who contributed in some way for the development of this work. To all of you, thank you very much!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Due to the advances in data acquisition and storage technologies, datasets from multiple modalities (e.g. medical, remote sensing, multimedia applications, among others) have grown at a very fast pace and they are commonly available to support research, education, entertainment, and several other activities. Attempts to the development of effective and efficient ways of handling and analyzing real-world applications are becoming increasingly widespread, yet they still face a number of practical challenges.

Annotation is the most effective way to organize the data and retrieve the desired information. However, the labor-intensive and time-consuming process of annotating data is a serious bottleneck in many pattern recognition applications when handling large datasets. Manually annotating such increasing volume of data becomes humanly infeasible and highly susceptible to errors. Moreover, successfully automatic annotation by a pattern classifier strongly depends on the quality of a much smaller training set.

The human knowledge is indispensable for the success of the learning phase and user's time and effort are precious resources. Can the user annotate a minimum number of samples and the classifier label the remaining ones with high accuracy? Active learning techniques have been investigated to answer this question. These techniques aim to identify the most representative samples for manual annotation in a few classifier learning iterations. At each iteration, the classifier selects samples from the dataset and suggests their labels to the user, who can accept/correct labels to the next iteration.

However, these techniques very often ignore the need for interactive response times during the active learning process. They usually adopt a common strategy, which requires at each learning iteration: classification of the entire dataset, reorganization of all samples according to some (sorting) criterion, and selection of the most informative ones to train the classifier. For a large dataset, it is impractical to process it entirely at every iteration.

Therefore, despite these efforts in active learning, there are important issues that remain open. The questions are: how many samples are needed for the learning process?

what are the most representatives samples? Can they be selected in an efficient manner? Given these issues, this PhD thesis presents active learning methods which can attain a significant reduction in the size of the learning set by applying an a priori process of identification and organization of a small relevant subset. Furthermore, the concomitant classification and selection processes enable the classification of a very small number of samples, while selecting the informative ones.

This thesis also addresses a real problem related to the diagnosis of intestinal parasites. The existence of a diverse class (i.e., the impurity class), being higher in number and also similar to some species of parasites, represents the major challenge and makes data reduction considerably less effective. This problem is circumvented with a different method that previously organizes the data and then properly balances the selection of diverse and uncertain samples. The method still allows interactive response times and yields a better and robust active learning approach for the diagnosis of parasites.

The proposed methods here have been extensively evaluated using different types of unsupervised and supervised classifiers, datasets from distinct applications and baseline learning approaches. Finally, the thesis demonstrates that further improvements are obtained with semi-supervised learning strategy.

Taking into account all the proposals, an overview of the main contributions of this thesis is presented in the following section.

1.1 Overview of the Contributions

Figure 1.1 illustrates an overview of the proposed active learning framework, highlighting the elements that compose our paradigm jointly with the developed learning strategies.

The *image acquisition and feature descriptor extraction processes* were not the focus of this doctoral research. From the non-annotated dataset, we proposed the active learning paradigm in order to select, more efficiently and effectively, a small number of the most useful samples for training a classifier. *Analysis and organization strategies* developed here, as far as we know, are unique in the sense that they perform data organization only once (in a *unsupervised preprocessing phase*). Previous data organization avoids to reprocess the large dataset at each learning iteration.

In general, the organization strategy adopted here analyzes the distribution of all non-annotated samples in the feature space to organize them. This data organization is based on graph clustering followed by sorting of the samples according to some criterion. Our methods generally follow the idea of selecting cluster roots, aiming to select samples from all classes faster. Then it is able to select the most diverse (difficult) samples for classification, because, after the first iteration, the classifier also participates of the sampling choice by a *selection strategy*.

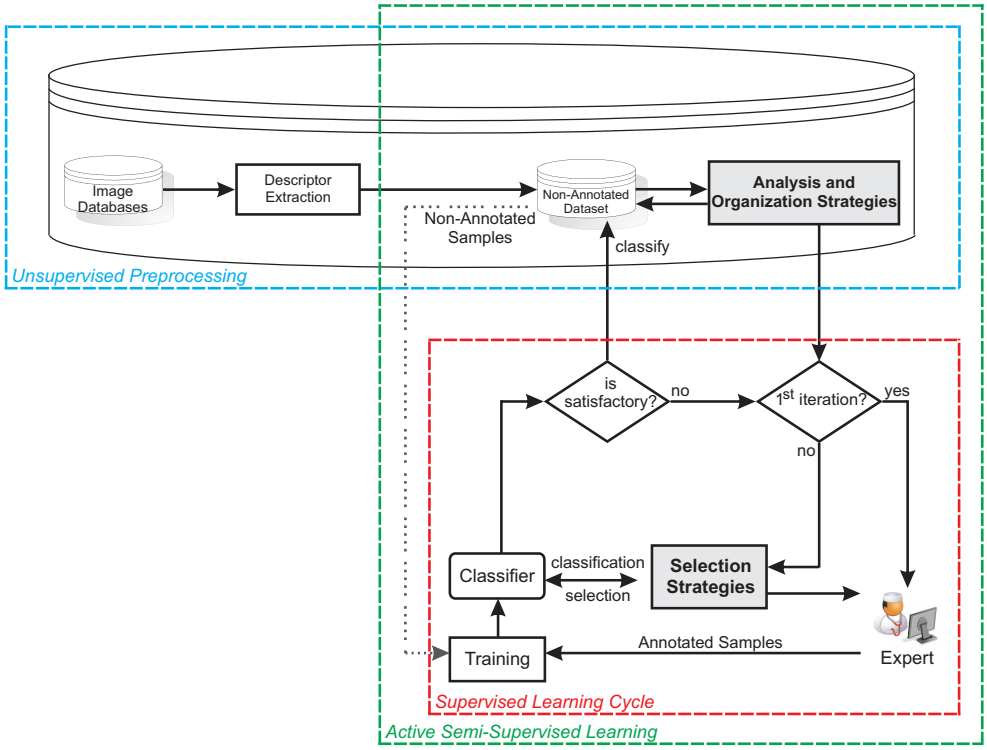


Figure 1.1: Pipeline of the proposed active learning framework.

At the first iteration, the roots of the clusters are displayed to the expert, who annotates their labels. These annotated samples constitute the training set of the first classifier instance. During the *learning process*, one sample (at a time) on the ordered set is labeled by the current classifier, and the sample is selected if it receives the label that satisfies the given selection criterion. It is important to emphasize that the classifier does not label all samples in the dataset. Both phases, *classification and selection*, are performed alternately until a desired number of samples per iteration be reached.

The actively selected samples along with expert-verified labels are then added to the training set. This process is iterative such that after each iteration of expert feedback the classifier is retrained. Since the selected samples are automatically labeled by the current classifier, the expert only has to verify the assigned classes and annotate the misclassified ones. As the classifier improves at each iteration, the number of samples incorrectly classified is considerably reduced. In this way, the expert's time and effort are also significantly reduced. Upon sensing that an acceptable accuracy has been reached[1], the expert can direct the final classifier to label the remaining of the dataset. In our

[1]The system can run cross validation using the labeled samples to indicate the learning status of the classifier at each iteration.

experiments, we considered that an expert would be satisfied whenever the mean accuracy on the training set remains stable along iterations or reaches a sufficiently high level for a given application.

In the proposed framework, non-annotated samples can also be included in the training set to design a more effective classifier by *active semi-supervised learning*.

Summing up, the main contributions of this work are:

- Proposal of a novel learning paradigm that is computationally and iteratively efficient, as it avoids to process the entire dataset at each learning iteration, affording interactive response time and verification of a considerably smaller part of the dataset, so allowing its application to large datasets;

- Development of new active learning strategies associated with the aforementioned paradigm which select the most useful (most diverse and most uncertain) samples for the learning process, quickly providing high classification accuracy, identification of samples from all classes and decrease of the human effort;

- Evaluation of the proposed active learning strategies with different clustering and classification techniques, as well as with baseline learning strategies and using datasets from different application domains, of different sizes, and with feature spaces of various dimensions and classes;

- Investigation and development of a promising active learning methodology toward the fully automation of the enteroparasitosis diagnosis via image analysis, which can significantly advance the area of clinical parasitology;

- Evaluation and validation of the developed methodology by an experienced expert in parasitology using a realistic scenario for this application, indicating that our solution is effective and suitable for laboratory routine. It is important to highlight that considering the low sensitivity rates from the traditional diagnosis procedure, based on visual analysis (48.3% up to 75.9%) [40], we may conclude that our solution is very relevant for the area of clinical parasitology;

- Specification of fast and accurate approach that combines active learning and semi-supervised learning strategies based on optimum-path forest classifiers, identifying and selecting samples from all classes more quickly while decreasing the propagated errors on the unlabeled set and keeping user interaction to a minimum.

This thesis is organized as a follows: Section 2 summarizes the main works and concepts in the field of image annotation and active learning presented in the literature.

Furthermore, we briefly discuss previous works that are required for an adequate understanding related to (supervised, unsupervised and semi-supervised) learning by optimum-path forest. Section 3 details the active learning paradigm as well as the strategies proposed. Section 4 discusses the experiments and the accomplished results. Finally, Section 5 presents the conclusions and some directions for future works.

Chapter 2

Background

This Chapter summarizes the main works and concepts in the field of image annotation and active learning, as well as briefly discuss background materials and previous works that are required for an adequate understanding.

2.1 Image Annotation and Active Learning

Image annotation is a process by which labels are associated with images, either manually, automatically or semi-automatically [91, 12, 103, 58]. The manual annotation approach presents some drawbacks such as being time consuming and laborious. Hence, the new trend towards automatic image annotation seems promising [42, 67, 65, 102, 31, 29, 54, 8].

The main idea of automatic image annotation techniques is to learn semantic concept models from labeled image samples, and use the concept models to label new images automatically. Once images are annotated with semantic labels, they can be retrieved by keyword. Assuming that low level features are extracted from image content and semantic labels are collected from image samples, conventional classification methods can be trained to map the features to the semantic labels. Once trained, the classifier can be used to annotate new image samples. Many works [102] explore the label annotation using classifiers, such as Bayesian [53, 8, 47], Support Vector Machines (SVM) [74, 39], Artificial Neural Network (ANN) [27, 73], k-Nearest Neighbor (k-NN) [90, 46], Decision Tree (DT) [55, 99, 97] and Optimum-Path Forest (OPF) [18, 17].

Despite the efforts in automatic image annotation, their success usually depend on a suitable image pre-processing and on a small training set, which is feasible for expert annotation. Such pre-processing should involve the design of discriminative features for a given problem, by exploring the prior knowledge about the problem and/or feature selection [78, 4] and deep learning techniques [28, 75, 104, 98, 2].

In conventional supervised learning, the algorithm passively accepts randomly selected

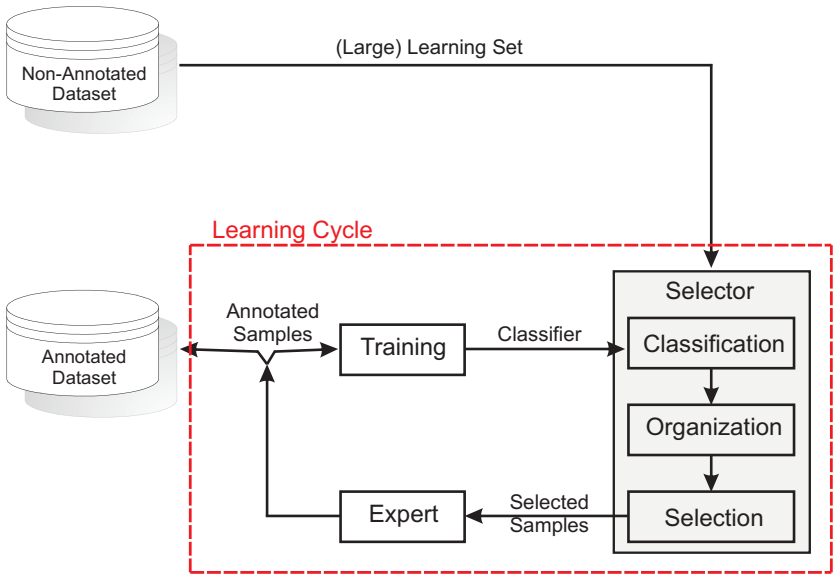


Figure 2.1: Pipeline of the traditional active learning paradigm.

samples from a given dataset to be manually labeled and used to train the classifier. As the dataset grows, an intelligent selection of a reasonably small training set can save considerable human effort and time on manual annotation besides providing a more effective classifier automatically annotating the remaining or future samples.

Active learning can accomplish both goals. Active learning is an iterative supervised learning that actively query the expert (or some other information source) for labels and which the learner participates of its learning process, choosing its learning samples at each iteration. Theoretical results show that it can significantly reduce the number of required training samples as compared to random selection for achieving similar classification accuracy [3, 20]. Active learning techniques can determine which non-annotated samples would be the most informative (i.e., improve the classifier the most) if they are annotated and used as training samples, so allowing to reach higher accuracies with fewer training labels annotated/corrected by the expert.

The idea of using active learning to assist in image annotation has received a lot of research attention [56, 85, 18, 76, 46, 49, 100, 89, 95, 94]. Figure 1 illustrates the complete pipeline of operations for data classification (unsupervised and/or supervised), organization, and selection, that are repeated at each iteration in the previous approaches. The first two are optional, but most recent methods seem to adopt them. At each iteration cycle, the expert is asked to annotate/correct a non-annotated/classified sample set chosen by the selector. As the samples are annotated/corrected by the expert, they are included in the training set to re-train the classifier for the next cycle. The entire learning set is

labeled by the current classifier, organized and finally a subset is selected and presented to the expert. The selector consists of three modules (classification, organization and selection) that are dependent on each other.

Besides the aforementioned inefficiency, most of the existing research on the traditional active learning approaches have focused on binary classification [93, 34, 38, 30]. Relatively few works have been devoted for multi-class active learning and these are typically based on ensemble or committee classifiers [76, 35, 62, 19, 60] or extensions of predominantly binary active learning methods to the multi-class scenario [50, 46].

Our work emphasizes four important mechanisms for active learning: reduction, organization, classification, and selection. Reduction is based on the analysis and disposal of data with no re-evaluation. As far as we know, our approach is unique in this aspect. Organization aims at prioritization for the selection of the highest priority samples, but this decision should be supported by the current instance of the classifier and take into account some pruning criterion. Another difference from previous works is that our method performs classification *after* organization.

A common approach for selecting data in active learning is to choose the most uncertain samples [57, 9, 15, 14, 92], which are the closest to the classification boundary. This simple and intuitive criterion (closest-to-boundary) performs well in some applications [49, 51]. However, some works indicate that better performance can be achieved by taking into account prior knowledge on data distribution. In these cases, clustering techniques have been incorporated into active learning [56, 66, 7, 64, 101, 86]. The clustering information is useful for active learning, since representative samples located at the center of the clusters are more likely to cover all classes and are good candidates to be selected first, for the manual annotation process. Furthermore, samples in the same cluster are likely to have the same label. This assumption could be used to accelerate active learning by reducing the number of annotated samples from a given cluster.

Although some of the aforementioned methods [101, 86, 64, 30, 13, 52] use clustering (unsupervised classification) for sample selection, they neither reduce the dataset based on clustering nor organize the reduced samples just once (a priori). They can be unified as depicted in Figure 2.1, with no data reduction, by reclassifying and/or reorganizing the entire dataset for sample selection at each learning iteration.

Therefore, despite these efforts in active learning, there are important questions that remain open. Namely, in an a priori setting, how to choose the best reduced set from a large learning set? How to organize this reduced set? This work presents a solution to these questions.

2.2 Learning by Optimum-Path Forest

In this Section, we present a methodology for learning based on *optimum-path forest* (OPF), which has been successfully applied to image processing and analysis problems [68, 45, 6, 77, 18], besides it was used in this work. Essentially, this methodology extends a previous approach, called *Image Foresting Transform* [32], for the design of image processing operators from the image domain to the feature space.

The main OPF algorithm is essentially a Dijkstra's algorithm based on multiple sources and more general path-value functions. In the OPF methodology, the training set is modeled as a graph, whose nodes are the samples and arcs connect nodes according to a given adjacency relation. To any given path on this graph (either a single sample or a sequence of distinct samples) we assign a cost given by a connectivity (path-value) function (e.g., the maximum/minimum arc weight along the path). The minimization (maximization) of a path-value map results in an optimum-path forest rooted at representative samples of classes/clusters (called prototypes). One way to view this construction process is to think that the prototypes compete among themselves for the remaining samples. The samples in the same class/cluster of a prototype (nodes of its optimum-path tree) will be those more closely connected to that prototype than to any other. When a new sample is processed, the method estimates, in an incremental way, which prototype would offer the optimum-path to this new sample, as though it were part of the training set, and the class/cluster of that prototype is assigned to the new sample.

In the OPF methodology class/clusters may present arbitrary shapes and have some degree of overlapping, and there is no need to use parametric models. This methodology provides effective supervised, unsupervised and semi-supervised learning techniques, which are described in the following Sections 2.2.1-2.2.3.

2.2.1 Supervised Learning

Given a training set with samples with distinct classes, it is desirable to design a pattern classifier which can assign the true class label to any new sample. Each sample is represented by a set of features and a distance function measures their dissimilarity in the feature space. The training samples are then interpreted as the nodes of a graph, whose arcs are defined by a given adjacency relation and weighted by the distance function. It is assumed that samples from the same class are connected by a path of nearby samples. Therefore, the degree of connectedness for any given path is measured by a connectivity function, which exploits the distances along the path.

In the supervised learning by OPF, the true label of the training samples is known and so it is exploited to identify key samples (prototypes) in each class. Optimum-paths are computed from the prototypes to each training sample, such that each prototype becomes

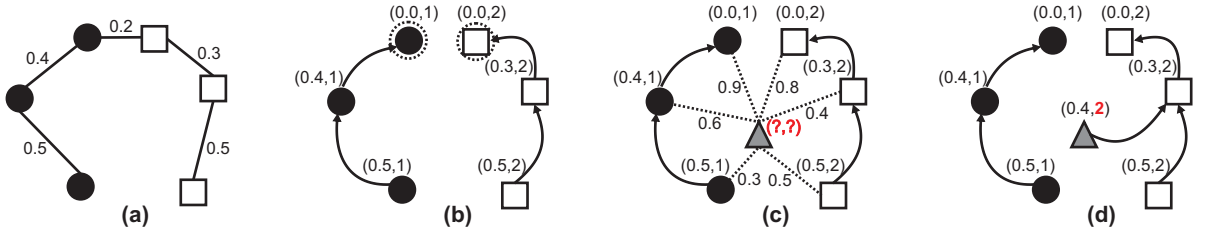


Figure 2.2: Pipeline of the supervised learning by OPF. (a) Complete weighted graph for a simple training set. (b) Resulting optimum-path forest for f_{max} and two given prototypes (circled nodes). The entries (x, y) over the nodes are, respectively, the cost and the label of the samples. The directed arcs indicate the predecessor nodes in the optimum path. (c) Test sample (gray triangle) and its connections (dashed lines) with the training nodes. (d) The optimum path from the most strongly connected prototype, its classification cost 0.4, and label 2 are assigned to the test sample. The test sample is classified in the class square, although its nearest training sample is from the class circle.

the root of an optimum-path tree composed of its most strongly connected samples. The labels of these samples are assumed to be the same as their root. The label assignment is based on optimum connectivity with respect to a set of prototypes rather than based on local distance decisions, as is the case in k-nearest neighbor approaches [36, 79] (See Figure 2.2). In this work, we applied the OPF classifier using a complete graph (implicit representation) and the maximum arc weight along a path as the connectivity function. The prototypes are chosen as samples that share an arc between distinct classes in a minimum-spanning tree of the training set.

The supervised OPF classifier has as advantage a very low computational training cost [63], given that it does not have to optimize parameters. Papa et al. [70] showed that its training phase can be considerably faster than the training phases of SVMs and ANNs, with accuracies better than or equivalent to the ones obtained by these approaches. This OPF classifier has been widely used in several applications, such as remote sensing, emotion recognition through speech processing, automatic vowel classification, biometrics, petroleum well drilling monitoring, medical image segmentation, and robust object tracking [72, 44, 71, 11, 69, 41, 61]. Considerable improvements have been continuously presented to make this OPF classifier more efficient for large datasets [68, 45].

2.2.2 Unsupervised Learning

In the unsupervised learning, we do not know the class label of the training samples. Therefore, in the data clustering, we expect that each cluster contains only samples of the same class and some other information about the application is needed to complete classification. The fundamental problem in data clustering is to identify natural groups

in the unlabeled training set. Natural groups are characterized by high concentrations of samples in the feature space, which form the domes of the probability density function (pdf). These domes can be detected and separated by defining the "influence zones" of their maxima. However, there are different ways to define these influence zones.

Clustering by Optimum-Path Forest (OPF) consists of identifying high concentrations of samples which can characterize relevant clusters for a specific application. This is a non-parametric approach which estimates the number of natural groups in a dataset as the number of maxima of its probability density function (pdf). Each maximum of the pdf will define a cluster as an optimum-path tree rooted at that maximum. The training forest becomes a classifier that can assign to any new sample the label of its most strongly connected root. It can handle plateaux of maximum, by electing a single root (one prototype per maximum), some overlapping among clusters, and groups with arbitrary shapes (Figure 2.3). The OPF clustering does not assume any shape for the clusters, as is assumed by k-means and k-medoids.

In the unsupervised learning algorithm by OPF, an unlabeled training set is interpreted as a graph whose nodes are samples and each node is connected with its k-closest neighbors in the feature space to form directed arcs. The basic idea is then to specify an adjacency relation and a path-value function, compute prototypes and reduce the problem into an optimum-path forest computation in the underlying graph. The pdf value at each node is estimated from the distance between adjacent samples, and a connectivity (path-value) function is designed such that the maximization of a connectivity map defines an optimum-path forest rooted at the maxima of the pdf. In this forest, each cluster is one optimum-path tree rooted at one maximum (root). Each root defines a cluster by conquering the most strongly connected samples according to a path-value function. The clusters are found by ordered label propagation from each maximum, as opposed to the mean-shift algorithm [10] which searches for the closest maximum by following the direction of the gradient of the pdf — a strategy that does not guarantee the assignment of a single label per maximum, and presents problems on the plateaus of the pdf.

According to [77], a suitable pdf $\rho(s)$ can be obtained as node weights of the k-NN graph. The pdf estimation is reduced to the choice of a suitable scale k for data clustering, which requires multiple applications of the algorithm for different values of k in order to select the best clustering result as the one that produces a minimum normalized cut in the k-NN graph. The best k for pdf estimation is found by optimization, but its search interval $[1, kmax]$ may produce different numbers of groups. The parameter $kmax$ represents an observation scale for the dataset. If $kmax$ is too high, it means that we are looking at the dataset from infinity and so, the result will be a single cluster. Higher values of k tend to produce a smaller number of clusters by merging the closest ones. As we approximate the dataset (reducing the value of $kmax$), the number of clusters increases up to some high

Figure 2.3: Example of the unsupervised learning by OPF. Circles represent the pdf's maxima, i.e. samples with higher density values, defining clusters as optimum-path trees rooted at each maximum.

number for $kmax = 1$. Still, the number of possible solution is low, because the method produces an identical number of clusters for several values of $kmax$. This shows the robustness of the method in finding natural groups in the dataset for distinct observation scales. In this work, we chose $kmax$ so as to obtain a number of groups higher than the number of classes known. In our experiments, we use the optimization approach described in [77] with an additional constraint that $k \geq 2c$, where c is the number of classes. This takes into account that one class may be represented by more than one cluster. Note that, we do not use any knowledge on the classes of samples, but we assume that we know how many classes are present in the dataset.

For large datasets, [6] suggests that the k-NN graph (and pdf estimation) be obtained from a subset of random samples from the entire learning set (i.e., a much smaller unsupervised training set). After optimum-path forest computation (training), the cluster labels can be efficiently propagated to the remaining samples by using adjacency radii computed for each training sample. This OPF methodology has already been successfully used for large datasets with about 1.5 million voxels when classifying gray-matter, white-matter, and cerebral spinal fluid in magnetic resonance images of the brain [6].

2.2.3 Semi-Supervised Learning

Semi-supervised learning (SSL) has become an increasingly popular learning approach, given that we have the limited availability of labeled data in contrast to an unbounded number of unlabeled ones. SSL targets the usual situation where labeled data are scarce and unlabeled data are abundant. The semi-supervised approach based on the Optimum-

Path Forest (OPF) methodology [1] has also been successful, overcoming traditional semi-supervised methods, such as Transductive Support Vector Machines (TSVM) [48], UniverSVM [16] and SemiL [105, 43].

In the semi-supervised OPF learning [1], the prototypes are selected from each class among the labeled training samples by using the same strategy as the supervised OPF [68]. Subsequently, all training samples are interpreted as a complete graph and each sample is assigned to the optimum-path tree of its most closely connected prototype. Therefore, the class of a prototype is propagated to all training samples (labeled or unlabeled) in its tree. Since the training set is then entirely labeled, the supervised OPF training algorithm is executed on it in order to select more and/or better prototypes. The resulting optimum-path forest is expected to generalize better than a forest created from only the initial labeled subset.

Figure 2.4 shows the pipeline of the semi-supervised learning by OPF. By computing a MST in the complete graph, we obtain a connected acyclic graph, whose nodes are all labeled samples of the learning set and the arcs are undirected and weighted by the distances between adjacent samples (Figure 2.4a). By removing the arcs between different labels, their adjacent samples becomes prototypes (Figure 2.4b). An optimum-path forest rooted at the prototypes is computed. Then each sample should be assigned to the optimum-path tree of its most closely connected prototype. Figure 2.4c illustrates an optimum-path forest, an unlabeled sample (diamond), and its possible connections with all samples in the training graph. The entries (x, y) over the nodes are, respectively, the cost and the label of the samples. Figure 2.4d shows the label propagation to the unlabeled sample from its most closely connected prototype. After the semi-supervised training, for each test sample (represented by the triangle in Figure 2.4e) is analized its possible connections with all samples in the training graph and it is defined an optimum path to the most closely connected prototype (Figure 2.4f). Figures 2.4g and 2.4h highlight one of the advantages of the semi-supervised approach. After the supervised training (i.e. without consider unlabeled samples), a test sample is classified in the class circle, although its nearest training sample is from the class square.

In [1], the authors assumed that the labeled samples in the training set are good enough to correctly assign the classes of most unlabeled samples in the training set, so they were randomly selected. In our research, we also exploit the combination of active and semi-supervised learning (Section 3.4) in order to select the most representative labeled samples, which will have an impact on decreasing propagated errors on the unlabeled samples, as well as on the construction of more robust classifiers (See Section 4.3.5).

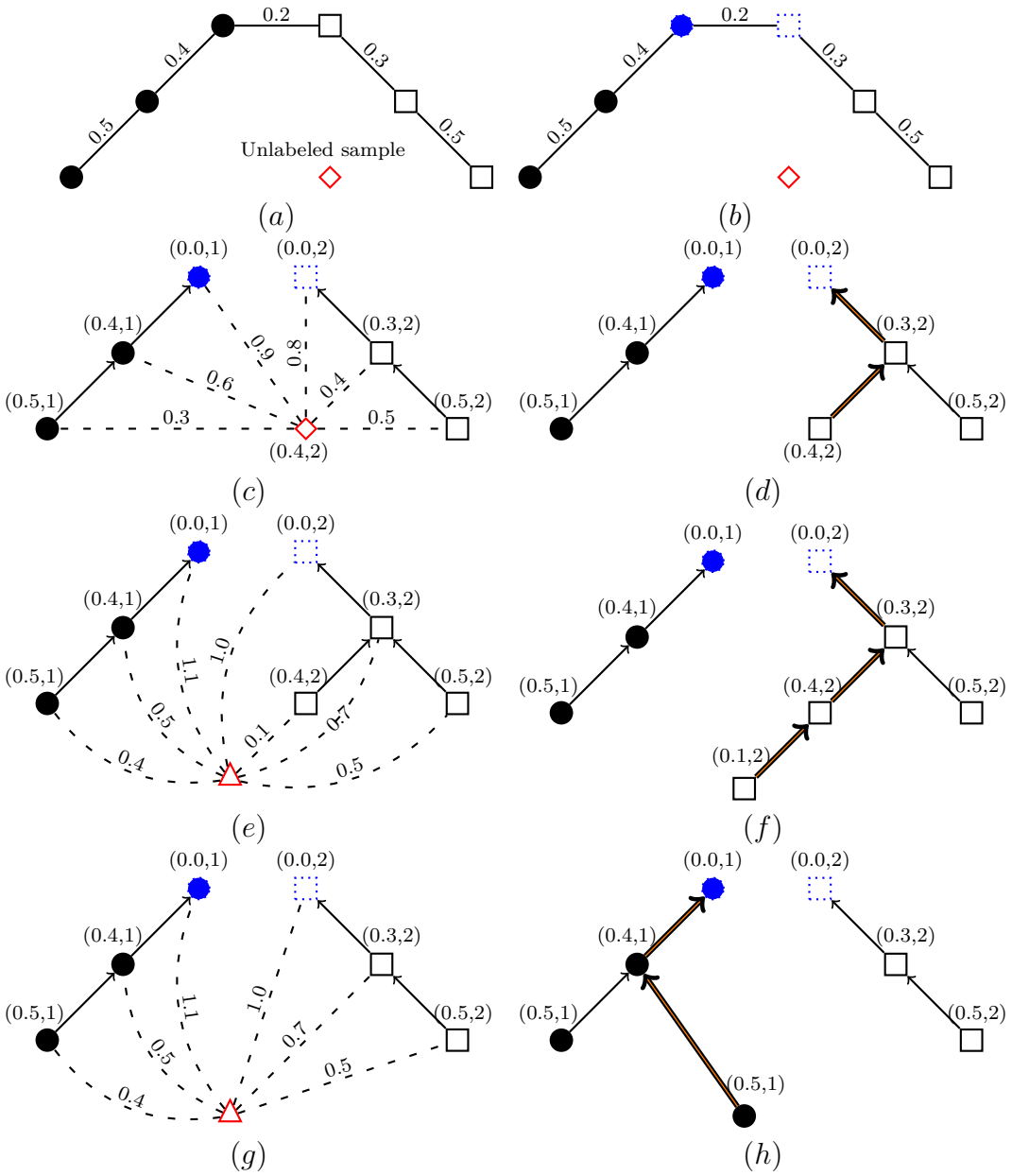


Figure 2.4: Pipeline of the semi-supervised learning by OPF.

Chapter 3

Proposed Paradigm

In this Chapter, we present an effective and efficient data reduction and organization paradigm for active learning, which enables the reduction and/or organization of the learning set a priori (only once).

3.1 Data Reduction and Organization Paradigm

We propose DROP - a Data Reduction and Organization Paradigm [82] - for active learning, in order to select, more efficiently and effectively, a small number of the most representative samples for training a classifier. The learning process aims to present for annotation samples from all classes at the first learning iteration and the most informative (most difficult) samples for classification at the subsequent iterations. In the proposed paradigm, the learning set is reduced into a subset with those relevant samples. The reduced learning set is also organized in pairs of samples such that, among the most difficult ones, the possibility to select sample pairs from distinct classes for annotation will be higher than pairs from the same class.

The major difference and advantage presented by the proposed paradigm (Figure 3.1) is that all non-annotated learning samples in the dataset undergo a reduction and/or organization process only once, unlike traditional active learning methods (as Figure 2.1). DROP analyzes the distribution of all learning samples in the feature space a priori, as preprocessing. Subsequently, during the learning process, a subset of previously arranged samples is selected at each iteration, without the need to classify and re-arrange all samples on the dataset. Thus, the selection process becomes faster, especially considering large datasets, since the improvement of the classifier at each iteration does not require rearranging all samples. Moreover, a remarkably faster selection process is completed by the choice of a small subset of samples and the classification of only these.

The learning process relies on the knowledge of both expert and classifier, at each

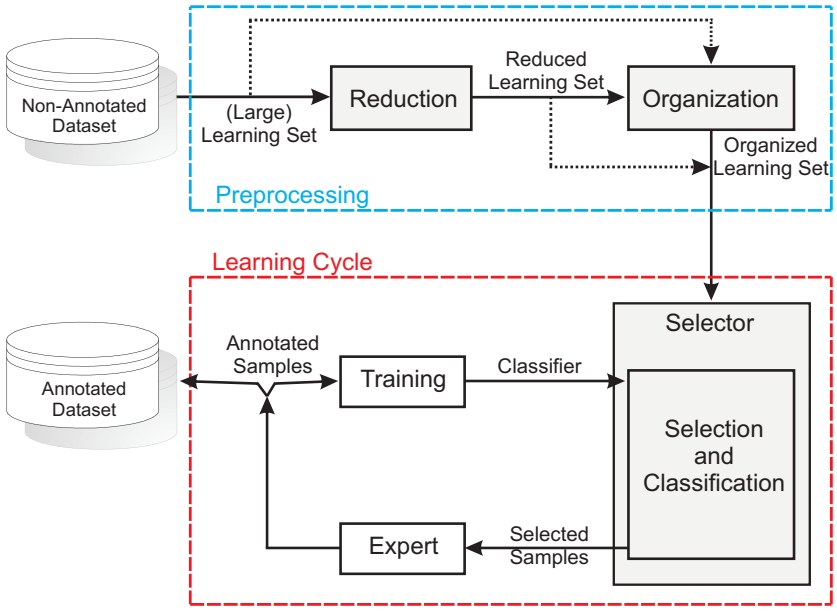


Figure 3.1: Pipeline of the proposed active learning paradigm.

iteration. The classifier actively participates of its learning process by classifying and supporting the choice of the most relevant samples. After this choice, once classified, these samples are displayed to the expert for confirmation of the labels assigned by the current classifier. As the classifier improves throughout the iteration, the expert is required to correct fewer misclassified samples. The expert's time and effort are increasingly lessened. Samples with confirmed/corrected labels are incorporated into the training set and a new classifier instance is generated. Upon sensing that an acceptable accuracy has been reached, the expert can direct the final classifier to annotate the remaining of the dataset. We considered that a expert would be satisfied whenever the measured accuracy remained stable or reached a sufficiently high level for a given application.

DROP being a paradigm, can be applied using different strategies for reduction, organization and selection processes. Figure 3.1 illustrates the execution pipeline of this active learning paradigm, highlighting its main processes which will be explained in the next sections.

3.2 Reduction Strategy

The reduction strategy adopted here aims to prevent the expert from having to annotate a large (and usually wasteful) number of training samples. It can prevent poor selection of samples (i.e. irrelevant selection) from a large learning set, depending on how well the reduced set represents the entire dataset, as well as on its size.

All learning methods will gradually improve when more expert annotations and more learning iterations are allowed. The reduction strategy becomes very important in a process where a goal is to limit the number of iterations to as few as possible. It is well known that on an actual field environment, a large number of learning iterations is tiresome and furthers human error in the annotation process which, consequently, affects the quality of the classifier. In this context, selecting samples that speed up the improvement of the classifier through the iterations becomes critical.

The reduction process is based on the analysis and disposal of data with no re-evaluation. It refines the larger learning set, by applying an a priori sharp identification process, downsizing it into a small subset with essentially the most informative samples. By reducing the learning dataset to a smaller and more informative set, the proposed strategy minimizes the number of learning iterations as well as the experts' time and effort, while decreasing the possibility of misannotation that could occur due to fatigue, in a longer annotation process.

As it was mentioned, any method can be incorporated into the proposed paradigm in order to reduce the learning set and later to organize the reduced one. In the next Section, we present an effective reduction method through data clustering [81].

3.2.1 Reduction by Clustering

A good training set must represent the distribution of the classes in the feature space while being as small as possible, given that efficiency and effectiveness of classification will depend on both factors. However, it is very difficult to determine which samples are the best to train a classifier or even how many of them we must use to obtain good results, since any information about the classes, ideal size, and composition of the training set is unknown prior to annotation.

The distribution of the samples in the feature space can be obtained from their probability density function (pdf). The domes of the pdf can represent clusters of samples that are more likely to belong to a same class. Therefore, data clustering based on a suitable pdf estimation can surrogate the distribution of the classes in the feature space in the case of unlabeled datasets. The maxima of this pdf should represent samples from all classes while its valleys should contain samples from distinct classes that are the most difficult for classification. This leads us to a natural strategy for data reduction, which must estimate a suitable pdf, separate the domes into clusters, and select the maxima and samples between clusters to compose a reduced dataset for active learning. Figure 3.2 illustrates an example of the pipeline of the proposed reduction strategy [81].

Let $\mathcal{Z}_2$ denote an unlabeled learning dataset, such that for every sample $s \in \mathcal{Z}_2$ there is a feature vector $\vec{v}(s)$. For $s, t \in \mathcal{Z}_2$, let $d(s,t)$ denote the distance between s and t

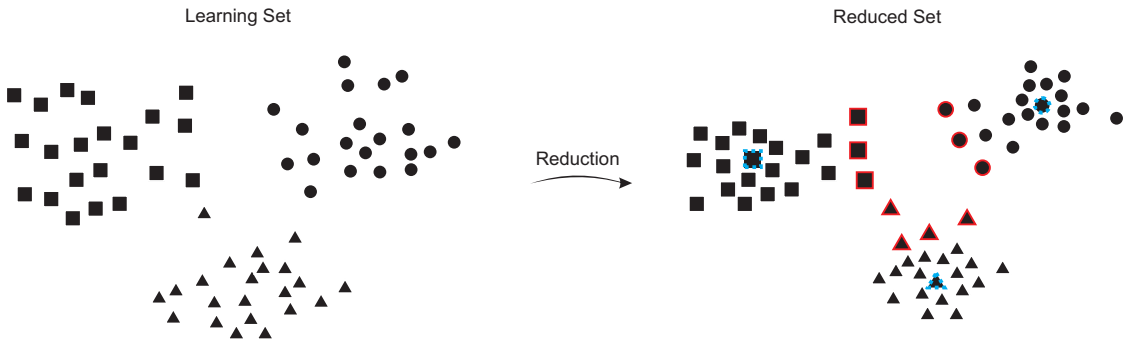


Figure 3.2: An example of pipeline of the proposed reduction strategy.

Algorithm 1: Data Reduction Strategy

input : A non-annotated learning dataset $\mathcal{Z}_2$ and a k-NN relation $\mathcal{A}$.
output: A reduced learning set $\mathcal{Z}'_2$ and a root set $\mathcal{R}$

```
1  Compute clusters of Z_2 using the adjacency relation A;
2  R ← cluster roots;
3  Z'_2 ← nil;
4  for each s ∈ Z_2 do
5      for each t ∈ A(s) do
6          if s.clusterid ≠ t.clusterid then
7              Z'_2 ← Z'_2 ∪ edge(s, t);
8              break;
9          end
10     end
11 end
```

in the feature space (e.g., $d(s,t) = ||\vec{v}(t) - \vec{v}(s)||$). The pair $(\mathcal{Z}_2, \mathcal{A})$ will denote a k-NN graph. That is, a k-NN relation $\mathcal{A}$ is defined over $\mathcal{Z}_2 \times \mathcal{Z}_2$ such that a sample $t \in \mathcal{Z}_2$ is said to be adjacent to a sample $s \in \mathcal{Z}_2$, if t is a k-nearest neighbor of s according to the distance function d.

Algorithm 1 presents the data reduction strategy. From the clustering of $\mathcal{Z}_2$ using a k-NN relation $\mathcal{A}$ (Line 1), we obtain the root set $\mathcal{R}$ (Line 2) as well as the boundary sample set $\mathcal{Z}'_2$ (Lines 4–11). The clustering method assigns cluster-ids to the samples, and a sample $s \in \mathcal{Z}_2$ is a boundary sample if there exists at least one adjacent sample $t \in \mathcal{A}(s)$ whose cluster-id is different from that of s. Hence, (s,t) is a boundary edge between different clusters if the sample t is one of the k-nearest neighbors of the sample s and $s.clusterid \neq t.clusterid$. The roots in $\mathcal{R}$ should increase the possibility of selecting

Figure 3.3: A possible clustering result by OPF (bigger dots indicate maxima of the pdf): a class with two groups, groups with distinct shapes, imbalanced classes with some overlapping, and a class that does not contribute to the boundary set.

samples from all classes and the boundary sample set $\mathcal{Z}_2'$ should contain the most difficult samples for classification.

For large datasets, [6] suggests that the k-NN graph (and pdf estimation) be obtained from a subset of random samples from $\mathcal{Z}_2$ (i.e., a much smaller unsupervised training set). After optimum-path forest computation (training), the cluster labels can be efficiently propagated to the remaining samples in $\mathcal{Z}_2$ by using adjacency radii computed for each training sample. Since the adjacency relation $\mathcal{A}$ is only defined for training samples, the reduced boundary sample set $\mathcal{Z}_2'$ can be obtained based on the mean adjacency radius ω of the training samples. Samples s and t will belong to $\mathcal{Z}_2'$ if $d(s,t) \leq \omega$ and $s.clusterid \neq t.clusterid$.

While data reduction, as preprocessing operation, is one of the key contributions of this work for achieving interactive response times to the expert's label corrections during active learning in large datasets, it should be clear why and in what conditions the reduced data will retain the most relevant samples for selection during active learning. The 2D plot in Figure 3.3 clarifies this issue by illustrating the pros and cons of our method in a general situation that includes: classes with multiple groups, groups with distinct shapes, a reasonable unbalanced number of samples per class, a reasonable overlapping between classes, and also classes that do not contribute to the boundary set between groups.

Given that data reduction is performed by grouping, the choice of the clustering method is paramount. Figure 3.3 shows, for example, a possible result for the OPF clustering algorithm. This approach is similar to the popular mean-shift method [10]: they solve clustering by grouping samples from the same dome of a joint probability density function (pdf), without assuming any type of group shape or even the number of

groups. Their difference is in the respective algorithms. Mean-shift cannot guarantee one group (representative sample) per dome of the pdf, when the maximum of the dome is a plateau, because the algorithm searches the closest maximum, independently for each sample. On the other hand, the OPF clustering algorithm estimates a tree root at each maximum, assigns to each root a distinct label, and propagates their group labels to the remaining samples as a wavefront downwards from the domes. In this way, the wavefronts of distinct labels meet at the valleys (actually, this can be seen as a dual of a watershed transform on the pdf manifold). This label propagation process creates one optimum-path tree (group), rooted at a maximum, per dome of the pdf.

By selecting group representative samples at the first iteration, the method is assuming that those representatives will cover all classes. The clustering parameters can usually be tuned for a given application to guarantee that this happens. By creating a reduced set with samples in the boundary between groups, the method assumes that this set will include those in the boundary between classes, which are the most difficult for classification and usually the most relevant for active learning. The concept of boundary between clusters depends on a pre-defined neighborhood between samples. One can adjust the neighborhood parameter to make the boundary larger, if necessary. Classes that do not create boundary samples are also assumed to be separated from the others based only on their representatives. However, one can propose a simple variant that selects more samples around those representatives, in the first iteration, if necessary. Therefore, the performance of the method might degrade when those premises are not fully satisfied (See Section 3.3.3). However, it has worked well for different applications with suitable features. Moreover, the data reduction strategy can perform a significant downsizing of the learning set (over fifty percent in our experiments).

It is important to highlight that different clustering methods (such as OPF, k-means, k-medoids or mean-shift) could certainly be used in the data reduction process. We evaluated our strategy using OPF and k-means clustering methods (see Section 4.3). The OPF clustering does not assume any shape for the clusters, as is assumed by k-means and k-medoids. The k-means algorithm [59] finds k clusters by the sum of the distance from the data samples to the nearest representative. In order to cover all/most classes as the centers of the groups, we define the value of k as $2c$. The boundary sample set is comprised of samples whose the closest counterpart falls lies on a different cluster. A first instantiation (Cluster-OPF-Rand) of the reduction strategy was developed to illustrate its effectiveness. It is based on the OPF methodology, while relying on clustering and classification for the learning process. In this particular instantiation, the organization of the reduced set occurs in a randomized fashion (see Section 4.3.1). Other instantiations were developed, encompassing different ways to organize the reduced set (Sections 3.3.1 and 3.3.2).

3.3 Organization and Selection Strategies

An effective and efficient organization strategy is essential to speed up the selection of the most useful samples for improving the classifier. We propose organization strategies that allows to achieve high accuracy quickly, further improving its efficiency.

The proposed paradigm DROP also enables the organization process to occur a priori. By organizing those samples only once (in a preprocessing process), the selection of samples becomes quite fast. Moreover, by interlacing the choice of samples and their classification into a joint process, the selection strategy decreases the number of samples that require expert annotation, while choosing the most relevant ones. In addition, classification does not occur for all samples in the database but to a small set of samples. Therefore, we can safely claim that our proposal is a powerful approach to handle massive datasets, since it does not require the classification and reorganization of the entire dataset at each iteration, unlike traditional approaches.

3.3.1 Decreasing Boundary Edges (DBE)

After applying our boundary reduction strategy, which performs a significant downsizing (by up to fifty percent in our experiments) of the learning set, it is important to organize the remaining samples in a prioritized fashion, so that the most relevant ones are more readily available for selection.

We propose a **D**ecreasing **B**oundary **E**dges (DBE) organization strategy [83], in order to effectively arrange the samples of the reduced set. Figure 3.4 illustrates the preprocessing (a priori reduction and organization) performed by DBE strategy. DBE organizes the reduced set based on the decreasing weight order of its boundary edges. The idea of prioritizing the largest edges formed by boundary samples is due to those samples being, more likely than not, of different classes.

This data organization leads to a considerable reduction of classification errors in the first iterations. It amounts to a major advantage of our strategy, since it requires very few iterations in the learning process to attain high accuracy. Moreover, once the learning set has been pre-arranged, unlike in the traditional methods, DBE does not require classification and reorganization of all non-annotated samples in the dataset at each iteration. For this reason, the selection strategy turns out to be very fast even for large datasets.

The selection strategy consists of choosing the samples (training set) that will be used to train the classifier throughout the iterations. In the first learning iteration, we display to the expert the roots of the clusters computed in the boundary reduction process, who annotates their labels. These annotated samples constitute the training set of the first instance of the classifier. During the learning cycle, the samples in the ordered list of

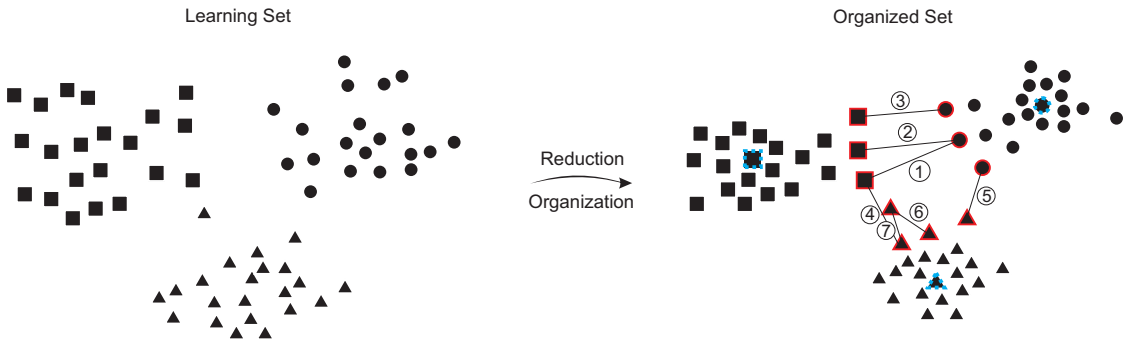


Figure 3.4: An example of pipeline of the preprocessing (a priori reduction and organization) performed by DBE strategy. The samples wrapped by a dashed line (roots) comprise the initial learning set. The numbered circles indicate the edges sorted in a decreasing weight order to be selected at each iteration.

edges of the reduced set are labeled by the current classifier and the samples on edges that receive different labels are selected. These two processes, classification and selection, are performed alternately until the number of samples ($2c$) to be displayed to the expert at each iteration is reached. Note that, this approach does not require the classification of all samples in the dataset, at each iteration.

Furthermore, since the selected samples are labeled by the current classifier, as the user verifies the assigned classes, he is only required to annotate the misclassified ones. In this way, the expert's time and effort are significantly reduced. As the classifier improves at each iteration, the number of samples incorrectly classified is increasingly reduced. After the labels are confirmed/corrected by the expert, the newly annotated samples are incorporated into the training set and a new classifier instance is generated.

Upon sensing that an acceptable accuracy has been reached, the expert can direct the final classifier to annotate the remaining of the dataset. In our experiments, we considered that an expert would be satisfied whenever the measured accuracy remained stable or reached a sufficiently high level for a given application.

Algorithm 2 shows DBE organization and selection strategies. After the preprocessing performed by graph clustering in the reduction process (Section 3.2.1), we obtain sets $\mathcal{R}$ and $\mathcal{Z}'_2$, comprised of the root of each cluster and boundary edges, respectively. The initial training set $\mathcal{Z}_1$ consists of the roots that form the set $\mathcal{R}$ (Line 1). An ordered list $\mathcal{L}$ is created with the edges from $\mathcal{Z}'_2$ in decreasing order of weights (Line 2). In Line 3, the expert annotates the classes of the roots in $\mathcal{Z}_1$. The loop on Lines 4–9 encompasses the processes of (re-)training and selection. At each iteration, $c-1$ edges from $\mathcal{L}$ are analyzed, where c is the number of classes. As edges are considered, their samples are labeled by the current classifier and the ones with distinct classes are selected to be displayed to

Algorithm 2: Organization and Selection Strategies - DBE

input : The reduced learning set $\mathcal{Z}'_2$, the root set $\mathcal{R}$ and the number of classes c
output : Trained classifier
auxiliaries: An annotated training set $\mathcal{Z}_1$, a sorted list $\mathcal{L}$ of boundary edges and a boundary sample set $\mathcal{Z}'_1$

1 $\mathcal{Z}_1 \leftarrow \mathcal{R}$;
2 $\mathcal{L} \leftarrow$ edges from $\mathcal{Z}'_2$ in decreasing order of weights;
3 Expert annotates the class of each root in $\mathcal{Z}_1$;
4 **while** *expert is not satisfied* **do**
5 (Re-)train the classifier with $\mathcal{Z}_1$;
6 $\mathcal{Z}'_1 \leftarrow 2 \cdot (c-1)$ new samples classified into distinct classes, following the order given by $\mathcal{L}$;
7 Expert accepts/corrects classes of samples in $\mathcal{Z}'_1$;
8 $\mathcal{Z}_1 \leftarrow \mathcal{Z}_1 \cup \mathcal{Z}'_1$;
9 **end**

the expert. In this way, the growth of the training set is controlled since only the most beneficial samples are retained. The learning cycle is repeated until the expert is pleased with the success rate on the selected set.

Note that different classifiers may be considered in the classification and selection processes. We evaluated DBE strategy using OPF classifiers (See Section 4.3.2).

3.3.2 Minimum Spanning Tree Boundary Edges (MST-BE)

We also propose a Minimum-Spanning Tree Boundary Edges (MST-BE) organization strategy [82]. The MST-BE strategy presents a better organization way, selecting the more relevant samples than by DBE one. Figure 3.5 illustrates the preprocessing (a priori reduction and organization) performed by MST-BE strategy.

In order to increase the possibility to select boundary samples from distinct classes in $\mathcal{Z}'_2$, the method interprets this set as a complete graph weighted by the distance $d(s,t)$ between samples in the feature space, computes a Minimum Spanning Tree (MST) on it, and organizes the MST edges by the decreasing weight order. The organization of the boundary set in decreasing order of distance between samples on its MST assumes that samples from the same class are usually the closest ones, and hence, they will be placed at the end of the resulting (organized) sample list, increasing the possibility of selecting samples from distinct classes sooner. Given that boundary edges with lower weights are more likely to be in the same class, MST-BE allows us to prioritize samples connected by edges with higher weights and classified in distinct classes during the selection strategy

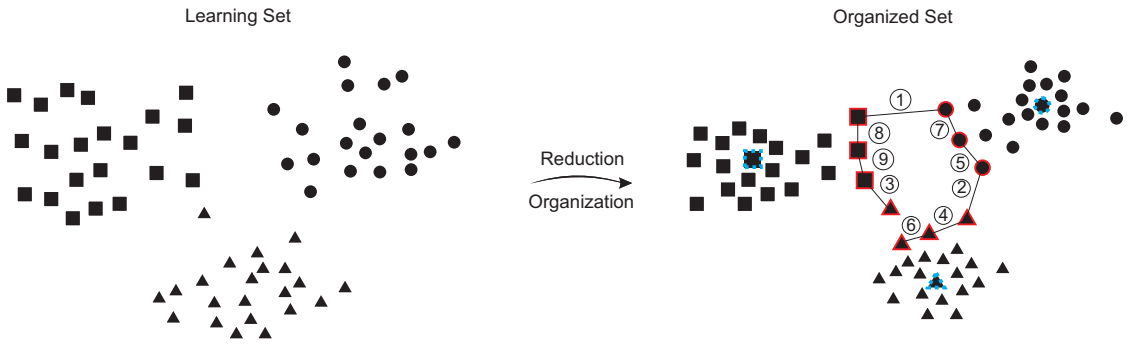


Figure 3.5: An example of pipeline of the preprocessing (a priori reduction and organization) performed by MST-BE strategy. The samples wrapped by a dashed line (roots) comprise the initial learning set. The numbered circles indicate the edges sorted in a decreasing weight order to be selected at each iteration.

for expert annotation/verification.

Therefore, this strategy increases the probability of the most informative samples being selected earlier and allows a more efficient and effective training of the classifier, leading to a considerable reduction in classifier errors in a few iterations. This amounts to a major advantage of our strategy, given that it provides better accuracies with less expert involvement in the learning process. Moreover, by reducing and organizing the samples previously, the selection of samples becomes quite fast.

Algorithm 3 shows MST-BE organization and selection strategies. After the preprocessing, performed by the clustering in the reduction process (Section 3.2), we obtain sets $\mathcal{R}$ and $\mathcal{Z}'_2$, comprised of the root of each cluster and boundary edges, respectively. The initial training set $\mathcal{Z}_1$ consists of the roots that form the set $\mathcal{R}$ (Line 1). The samples in $\mathcal{Z}'_2$ are ranked in order of importance by computing an MST (Line 2) and creating an ordered list $\mathcal{L}$ with the edges in decreasing order of weights (Line 3). In the first learning iteration, the roots of the clusters in $\mathcal{Z}_1$ are displayed to the expert, who annotates their labels (Line 4). During the learning cycle (loop on Lines 5–10) occurs the processes of (re-)training and selection. At each iteration, $c-1$ edges from $\mathcal{L}$ are analyzed, where c is the number of classes. As edges are considered, their samples are labeled by the current classifier and the ones that receive different labels are selected to be displayed to the expert. The classification and selection processes are performed alternately until the desired number of samples ($2c$) to be displayed to the expert at each iteration is reached.

Different classifiers may also be considered in the classification and selection processes. We evaluated our paradigm using SVM and OPF classifiers (see Section 4.3.3).

Algorithm 3: Organization and Selection Strategies - MST-BE

input : The reduced learning set $\mathcal{Z}'_2$, the root set $\mathcal{R}$ and the number of classes c
output : Trained classifier
auxiliaries: An annotated training set $\mathcal{Z}_1$, a sorted list $\mathcal{L}$ of MST edges and a boundary sample set $\mathcal{Z}'_1$

```
1  Z_1 ← R;
2  Compute a MST of Z'_2;
3  L ← edges from MST in decreasing order of weights;
4  User annotates classes of roots in Z_1;
5  while user is not satisfied do
6  |  (Re-)train the classifier with Z_1;
7  |  Z'_1 ← 2 · (c − 1) new samples classified into distinct classes, following the order
   |  given by L;
8  |  User accepts/corrects classes of samples in Z'_1;
9  |  Z_1 ← Z_1 ∪ Z'_1;
10 end
```

3.3.3 Root Distance Based Sampling (RDS)

In laboratory routine, the diagnosis of intestinal parasites currently relies on the visual analysis of fecal samples using optical microscopy. This form of analysis is often compromised by the presence of fecal impurities, incorrect human procedures, and lack of human knowledge. Usually, visual diagnosis takes several minutes of a specialist per slide [37] - an exhaustive process whose abbreviation may seriously compromise the quality of the diagnosis. We have developed an automated system for this application, which can considerably improve the diagnosis sensibility and reliability [87, 88].

In our system, each lab exam produces about 2,700 images of 4M pixels each for analysis, and each image may contain from tens to thousands of objects to be labeled either as an impurity or as some species of parasite among the 15 most common ones in Brazil. This image acquisition process can quickly generate a large dataset, becoming unfeasible for full manual annotation. Given that random sampling is not usually the best alternative [3, 20], this problem calls for an active learning method that can select a reasonably small training set consisting of the most useful samples for expert verification (manual annotation first, and, subsequently, label correction or confirmation) for a few learning iterations. The resulting pattern classifier should then be able to correctly label the remaining and future ensuing samples.

Active learning is also desirable for re-evaluation and improvement of the system's performance, which can benefit from the growth of the dataset, after some number of

new exams. During active learning, the classifier actively participates in its own learning process by suggesting labels for expert supervision at each iteration. However, in a real problem of diagnosis of parasites, impurities are exceptionally abundant, form several groups in the feature space, and are quite similar to some species of parasites (see Figure 4.4b), resulting in a major challenge for existing methods in the literature.

In this context, under a scenario with the presence of the fecal impurity class, the proposed strategies with a reduction process (i.e. Cluster-OPF-Rand, DBE and MST-BE strategies) showed to be considerably less effective. The data reduction can discard crucial samples to the learning process. In this case, it is important to be careful because some parasite species and/or impurities can be out of the cluster border (See Figure 3.6).

Therefore, we also searched for a more robust solution in the presence of a diverse class (such as impurities in the diagnosis of parasites), which previously organizes the data but without discarding any of them. We propose a new active learning strategy, called **R**oot **D**istance-Based **S**ampling (RDS) [84] that pre-organizes the data and then properly balances the selection of diverse and uncertain samples for training. As mentioned, our strategy follows the idea of uncertainty and diversity sampling [5, 33], aiming to select samples from all classes faster (diversity) and, at the same time, the most difficult samples (uncertainty) for classification. Moreover, after the first iteration, the classifier also participates of the sampling choice in the selection process.

In our strategy, data organization is based on clustering, followed by the sorting of the samples within each group based on their distance to their representative (root) sample in the group (Figure 3.7). In the first iteration, the expert labels the root samples used to train the first instance of the classifier. In the subsequent iterations, the current classifier selects samples from each group according to the corresponding ordered list so long as their classification does not match the class of the corresponding root. When this condition is not satisfied, RDS selects uncertain samples in their decreasing distance to the cluster's root. This strategy allows us to explore data diversity by covering all classes faster and, at the same time to select uncertain samples, which are more useful for training the classifier for the succeeding iteration. Differently from most active learning approaches, our strategy avoids reprocessing the large dataset at each learning iteration, enabling the halting of sample selection after a desired number of samples per iteration and so providing interactive response time.

The organization strategy is represented by Algorithm 4. From the clustering of $\mathcal{Z}_2$ (Line 1), we obtain the root set $\mathcal{R}$ (Line 2), as well as sets $\mathcal{C}_i$ of samples from each cluster i, $i = 1, 2, \ldots, nc$, where nc is the number of clusters. For each cluster set $\mathcal{C}_i$, we compute the distance $d(r_i, q)$ between its root r_i and each sample q in $\mathcal{C}_i$ (Lines 3–7). As output, we obtain organized learning sets $\mathcal{L}_i$, which are composed of samples in increasing order of these (Line 8).

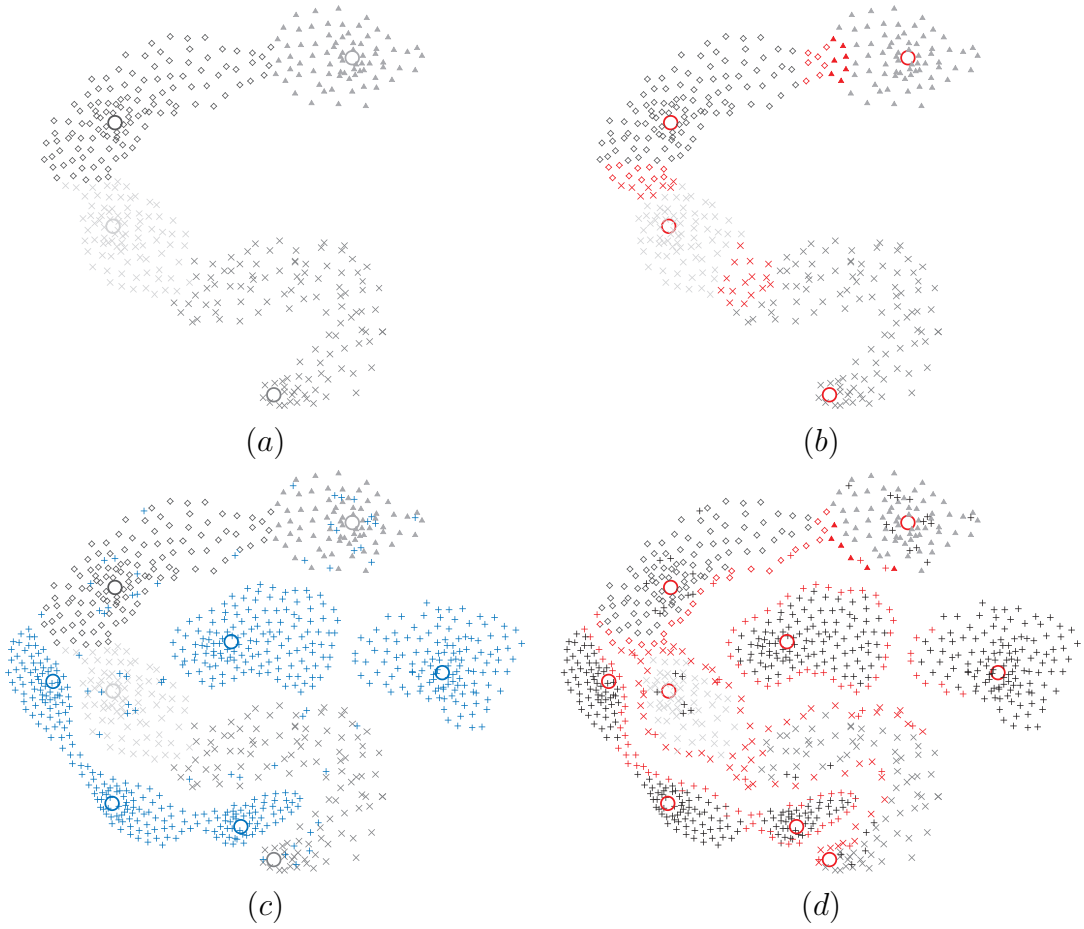
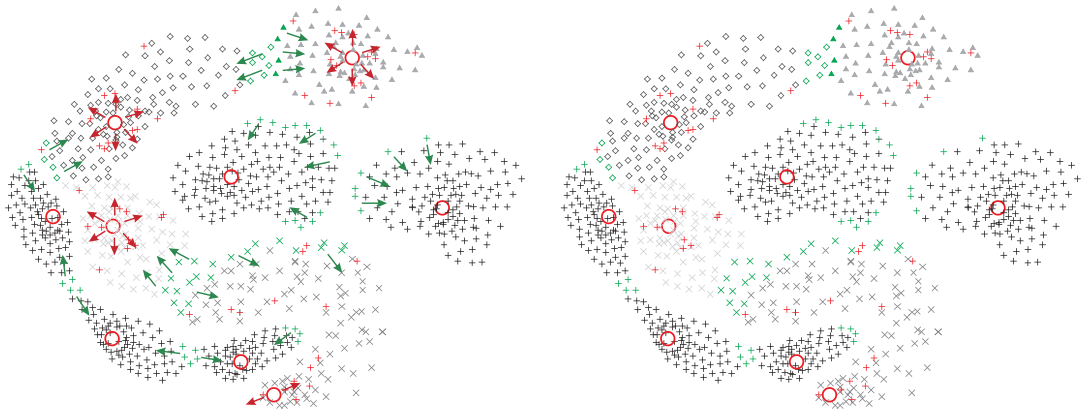


Figure 3.6: An example of pipeline of the preprocessing (a priori reduction and organization) performed by MST-BE strategy on the diagnosis of intestinal parasites. (a) under a scenario without the presence of the fecal impurity class. (b) cluster roots and boundary between distinct cluster samples that form the reduced learning set. (c) under a scenario with the presence of the fecal impurity class. (d) samples (cluster roots and boundary between distinct cluster samples) from the reduced learning set do not correspond to the boundary between different class samples. This reduced set includes too many needless impurities, besides it does not contain crucial (impurity) samples which were discarded in the reduction process.

During the learning cycle, one sample at a time on each ordered set $\mathcal{L}_i$ is labeled by the current classifier, and becomes selected if it receives a different label than the one of root r_i from $\mathcal{C}_i$. Note that, the classifier does not label all the samples in the dataset. Both phases, classification and selection, are performed alternately until the number of samples to be displayed to the expert at each iteration is reached. If there are no more



(a) (b)

Figure 3.7: An example of pipeline of the preprocessing (a priori organization) performed by RDS strategy. (a) selection of samples of each cluster (in each ordered list), i.e. selection of samples closer to the roots and whose labels are distinct from those of the roots, and samples in the decreasing distant order from their roots. (b) The most uncertain samples and samples with greater diversity selected by RDS strategy.

Algorithm 4: Organization Strategy

input : A non-annotated learning dataset $\mathcal{Z}_2$.
output : Organized learning sets $\mathcal{L}_i$ and a root set $\mathcal{R}$.
auxiliaries: Sets $\mathcal{C}_i$ with samples of the cluster $i = 1, 2, \ldots, nc$ and the number nc of clusters.

```
C_i, i = 1,2,...,nc, ← Compute clusters of Z_2;
R ← cluster roots;
for each r_i ∈ R, i = 1,2,...,nc do
    for each q ∈ C_i do
        Compute d(r_i, q)
    end
end
L_i ← Organize samples in C_i, i = 1,2,...,nc by their increasing order of distance to r_i.
```

samples from $\mathcal{L}_i$ whose label is different from the label of r_i, we continue with the sampling criterion of uncertain samples, selecting samples from $\mathcal{L}_i$ in decreasing order of distance from r_i (i.e., the ones closer to the boundary between clusters are selected first).

Since the selected samples are automatically labeled by the current classifier, the expert just has to verify the assigned classes and annotate the misclassified ones. After the

Algorithm 5: Selection Strategy

input : Organized learning sets $\mathcal{L}_i$, $i = 1, 2, \ldots, nc$, and the root set $\mathcal{R}$
output : Trained classifier
auxiliaries: An annotated training set $\mathcal{Z}_1$ and a sample set $\mathcal{Z}'_1$

1 $\mathcal{Z}_1 \leftarrow \mathcal{R}$;
2 Expert annotates classes of roots in $\mathcal{Z}_1$;
3 **while** *expert is not satisfied* **do**
4 (Re-)train the classifier with $\mathcal{Z}_1$;
5 $\mathcal{Z}'_1 \leftarrow$ new samples classified into distinct classes, following the order given by $\mathcal{L}_i$, $i = 1, 2, \ldots, nc$, or samples in the decreasing distant order from the root r_i;
6 Expert accepts/corrects classes of samples in $\mathcal{Z}'_1$;
7 $\mathcal{Z}_1 \leftarrow \mathcal{Z}_1 \cup \mathcal{Z}'_1$;
8 **end**

labels are confirmed/corrected by the expert, the newly annotated samples are incorporated into the training set and a new classifier instance is generated. As the classifier improves at each iteration, the number of samples incorrectly classified is increasingly reduced. In this way, the expert's time and effort are also significantly diminished.

Algorithm 5 shows the selection strategy. After the pre-processing performed by the clustering method (Section 2.2.2) in the organization process, we obtain sets $\mathcal{R}$ and $\mathcal{L}_i$, comprised of the root of each cluster and ordered samples from each cluster $\mathcal{C}_i$, respectively. The initial training set $\mathcal{Z}_1$ consists of the roots that form the set $\mathcal{R}$ (Line 1). In Line 2, the expert annotates the classes of the roots in $\mathcal{Z}_1$. The loop in Lines 3–8 encompasses the processes of (re-)training and selection. At each iteration of the loop, a new classifier is trained using the current training set $\mathcal{Z}_1$, and the samples $\mathcal{Z}'_1$ are selected, according to the selection strategy described above, to be displayed to the expert.

The growth of the training set is controlled since only the most beneficial samples are retained. Selecting samples of each cluster (in each ordered list $\mathcal{L}_i$), allows us to obtain samples with greater diversity. On the other hand, the selection of samples closer to the roots and whose labels are distinct from those of the roots, and samples in the decreasing distant order from their roots, allows us to obtain the most uncertain samples for the classifier.

3.4 Active Semi-Supervised Learning Strategy (ASSL)

Given the limited availability of labeled samples in contrast to an unbounded number of unlabeled ones, semi-supervised learning (SSL) has become an increasingly popular learning approach. However, most of researches in this area typically assumes that the labeled

set is given and fixed. Recent applications involving very large amounts of unlabeled samples would certainly benefit from a combination of active learning and semi-supervised learning strategies, as this would allow for the identification and labeling of a small number of better representative samples. Despite some efforts in active semi-supervised learning, their success depends on an approach suitable to be applied to real large datasets.

We introduce a strategy, called Active Semi-Supervised Learning (ASSL) [80]. ASSL is a novel integration of semi-supervised learning (proposed by [1] and described in Section 2.2.3) and a priori-reduction and organization criteria (proposed by [82]) for active learning. It differs from standard active learning in which all samples in the database have to be classified and/or reorganized at each learning iteration. In the ASSL, the learning set is substantially reduced and the organization of samples takes place only once. The active learning strategy used here (and presented in Section 3.3.2) reduces the possibility of selecting an irrelevant sample from a large learning set, since a well chosen size reduction process and an a priori ordering allow essentially good informative samples.

Figure 3.8 illustrates an example of pipeline performed by ASSL strategy. In the semi-supervised learning, the training set $\mathcal{Z}_1$ is composed of labeled and unlabeled samples. The labeled samples will be selected from the reduced set, root ($\mathcal{R}$) and boundary ($\mathcal{Z}_2'$) sets, which are initially labeled by the clustering strategy. Later, those selected samples will have their labels verified/corrected by the expert. The unlabeled samples will be selected from $\mathcal{Z}_2''$ comprised of the remaining samples from $\mathcal{Z}_2 \backslash R \cup Z_2'$. In the first learning iteration, the roots of the clusters computed during the data reduction process are displayed to the expert, who annotates their labels. These annotated samples, constitute the first labeled set. The unlabeled set is selected in a randomized way with twice as many elements as the labeled set. The union of these sets constitute the training set for the first instance of the semi-supervised classifier.

During the learning cycle, the samples in the ordered list of edges of the MST are labeled by the current classifier and the samples on edges that receive different labels are selected. These two phases, classification and selection, are performed alternately until the number of labeled samples to be displayed to the expert at each iteration is reached. It does not require the classification of all samples in the dataset, at each iteration. Furthermore, since the selected samples are automatically labeled by the current classifier, as the expert verifies the assigned classes, he is only required to annotate the misclassified ones. In this way, the expert's time and effort are significantly reduced. As the classifier improves at each iteration, the number of samples incorrectly classified is increasingly reduced. After the labels are confirmed/corrected by the expert, the newly annotated samples as well as the randomly chosen unlabeled ones are incorporated into the training set and a new classifier instance is generated. Upon perceiving that an acceptable accuracy has been reached, the expert can direct the final classifier to annotate what remains of the

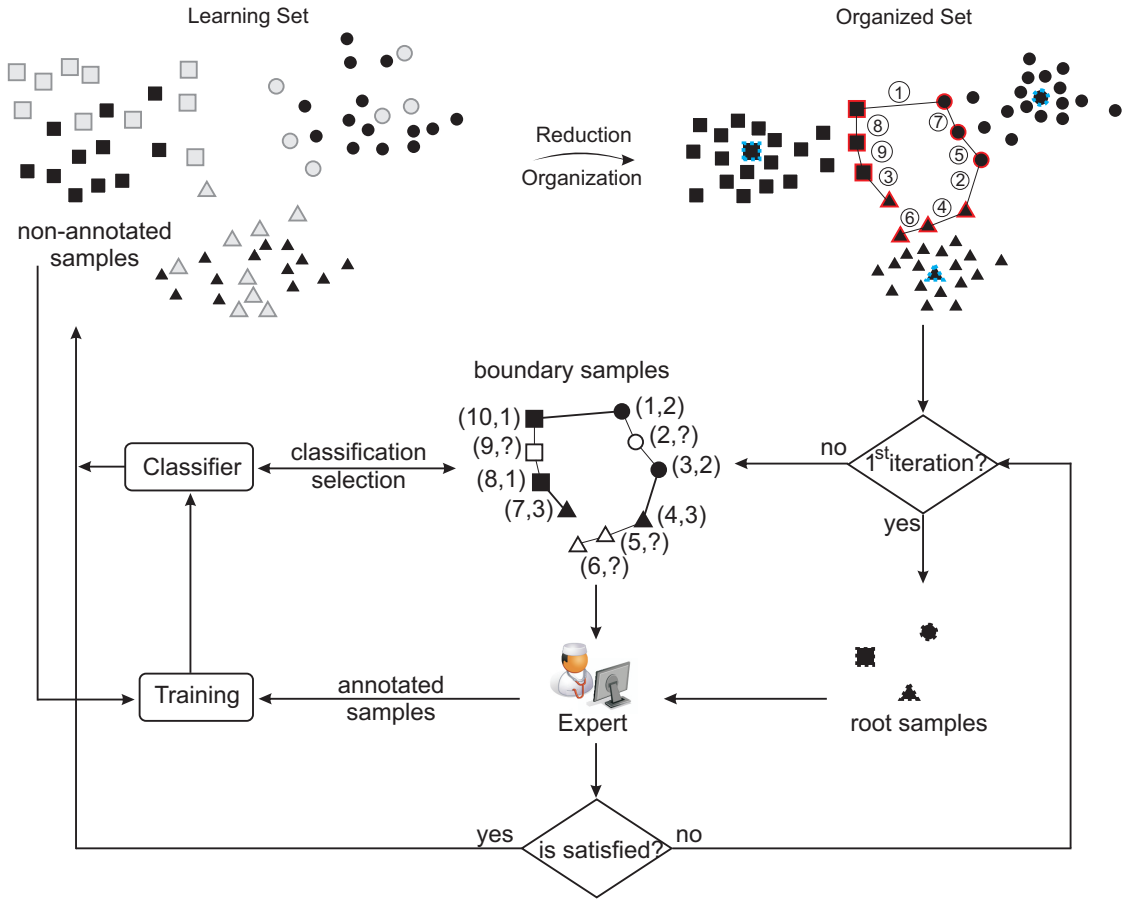


Figure 3.8: An example of pipeline performed by ASSL strategy.

dataset. In our experiments, we considered that an expert would be satisfied whenever the measured accuracy remained stable or reached a sufficiently high level for a given application.

Algorithm 6 shows ASSL active semi-supervised learning strategy. After the preprocessing carried out by the clustering method, in the reduction process (Section 3.2.1), we obtain sets $\mathcal{R}$ and $\mathcal{Z}'_2$, comprised of the root of each cluster and boundary edges, respectively. The initial labeled set $\mathcal{Z}'_1$ consists of the roots that form the set $\mathcal{R}$ (Line 1). In Line 2, the expert annotates the classes of the roots in $\mathcal{Z}_1$. The initial unlabeled set $\mathcal{Z}''_1$ consists of randomized samples from the remaining unlabeled sample set $\mathcal{Z}''_2$ (Lines 3–4). In Line 5, we obtain the first training set $\mathcal{Z}_1$ formed by the labeled $\mathcal{Z}'_1$ and unlabeled $\mathcal{Z}''_1$ sets. The loop on Lines 6–13 encompasses the processes of (re-)training and selection. At each iteration, edges from $\mathcal{Z}'_2$ are analyzed. As edges are considered, their samples are labeled by the current semi-supervised classifier and the ones with distinct classes are selected to be displayed to the expert. In this way, the growth of the training set is

Algorithm 6: Active Semi-Supervised Learning Strategy - ASSL

input : A learning dataset $\mathcal{Z}_2$, the boundary set $\mathcal{Z}'_2 \subset \mathcal{Z}_2$ of sorted MST edges, the root set $\mathcal{R} \subset \mathcal{Z}_2$ and the number of classes c
output : Trained semi-supervised classifier
auxiliaries: The unlabeled set $\mathcal{Z}''_2$, the training set $\mathcal{Z}_1$, the selected labeled boundary set $\mathcal{Z}'_1$, and the selected unlabeled set $\mathcal{Z}''_1$

$\mathcal{Z}'_1 \leftarrow \mathcal{R}$;
Expert annotates classes of roots in $\mathcal{Z}'_1$;
$\mathcal{Z}''_2 \leftarrow \mathcal{Z}_2 \backslash \mathcal{R} \cup \mathcal{Z}'_2$;
$\mathcal{Z}''_1 \leftarrow (2 \cdot |\mathcal{Z}'_1|)$ random samples from $\mathcal{Z}''_2$;
$\mathcal{Z}_1 \leftarrow \mathcal{Z}'_1 \cup \mathcal{Z}''_1$;
while *user is not satisfied* **do**
    (Re-)train the semi-supervised classifier with $\mathcal{Z}_1$;
    $\mathcal{Z}'_1 \leftarrow$ new samples classified into distinct classes, following the order given by $\mathcal{Z}'_2$;
    Expert accepts/corrects classes of samples in $\mathcal{Z}'_1$;
    $\mathcal{Z}''_1 \leftarrow (2 \cdot |\mathcal{Z}'_1|)$ random samples from $\mathcal{Z}''_2$;
    $\mathcal{Z}_1 \leftarrow \mathcal{Z}_1 \cup \mathcal{Z}'_1 \cup \mathcal{Z}''_1$.
end

controlled since only the most beneficial labeled samples are retained and their labels are propagated to the unlabeled samples. The learning cycle is repeated until the expert is pleased with the success rate on the selected set.

A first instantiation (ASSL-OPF) of the active semi-supervised learning strategy was developed to illustrate its effectiveness. It is based on the OPF methodology, while relying on clustering and classification for the learning process. In this particular instantiation, we used the active learning strategy MST-BE (proposed by [82] and presented in Section 3.3.2), as well as the semi-supervised learning strategy OPFSemi (proposed by [1] and described in Section 2.2.3). Other instantiations can be developed, encompassing different active learning methods and/or semi-supervised learning ones.

Chapter 4

Experiments and Results

In this Chapter, we describe the scenarios (Section 4.1), datasets (Section 4.2) and results (Section 4.3) used in the experiments.

4.1 Scenarios

Since our paradigm aims at selecting a suitable set of samples to constitute the training set used throughout the iterations, an appropriate measure of quality was required. To achieve this goal, we compared the performance of each method (measuring the accuracy on an unseen test set and the number of annotated samples during the learning process). We also considered the computational time for selecting the most representative samples throughout the learning, as well as the time gain for classification on the reduced learning set. For the active semi-supervised learning, we also presented the percentage of propagated errors on the unlabeled set, as well as the number of known classes.

The proposed paradigm was evaluated against two baseline learning methods: Al-SVM [95], which selects samples from the entire learning set at each iteration using an SVM classifier, and the Rand method, in which samples are randomly selected from the entire learning set. As one might expect, the wider choice of samples here does not necessarily yield gains over the results of the proposed methods due to the benefits of clustering.

We show that our paradigm is effective in lessening the expert's effort in data annotation in order to produce an accurate classifier. To demonstrate the effectiveness of the proposed paradigm using different clustering techniques, we performed comparisons using the OPF and $k$means algorithms for the reduction process. The $k$means algorithm is well known and widely used. See [59] for a detailed implementation. The $k$means algorithm finds k representatives r_1,...,r_k of the learning set, so as to minimize the sum of the distance from the data samples to the nearest representative. Choosing the number k of clusters is a general problem for all clustering algorithms, and a variety of more or

less successful methods have been devised for this problem. In this work, in order to obtain representative samples that cover all/most classes, we define the value of k as $2c$, where c is the number of classes. While $k$means requires specification, OPF can estimate the number of clusters (higher than $2c$). Other clustering techniques could potentially be employed in the organization process as well.

Similarly, alternative supervised classifiers may be considered in the classification and selection processes. We compared the performance of our paradigm using SVM and OPF classifiers, due to the extensive use of the former, and the considerable advantages (such as speed, simplicity, being multi-class, parameter independence) over SVMs offered by the latter [68] when handling large datasets. In order to facilitate the comparison among the methods, when applicable, they were labeled as a triple, consisting of the clustering method, active learning method, and classification method, separated by an underscore character. The methods were denoted as $k$means_method_OPF, OPF_method_OPF, $k$means_method_SVM, OPF_method_SVM, and Al-SVM.

To attain unbiased analysis, we considered the size of the selected set of each iteration as being the same for all approaches. For sample selection, we established the number of samples per iteration as 2 times the number of classes. The results reported in the Section 4.3 were compiled from the average of experiments run 10 times, with randomly generated sets of samples for the learning and test sets, for accuracy measures. For all datasets used, we chose 80% of the available samples for learning and 20% for testing.

4.2 Datasets

The experiments were conducted on real-world datasets from very diverse domains. The first four (Faces, Statlog, Pendigits and Covtype) datasets are public. The last two (Cowhide and Parasites) ones are proprietary and were obtained from real applications. Table 4.1 presents details (number of samples, attributes, and classes) of the datasets used in the experiments.

- Faces: this dataset, obtained from University of Notre Dame [23], was originally designed to study the effect of time on face recognition. The images were acquired in several sessions weekly with the participation of distinct individuals. In these sessions, different expressions (neutral, smiling, sad) were captured. Figure 4.1 displays samples from this dataset.

- Statlog: this is the Landsat Satellite dataset obtained from the UCI Machine Learning Repository [26], which consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image and the classification associated with the central pixel in each neighborhood;

Table 4.1: Number of samples, attributes and classes of the datasets

Dataset	**Samples**	**Attributes**	**Classes**
Faces	1,864	162	54
Statlog	2,310	19	7
Pendigits	10,992	16	10
Covtype	581,012	54	7
Cowhide	1,690	160	5
Parasites	1,660	262	15

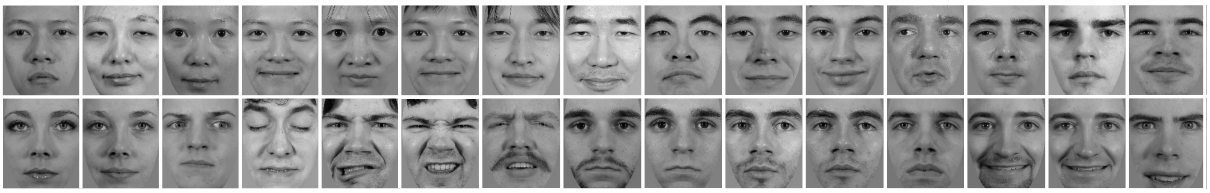


Figure 4.1: Examples of images from the Faces dataset.

- Pendigits: this is the Pen-Based Recognition of Handwritten Digit dataset obtained from the UCI Machine Learning Repository [25], that consists of 10,992 samples in 16 dimensions, distributed in 10 classes corresponding to the digits [0...9]. The 16 dimensions are drawn by re-sampling from handwritten digits. This digits database was built from a collection of 250 samples from 44 writers.

- Covtype: this is the Covertype dataset obtained from the UCI Machine Learning Repository [21], that consists of 581,012 samples, 7 classes, and 54 features. This is a very heterogeneous set (see Table 4.2).

- Cowhide: this is a proprietary dataset [22] obtained from a real application for the classification of defects in cowhide. The main reason for selecting samples of cowhide defects is the great complexity of their evaluation, especially in areas close to the vicinity of different defects; in addition, it is a dataset of current use. Five types of regions of interest in the Wet-Blue[1] processing stage were selected, namely, scabies, ticks, hot-iron, cut, and regions without defect (Figure 4.2).

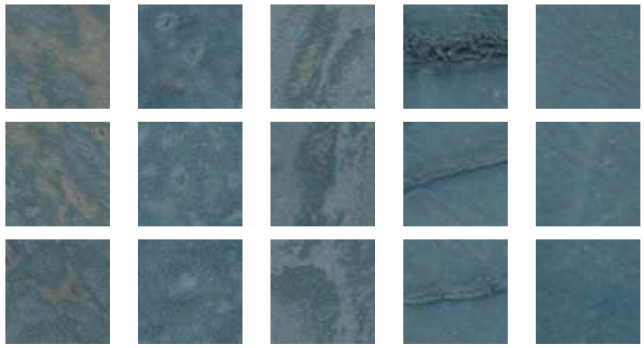
- Parasites: this is a proprietary dataset [24] composed of images of parasites, provided by a research laboratory at the University of Campinas, where fecal parasitological examination is performed for diagnosis of enteroparasitoses present in humans. A particularity of this set is that each class contains a different number of images

[1]Wet-Blue leather is an intermediate stage between untanned and finished leather.

varying from 33 to 163 depending on the parasite species found on microscope slides. Figure 4.3 displays specimens from this dataset.

Table 4.2: Class and number of samples in each class.

Class	Samples
1	211,840
2	283,301
3	35,754
4	2,747
5	9,493
6	17,367
7	20,510



(a) (b) (c) (d) (e)

Figure 4.2: Images of a defect in wet-blue. (a) Scabies, (b) Tick, (c) Hot-iron, (d) Cut, (e) without defect.

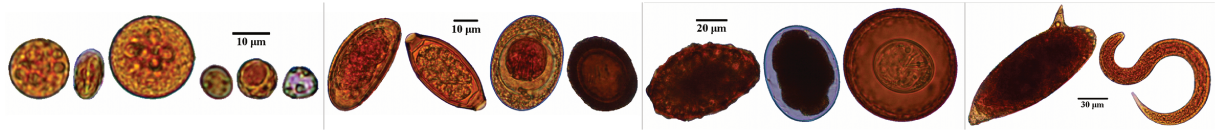


Figure 4.3: Examples of images from each class of the structures of intestinal parasites in the Parasites dataset.

4.2.1 Datasets for the Diagnosis of Parasites

The automated system for the diagnosis of parasites [87, 88] consists of a parasitological technique to process fecal samples and produce suitable microscopy slides; a customized

motorized microscope with digital camera to acquire images from the slides; and computer methods for microscope and camera control, image segmentation, feature extraction, and sample recognition. We used this system to automatically acquire images from microscopy slides and form three datasets. The first dataset (d_1) containing 1,944 parasites (without impurities) as segmented by the system and carefully labeled by an experienced expert in parasitology. The second dataset (d_2) with 5,948 samples, containing 1,944 parasites and 4,004 impurities. The third dataset (d_3) consists of 141,059 unlabeled samples. In this case, however, we used different versions of the stool processing technique, the classes are unbalanced, not all classes are present, and the proportion between impurities and parasites is much higher. Indeed, the third dataset better reflects the circumstances in a laboratory routine.

The fecal samples were obtained from several regions of the state of São Paulo: university hospitals at the University of Campinas (UNICAMP) and at the São Paulo State University (UNESP), as well as the Ouro Verde Hospital, and were processed in our laboratory at the Institute of Computing, UNICAMP.

The parasites are from the 15 most common species in Brazil (see Figure 4.4a): *Ascaris lumbricoides* eggs, *Enterobius vermicularis* eggs, Ancylostomatidae eggs, *Trichuris trichiura* eggs, *Hymenolepis nana* eggs, *Hymenolepis diminuta* eggs, *Taenia* spp. eggs, *Schistosoma mansoni* eggs, *Strongyloides stercoralis* larvae, *Entamoeba histolytica/E.dispar* cysts, *Giardia duodenalis* cysts, *Entamoeba coli* cysts, *Endolimax nana* cysts, *Iodameba bütschlii* cysts, and *Blastocystis hominis* cysts. The impurities constitute the greatest challenge, being higher in number, diverse, and also often similar to some species of parasites (see Figure 4.4b).

4.3 Results

This Section discusses the results of the experiments performed on the aforementioned datasets for each proposed strategy.

4.3.1 Cluster-OPF-Rand Method

A first instantiation, Cluster-OPF-Rand, of the reduction strategy was developed in order to illustrate its effectiveness. Cluster-OPF-Rand is based on the Optimum-Path Forest (OPF) methodology, while relying on clustering and classification for the learning processes. For evaluation, we developed a baseline approach (OPF-Rand) using the OPF classifier and random selection of training samples from the entire dataset. At each learning iteration, the same member of random samples is selected from the entire dataset for OPF-Rand, and from the reduced dataset, for the Cluster-OPF-Rand. This number of

(a) (b)

Figure 4.4: Examples of image samples in the datasets. (a) from each class of parasites. (b) from impurities.

samples is equal to the number suggested by Cluster-OPF-Rand based on the clustering results - a fair choice. These samples are classified and presented to the expert for annotation. The expert annotates the misclassified samples and they are added to the training set to improve the OPF classifiers used in each method for the next iteration. Thus, one can easily note the gain obtained by using clustering for dataset reduction, which induces the knowledge of a large number of classes, resulting in an early increase in accuracy. Moreover, clustering also allows for the choice of random samples from the reduced set comprised of good representative samples, instead of a much larger set of data (as in OPF-Rand).

Notice that the proposed method creates a new classifier instance at each iteration. We would like to verify the ability of Cluster-OPF-Rand in choosing the most representative samples from a reduced set, as well as, in which iteration, whether the expert might be pleased with the classification accuracy. Therefore, we monitor the mean accuracy of each instance on the unseen samples of the test set. Furthermore, for each sample set selected at each iteration, we simulate the expert interaction by correcting the misclassified labels given by the current classifier instance. Tables 4.3-4.5 help compare the mean accuracy and the total number of annotated images used to increase the training set.

Table 4.3: Accuracies and total annotated images for Cluster-OPF-Rand and OPF-Rand on the Faces dataset.

Faces	Accuracy (%)		Total Annotated Images (%)	
Iteration	Cluster-OPF-Rand	OPF-Rand	Cluster-OPF-Rand	OPF-Rand
1	94.85	85.11	6.51	6.51
2	97.27	94.21	7.59	8.51
3	98.06	97.35	8.11	9.40
4	98.57	98.35	8.41	9.78
5	98.85	98.78	8.68	9.98

Table 4.4: Accuracies and total annotated images for Cluster-OPF-Rand and OPF-Rand on the Parasites dataset.

Parasites	Accuracy (%)		Total Annotated Images (%)	
Iteration	Cluster-OPF-Rand	OPF-Rand	Cluster-OPF-Rand	OPF-Rand
1	92.68	79.44	1.98	1.98
2	94.12	88.50	2.54	2.66
3	94.94	91.60	2.91	3.06
4	95.30	92.67	3.12	3.29
5	95.21	93.64	3.36	3.54

In summary, Cluster-OPF-Rand enables to achieve the desired results, by using the knowledge of both expert and classifier, at each learning iteration, along with the reduction strategy developed. Experiments with datasets from distinct applications showed that Cluster-OPF-Rand attains higher accuracy sooner than those presented by the baseline OPF-Rand and random selection of training samples from the entire dataset. By reducing the learning dataset to a smaller and more representative set, Cluster-OPF-Rand also minimizes the number of learning iterations as well as the annotation effort.

Using the Faces dataset (Table 4.3), both methods achieve similar accuracies and both can be improved with more expert annotations and more learning iterations. However, Cluster-OPF-Rand allows the learning process to stop earlier in comparison with OPF-Rand. Furthermore, it is important to highlight that, out of 1,469 samples only 132.94 (about 9.05%) had to be annotated for the proposed method to achieve accuracy above 99%, in its last (9^{th}) iteration using all samples on the reduced set. These results are similar to those for the remaining datasets (Tables 4.4 and 4.5). This shows that our method can outperform OPF-Rand in effectiveness.

Considering the Parasites dataset (Table 4.4), in the first iteration, Cluster-OPF-Rand achieves accuracies above 92% with less than 2% of the learning samples annotated by

Table 4.5: Accuracies and total annotated images for Cluster-OPF-Rand and OPF-Rand on the Pendigits dataset.

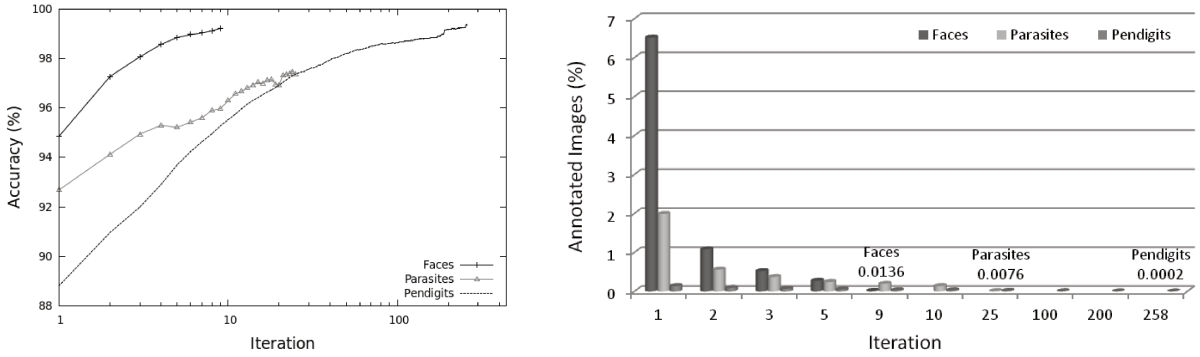
Pendigits	Accuracy (%)		Total Annotated Images (%)	
Iteration	Cluster-OPF-Rand	OPF-Rand	Cluster-OPF-Rand	OPF-Rand
1	88.80	70.36	0.13	0.13
2	90.96	82.97	0.22	0.25
3	91.99	87.49	0.29	0.30
4	92.89	89.72	0.35	0.35
5	93.70	91.25	0.40	0.40

the expert, while the randomized method OPF-Rand reaches similar accuracies only from the fourth iteration on and requiring the expert to annotate more than 3% of the learning samples. Furthermore, out of 1,323 samples only 77.7 (about 5.87%) had to be annotated for Cluster-OPF-Rand to achieve an accuracy above 97%, in its last (25^{th}) iteration using all samples in the reduced set.

For the Pendigits dataset (Table 4.5), our method obtains high accuracies in all learning iterations. In the first one, it presents an accuracy of 88.80%. In the remaining iterations, the accuracies tend to increase continuously, reaching over 99%. Furthermore, out of 8,791 samples only 79.9 (about 0.90%) had to be annotated for the proposed method to achieve accuracy above 97% in the 30^{th} iteration. In a practical situation, an expert would be very pleased at this point, mainly considering that the randomized method (OPF-Rand) learning process consists of 440 iterations, when using all available learning samples.

Figure 4.5a-b illustrates the mean accuracies and the number of samples annotated by the expert at each iteration for each dataset using Cluster-OPF-Rand, respectively. We used logarithmic scales, due to the size of these datasets. Our method requires a greater effort by the expert in the first few iterations, since the selected samples are the most difficult to classify. However, looking at the end of the learning process, one can observe that the proposed method demands less effort from the expert, who annotates much fewer samples after some iterations (reaching almost no annotations at all).

The reduction strategy becomes very important in a process where a goal is to limit the number of iterations to as few as possible. In this context, selecting samples that speed up the improvement of the classifier through the iterations becomes critical. The more difficult to classify the selected samples in the current iteration are, the more useful they are to improve the classifier for the next iteration. Therefore, the selection of hard to classify samples coupled with the early knowledge of all classes allow for higher accuracy sooner.



(a) (b)

Figure 4.5: Comparison of Cluster-OPF-Rand on the three datasets. (a) Mean accuracy on the test sets. (b) Total annotated samples in each iteration (in percentage).

Note that, in the first iteration with all datasets (Tables 4.3-4.5), Cluster-OPF-Rand provides higher accuracies than OPF-Rand. Using roots of each cluster for the first classifier instance becomes really important due to its use in the next iteration. This reduces the time and effort by the expert who mainly has only to confirm the labels of the samples that have already been classified. Hence, this first instance of the classifier should be based on the knowledge of as many classes as possible (ideally, all of them). In later learning iterations, the performance gain depends on the choice of good samples. With the proposed method, it is possible to improve these choices by reducing a large dataset to a small subset consisting of boundary cluster samples for the training of the subsequent classifiers.

It is clear that Cluster-OPF-Rand, in addition to providing high accuracies, requires fewer learning iterations than those demanded by OPF-Rand. Additionally, it relies on fewer interactions with the expert whose effort is reduced to almost none after a few iterations. Therefore, clustering improves the knowledge of samples from most/all classes. From the results presented, we can see that clustering roots allow us to obtain high accuracy since the first iteration. In the remaining iterations, the growth of accuracy is faster for Cluster-OPF-Rand, which also proves beneficial for the reduction strategy proposed.

4.3.2 DBE Method

In this Section, to validate the DBE strategy, we compared its performance against the baseline approach (Rand) based on the OPF classifier and random selection of samples. We also compared the OPF and $k$means clustering techniques, denoting the DBE methods as OPF_DBE and $k$means_DBE. Similarly, alternative supervised classifiers can also be

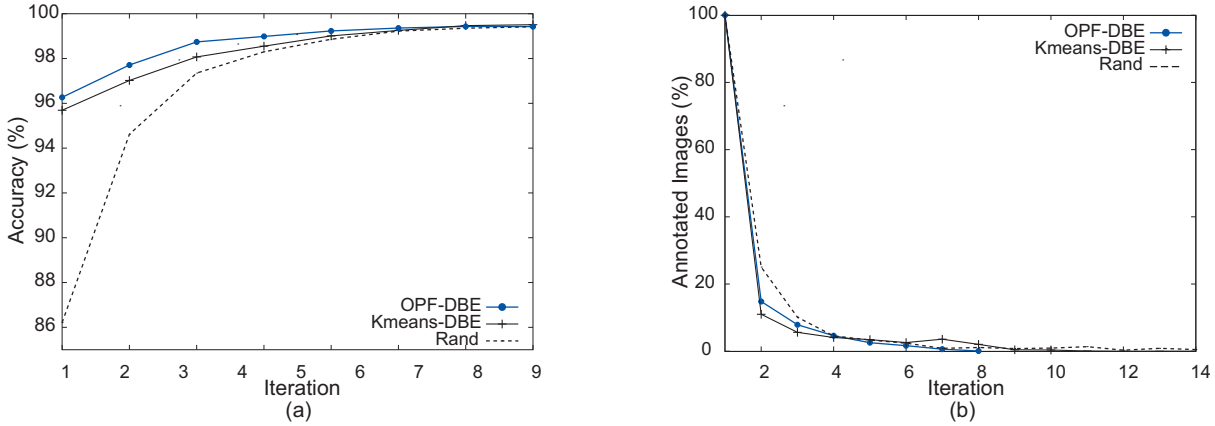


Figure 4.6: Comparison for DBE strategy on the Faces dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

considered in the classification and selection processes of the learning process. We chose OPF-based classification since it offers considerable advantages and has been used very successfully for different applications [68].

To compare the effectiveness of each method, we considered the accuracy measured (on an unseen test set obtained from each dataset) throughout the learning iterations as well as the percentage of annotated images in each iteration. Figures 4.6 and 4.7 show these results for the datasets Faces and Parasites, respectively.

Both DBE methods started off with a better performance than the Rand method, for all datasets analyzed. Moreover, the DBE ones achieved high accuracies earlier (see Figures 4.6a and 4.7a). To reach the same accuracies, the Rand method required more learning iterations than the DBE ones. Regarding the number of annotated images, the DBE methods required more effort from the expert in the beginning, but after a few iterations DBE did not involve much work (see Figures 4.6b and 4.7b).

Considering the Faces dataset, as early as on the third iteration, DBE reached accuracies above 98% using both OPF and $k$means clustering, while the randomized method required two extra iterations to achieve similar accuracies. One could safely assume that an expert would be pleased at this point, in any practical situation. One can also observe that under DBE methods, from the first iteration onward, the number of images annotated by the expert decreased drastically, reaching 0% by the end of their learning cycle. The Rand method required more interactions with the expert for the annotation of mislabeled samples for a large number of iterations, namely, up until its 14^{th} and last iteration. See Figure 4.6b. The reason for this behavior is the absence of any strategy for reduction or prioritization of samples and simply the reliance on randomness, on a large dataset. Regarding the DBE methods, using OPF and $k$means clustering in the reduction

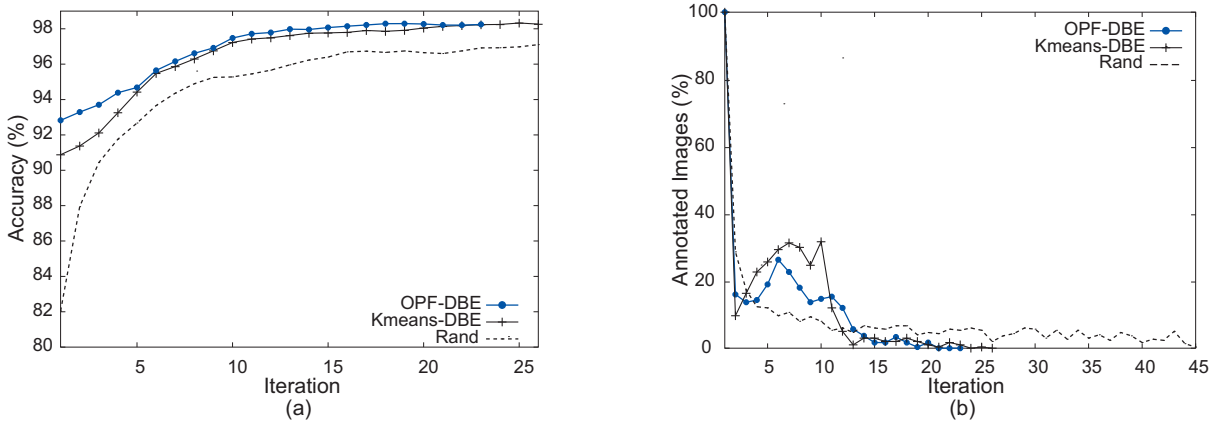


Figure 4.7: Comparison for DBE strategy on the Parasites dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

process, the OPF-DBE reaches better performance (more quickly) than the *k*means-DBE. Moreover, out of 1,469 learning samples, only about 10% had to be annotated by the expert for OPF-DBE (Table 4.6) in order to achieve an accuracy above 99%, on its last iteration after processing all samples in the reduced dataset.

Overall, these results are similar to those observed for the Parasites dataset (Figure 4.7). While DBE methods behave with a broadly superior performance, for the Rand method the learning process is fairly slow in reaching equally high accuracies and requires more iterations. This shows that the proposed strategy outperforms Rand in effectiveness. On the 10^{th} iteration, DBE methods reach accuracies above 97%, while Rand can achieve similar accuracies only after the 26^{th} iteration. It is worth noting that while the reduction strategy can decrease by up to fifty percent the number of samples available for learning (dropping the number of required iterations from 45 to 23 for OPF-DBE and from 45 to 26 for *k*means-DBE, see Figure 4.7b), the reduced set is comprised mostly of relevant samples since accuracies are high and the number of images annotated using DBE methods are smaller than with Rand. OPF-DBE reached an accuracy above 98%, with the annotation of only 6.99% of the available 1,323 samples by its last iteration. See Table 4.6. However, the Rand method required interactions with the expert (and annotations) until its last (45^{th}) learning iteration (Figure 4.7b).

Observe that the annotation of a smaller number of samples was not always achieved by DBE strategy since, in the first few iterations (Figure 4.7b), it selects more difficult samples to be classified. It is important to highlight that the harder to classify the selected samples are the more useful they will be to improve the classifier for the next iteration. Selecting images that speed up the improvement of the classifier throughout the iterations is essential. DBE enables one to obtain a better classifier with a very small number of

Table 4.6: Total size of the learning set $|\mathcal{Z}_2|$, total number of annotated images, accuracies (in percentage) and computational time gains on two datasets for OPF-DBE.

datasets	$\lvert\mathcal{Z}_2\rvert$	annotated	accuracy	time gain
Faces	1,469	10.08%	99.42%	9.85×
Parasites	1,323	6.99%	98.24%	14.63×

iterations and interactions due to the reduction, organization and selection strategies. This is particularly useful in a process where a goal is to limit the number of iterations to as few as possible.

Note that all methods will gradually improve when more expert annotations and more learning iterations are allowed. However, DBE enables the learning process to stop earlier in comparison with the randomized method. It is well known that on an actual field environment, a large number of learning iterations is tiresome and furthers human errors in the annotation process which, consequently, affects the quality of the classifier. In our experiments, we discarded the possibility of misannotations by the expert, which, by the way, would increase the number of images annotated by the expert on the Rand method, due to its intensive interaction through many late learning iterations.

Let us present evidence that both proposed strategies inherent to DBE have specific and complementary roles. The reduction strategy through clustering improves the knowledge of samples from most/all classes ever since the first iteration, as well as refines the larger learning set by decreasing it into a smaller relevant set. Observe on all graphs (Figures 4.6a and 4.7a), that the first iteration of DBE provides higher accuracies than Rand. Therefore, the roots of the clusters caused a very positive impact on the performance of the first classifier. As consequence, it improves selection of new samples and the performance gain of DBE over Rand continues along the subsequent iterations.

The organization strategy of prioritizing samples based on sorting criteria enables the choice of more useful samples from the reduced set and consequently the classifier learns more quickly. The growth of accuracy being faster for DBE than for Rand shows the benefits of the proposed sorting strategy. Hence, the reduction and sorting strategy allows for a better selection from a small relevant subset.

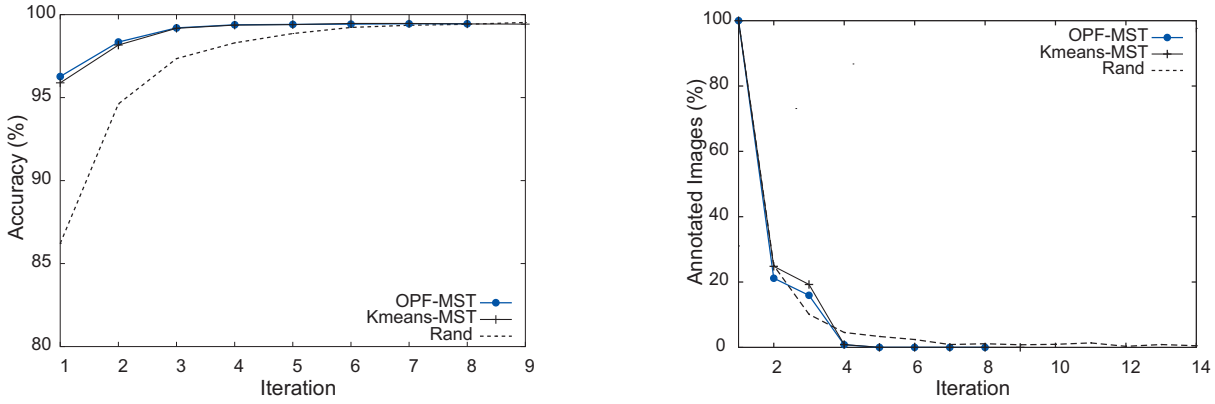
Although the OPF classifier is many times faster than other popular methods, such as SVM and ANN-MLP, the classification time on $\mathcal{Z}_2'$ amounted to about 1/10 and 1/15 of the classification time on $\mathcal{Z}_2$ for the Faces and Parasites datasets, respectively. This gives us an idea of the efficiency gain of OPF-DBE with respect to other active learning approaches that re-classify the entire $\mathcal{Z}_2$ at each iteration of the learning process. Table 4.6 presents this gain, as well as mean accuracies, total annotated images and the total size of the learning set $\mathcal{Z}_2$ on each dataset for OPF-DBE.

4.3.3 MST-BE Method

In this Section, we evaluated the proposed strategy (MST-BE) using the traditional random selection process as baseline. To demonstrate the effectiveness of the proposed strategy using different clustering techniques, we performed comparisons using OPF and *k*means algorithms for the reduction process. We denoted the MST-BE methods as OPF-MST and *k*means-MST. In this work, we also compared the performance of our strategy using SVM and OPF classifiers. The MST-BE methods were denoted as MST-OPF and MST-SVM. To compare the effectiveness of each method, we considered the accuracy measured (on an unseen test set obtained from each dataset) throughout the learning iterations as well as the percentage of annotated images in each iteration.

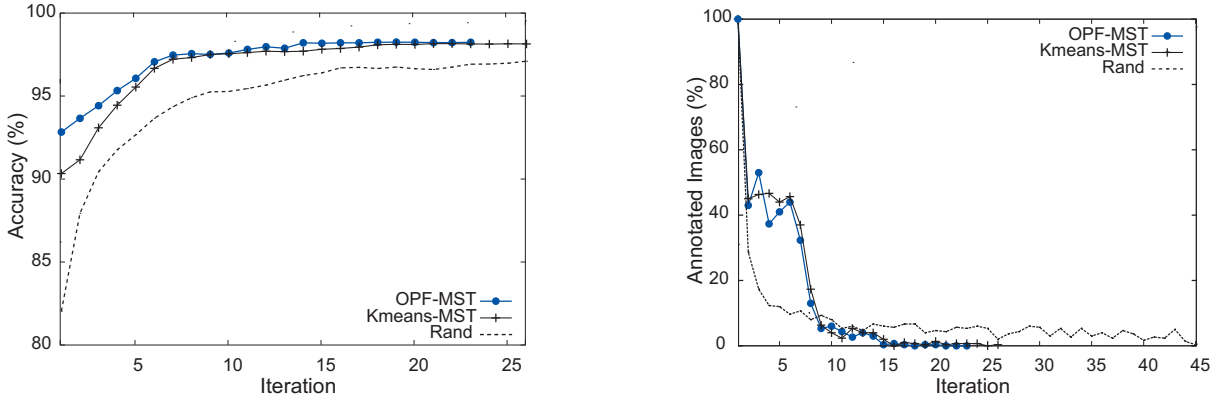
Figures 4.8, 4.9 and 4.10 show the results using OPF-MST, *k*means-MST and Rand for the datasets Faces, Parasites and Pendigits, respectively. As previously mentioned, alternative supervised classifiers may be considered in the classification and selection processes. Initially, all results have been generated by using the OPF classifier, since it has demonstrated considerable advantages in effectiveness and efficiency. Both MST-BE methods (OPF-MST and *k*means-MST) started off with a better performance than the Rand method, for all datasets analyzed. Moreover, the MST-BE ones achieved high accuracies earlier (see Figures 4.8a, 4.9a and 4.10a). To reach the same accuracies, the Rand method required more learning iterations than the MST-BE ones. Regarding the number of annotated images, the MST-BE methods required more effort from the expert in the beginning, but after a few iterations MST-BE did not involve much work (see Figures 4.8b, 4.9b and 4.10b).

Considering the Faces dataset, as early as on the third iteration, MST-BE reached accuracies above 99% using both OPF and *k*means clustering, while the randomized method required five extra iterations to achieve similar accuracies. One can also observe that under MST-BE methods, from the first iteration onward, the number of images annotated by the expert decreased drastically, reaching 0% by the end of their learning cycle. The Rand method required more interactions with the expert for the annotation of mislabeled samples for a large number of iterations, namely, up until its 14^{th} and last iteration (see Figure 4.8b). The reason for this behavior is the absence of any strategy for reduction or prioritization of samples and simply the reliance on randomness, on a large dataset. Regarding the MST-BE methods, using OPF and *k*means clustering in the reduction process, the OPF-MST reaches better performance (more quickly) than the *k*means-MST. Moreover, out of 1,469 learning samples, only about 10.50% had to be annotated by the expert for OPF-MST (Table 4.7) in order to achieve an accuracy above 99%, on its last (8^{th}) iteration after processing all samples in the reduced dataset. Furthermore, OPF-MST presented a speed up of about 9.31 times in classification time over the traditional method where the entire learning set is classified at each iteration.



(a) (b)

Figure 4.8: Comparison for MST-BE strategy using the OPF classifier on the Faces dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.



(a) (b)

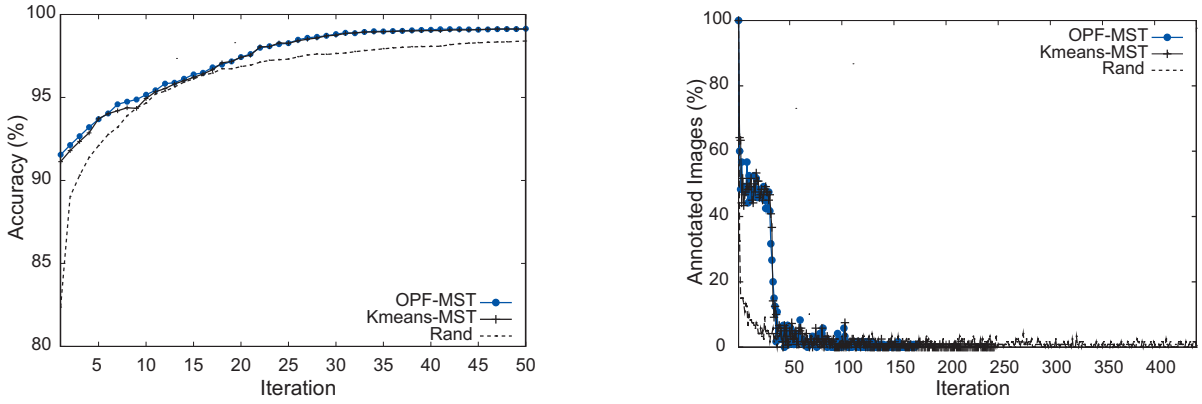
Figure 4.9: Comparison for MST-BE strategy using the OPF classifier on the Parasites dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

Overall, these results are similar to those observed for the Parasites and Pendigits datasets (Figures 4.9 and 4.10). While MST-BE methods behave with a broadly superior performance, for the Rand method the learning process is fairly slow in reaching equally high accuracies and requires more iterations. This shows that the proposed strategies outperform Rand in effectiveness.

Using the Parasites datasets, on the 6^{th} iteration, OPF-MST reaches accuracy above 97%, while *k*means-MST and Rand can achieve similar accuracies only after the 7^{th} and 26^{th} iteration, respectively. It is worth noting that while the reduction strategy can

Table 4.7: Total size of the learning set $\mathcal{Z}_2$, total number of annotated images, mean accuracies ± standard deviations and computational time gains on three datasets for OPF-MST using the OPF classifier.

datasets	$\lvert\mathcal{Z}_2\rvert$	annotated	accuracy ± std	time gain
Faces	1,469	10.08%	99.42% ± 0.29	9.31
Parasites	1,323	6.99%	98.24% ± 0.62	10.35
Pendigits	8,791	4.21%	99.42% ± 0.04	9.40



(a) (b)

Figure 4.10: Comparison for MST-BE strategy using the OPF classifier on the Pendigits dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

decrease by up to fifty percent the number of samples available for learning (dropping the number of required iterations from 45 to 23 for OPF-MST and from 45 to 26 for $k$means-MST, see Figure 4.9b), the reduced set is comprised mostly of relevant samples since accuracies are high and the expert's effort in data annotation using MST-BE methods is smaller than with Rand. OPF-MST reached an accuracy above 98%, with the annotation of only 8.91% of the available 1,323 samples by its last (23^{th}) iteration (see Table 4.7). However, the Rand method required interactions with the expert (and annotations) until its last (45^{th}) learning iteration (Figure 4.9b). In regard to computational time, OPF-MST was 10.35 times faster than the traditional method where all samples in the learning set are classified at each iteration.

With the Pendigits dataset (Figure 4.10), out of 8,791 samples only about 4% had to be annotated for the OPF-MST to achieve accuracy above 99%. Regarding to computational time, OPF-MST was about 9.40 times faster than the traditional method where the entire learning set is classified at each iteration.

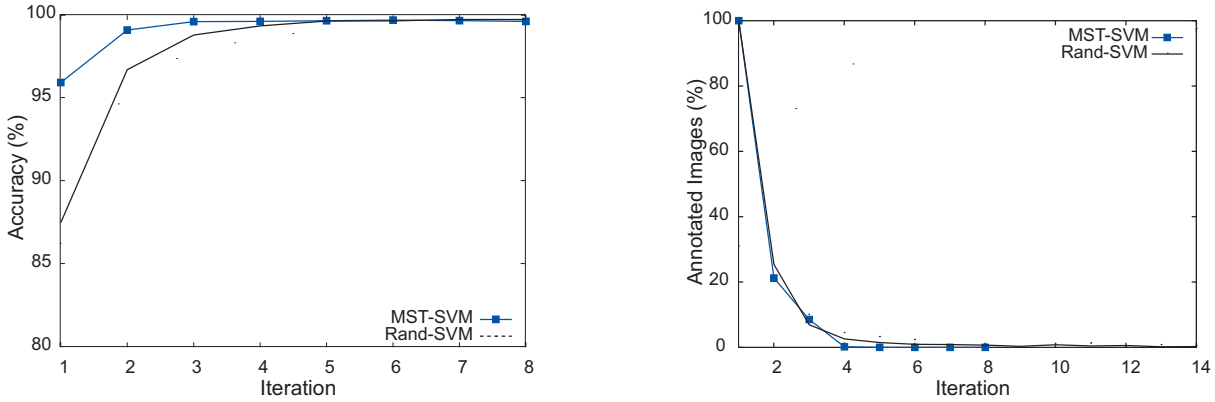
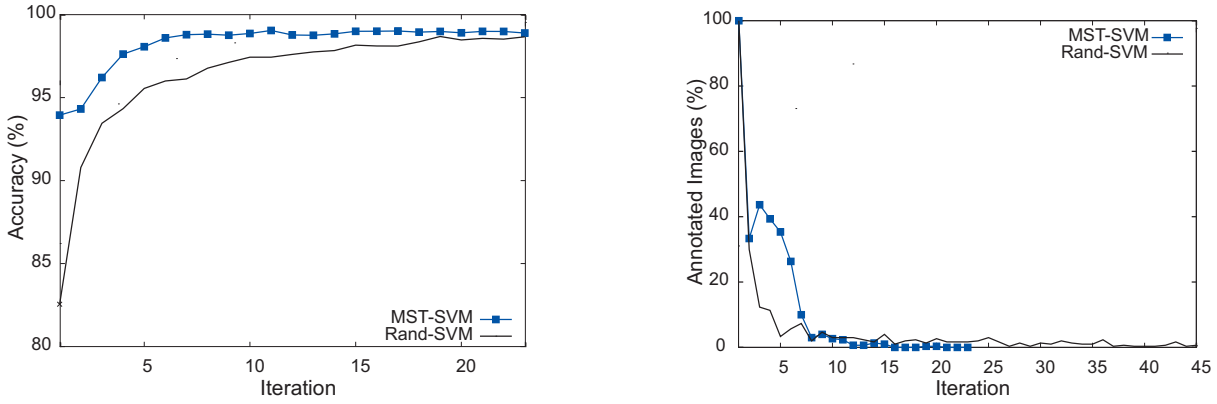


Figure 4.11: Comparison for MST-BE strategy using the OPF clustering on the Faces dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

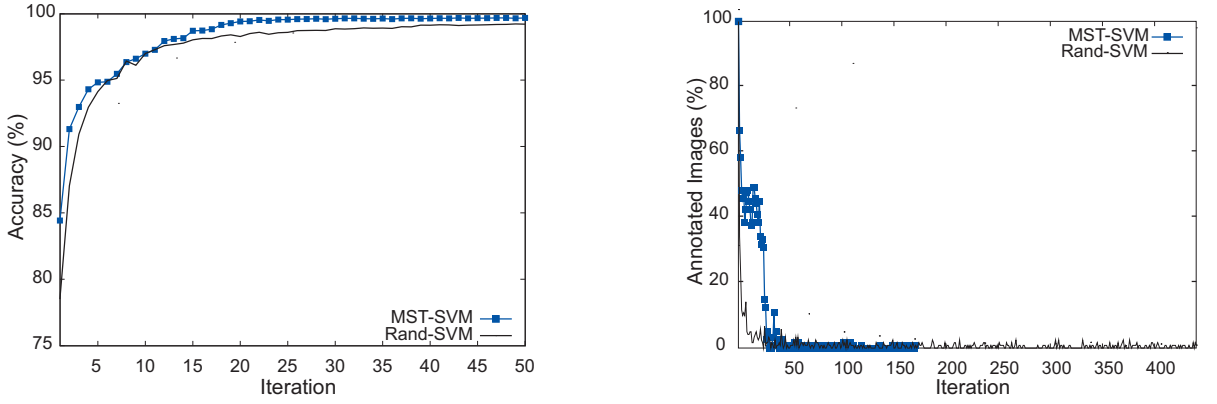
Observe that the annotation of a smaller number of samples was not always achieved by MST-BE methods since, in the first few iterations (Figures 4.8b-4.10b), it selects more difficult samples to be classified. It is also important to highlight that the harder to classify the selected samples are, the more useful they will be to improve the classifier for the next iteration. Requiring a greater effort from the expert in the beginning of the learning process is a small price to pay and corresponds to any expert's expectation. The experts are willing to spend extra effort interacting in the first few iterations to make the learning faster. It is important to emphasize that, in the proposed paradigm, the expert's time and effort are reduced to none after just a few iterations. From the expert's point of view, the expectation of annotating misclassified samples after many iterations is tiresome and induces the perception of non-convergence, as it occurs with the randomized method. Moreover, it is well known that on an actual field environment, due to the expert's fatigue, errors may be introduced in the annotation process, which, consequently, affects the quality of the classifier. In our experiments, we discarded the possibility of misannotations by the expert, which, by the way, would increase the number of images annotated on the Rand method.

To verify the effectiveness of the proposed paradigm using different classifiers, we also performed comparisons between MST-BE and Rand methods using SVM-based classification. Figures 4.11, 4.12 and 4.13 present the graphs for MST-SVM and Rand-SVM using the Faces, Parasites and Pendigits datasets, respectively. In this case, the OPF-based clustering was used in the reduction process, due to the best results presented over the $k$means clustering. In addition to achieving higher accuracies sooner (Figures 4.11a, 4.12a and 4.13a), MST-SVM requires fewer learning iterations than those presented by



(a) (b)

Figure 4.12: Comparison for MST-BE strategy using the OPF clustering on the Parasites dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.



(a) (b)

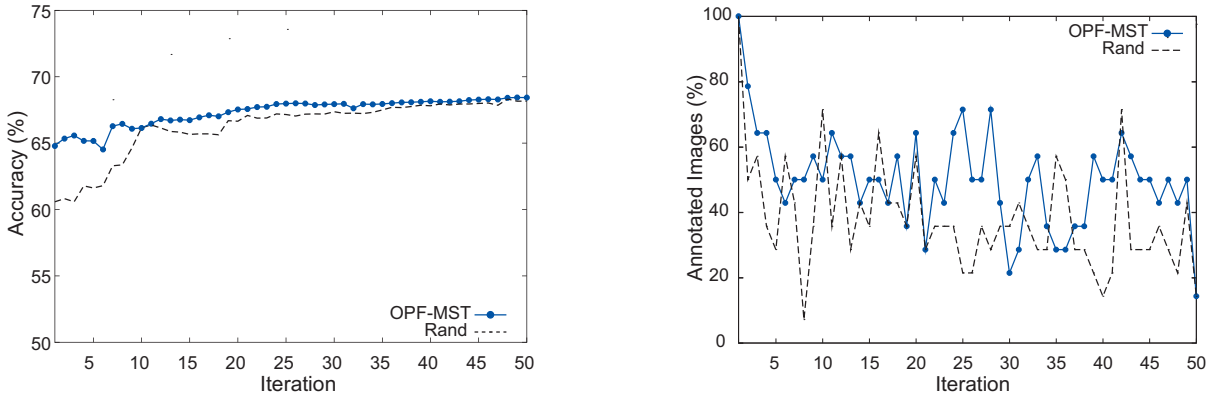
Figure 4.13: Comparison for MST-BE strategy using the OPF clustering on the Pendigits dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

Rand-SVM. In MST-SVM, after a few iterations, the expert no longer needs to correct the label of any sample (see Figures 4.11b, 4.12b and 4.13b). As we can see (Table 4.8), the MST-BE methods using both OPF and SVM classifiers obtained similar accuracies. However, when we compared their classification time on the reduced learning set $\mathcal{Z}'_2$, OPF classifier presents a gain of 8.30, 3.81 and 5.52 times for Faces, Parasites and Pendigits, respectively.

In order to demonstrate the applicability and effectiveness of the proposed method in a practical situation (with thousand of samples), we also performed experiments with

Table 4.8: Mean accuracies ± standard deviations in the third iteration evaluated on three datasets for MST-BE using OPF clustering in the reduction process and OPF and SVM classifiers in the classification process.

datasets	OPF	SVM
Faces	98.35% ± 0.44	99.11% ± 0.59
Parasites	93.66% ± 0.64	94.31% ± 0.74
Pendigits	92.13% ± 0.81	91.31% ± 1.40



(a) (b)

Figure 4.14: Comparison for MST-BE strategy using the OPF classifier on the Covtype dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

the Covtype dataset (Figure 4.14), a larger and more challenging dataset, given its size and high heterogeneity, i.e., its great variability in the number of samples in each class (see Table 4.2). Hence, for this dataset, we present only the results using the faster technique (OPF) for both reduction and classification processes. The results were similar to those observed previously. It is worth noting that the proposed reduction strategy performed a significant downsizing of the learning set (over ninety percent). Moreover, out of 464,807 samples only less than 1% had to be annotated for the OPF-MST to achieve accuracy above 75%. Notice that all methods will gradually improve when more expert annotations and more learning iterations are allowed. Selecting images that speed up the improvement of the classifier throughout the iterations is essential. MST-BE enables one to obtain a better classifier with a very small number of iterations and interactions due to the reduction, organization and selection strategies. This is particularly useful in a process where a goal is to limit the number of learning iterations to as few as possible. Regarding to computational time, OPF-MST was about 24.68 times faster than the traditional method

where the entire learning set is classified at each iteration.

4.3.4 RDS Method

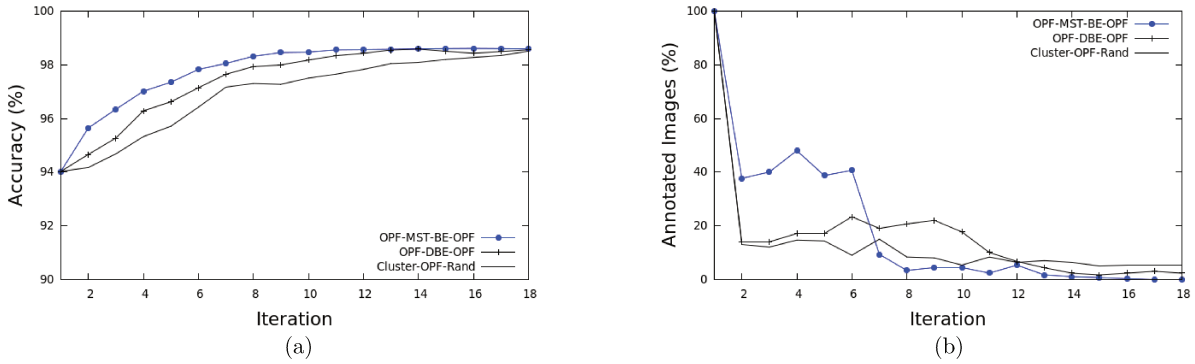
In this Section, initially, we present a comparison between the proposed methods (Cluster-OPF-Rand, DBE, and MST-BE). To compare the effectiveness of each method, we considered the accuracy measured (on an unseen test set obtained from the dataset) throughout the learning iterations as well as the percentage of annotated images in each iteration, using the first (d_1) dataset (described in Section 4.2.1) without impurities. Figure 4.15 shows that these methods were advances throughout this research towards an automatic classification of human intestinal parasites.

Therefore, RDS was validated against two state-of-the-art methods and evaluated with distinct clustering and classification techniques. Its version with both techniques based on optimum-path forest presented the best result. We have also evaluated this best method in a realistic scenario, which better reflects the situation in laboratory routine, using a dataset with over 140,000 unlabeled samples with unbalanced classes, not all classes, and a lot of impurities. In this case, the expert participated of the active learning process which involved label verification of only a small portion of the dataset.

We used the first dataset (d_1) to compare the RDS's performance (accuracy on an unseen test set) against two baseline active learning methods: Al-SVM [95], which selects samples from the entire learning set at each iteration using an SVM classifier, and the most competitive one, MST-BE [82], which also uses clustering to reduce and organize the learning set a priori, and interrupts sample selection when the classifier detects the desired number of samples per iteration. We also compared RDS with a random method in which samples were randomly selected from the entire dataset. For clustering, we evaluated the OPF and $k$means techniques. For classification, in the selection process, we used the SVM and OPF classifiers.

In order to facilitate the comparison among methods, when applicable, they were labeled as a triple, consisting of the clustering method, active learning method, and classification method, separated by an underscore character. The methods were denoted as Kmeans_RDS_OPF, OPF_RDS_OPF, Kmeans_RDS_SVM, OPF_RDS_SVM, Kmeans_MST-BE_OPF, OPF_MST-BE_OPF, Kmeans_MST-BE_SVM, OPF_MST-BE_SVM, and Al-SVM.

For sample selection, we established the number of samples per iteration as 2 times the number of classes. First, we evaluated the methods using two versions of the labeled dataset, with and without impurities. This is important to evaluate the robustness of the methods with respect to the presence of such diverse class. The unlabeled dataset with 141,059 samples was used only to validate the best method by the parasitology specialist.



(a) (b)

Figure 4.15: Comparison between Cluster-OPF-Rand, DBE, and MST-BE methods using the OPF methodology (clustering and classification techniques) on the d_1 dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

Table 4.9: Mean accuracies ± standard deviations of the methods on the Parasites dataset (d_1) without impurities for the 5^{th} iteration.

methods	**OPF_ MST-BE_ OPF**	**OPF_ MST-BE_ SVM**	**Kmeans_ MST-BE_ OPF**	**Kmeans_ MST-BE_ SVM**	**OPF_ RDS_ OPF**	**OPF_ RDS_ SVM**	**Kmeans_ RDS_ OPF**	**Kmeans_ RDS_ SVM**	**Al- SVM**	**Rand_ OPF**
accs	97.35%	98.39%	96.32%	97.41%	96.99%	97.97%	95.64%	96.06%	86.95%	82.50%
std dev	±0.64	±0.33	±0.74	±0.55	±0.99	±0.47	±1.29	±1.96	±2.01	±2.13

Tables 4.9-4.10 show the results of the comparisons among the methods for the case of the first dataset (d_1) without impurities in the 5^{th} iteration and the case (dataset d_2) with impurities in the 10^{th} iteration, respectively. The active learning methods (RDS and MST-BE) are superior to Al-SVM and Rand_OPF methods, for both OPF and $k$means clustering techniques (used for data organization), as well as for both OPF and SVM classifiers (used for training sample selection).

In the absence of impurities, RDS and MST-BE are equivalent, when the same clustering technique is used. However, comparing the clustering techniques, the methods using OPF clustering are more accurate than those based on $k$means. For comparison between the pattern classifiers, in the absence of impurities, RDS methods using OPF and SVM classifiers are equivalent when the clustering technique is fixed.

Considering the scenario with impurities, similarly to what was observed without impurities, RDS and MST-BE achieved high accuracies earlier in comparison with Al-SVM and Rand_OPF methods. The methods using OPF clustering also presented better results. However, in the presence of impurities, RDS outperformed MST-BE. For instance, as early as on the fourth iteration, RDS reached accuracies above 90% using both OPF

Table 4.10: Mean accuracies ± standard deviations of the methods on the Parasites dataset (d_2) with impurities for the 10^{th} iteration.

methods	**OPF_ MST-BE_ OPF**	**OPF_ MST-BE_ SVM**	**Kmeans_ MST-BE_ OPF**	**Kmeans_ MST-BE_ SVM**	**OPF_ RDS_ OPF**	**OPF_ RDS_ SVM**	**Kmeans_ RDS_ OPF**	**Kmeans_ RDS_ SVM**	**Al- SVM**	**Rand_ OPF**
accs	89.18%	85.96%	83.19%	81.40%	91.58%	90.27%	87.86%	84.90%	77.93%	74.07%
std dev	1.18±	1.72±	1.51±	1.83±	0.90±	1.79±	1.50±	1.53±	1.61±	2.10±

clustering and classification methods, while MST-BE required over ten extra iterations to achieve similar accuracies. For the Al-SVM method, the learning process was fairly slow in reaching equally high accuracies, and required over fifty iterations. Moreover, considering only the classification times on the reduced learning set of the proposed method, rather than on the entire learning set of the traditional methods, RDS provided efficiency gains of 13 times on the tested datasets.

Comparing RDS method with its variant classification methods, we can say that the OPF and SVM classifiers presented different performance. OPF classifier showed a more stable behavior. Besides, its learning times in the classification and selection processes are up to 5 times shorter than SVM learning times. Moreover, if the training set grows large (e.g., 7,000 samples), the response time is considerably affected, due to the time to retrain the OPF classifier, but it is still interactive time. Given that the training time for SVM is much higher due to its parameter optimization, the response time would be impractical. Therefore, the importance of investigating fast training algorithms for the existing classifiers is paramount.

In general, the RDS method (using both OPF clustering and classifier) had the best performance (achieving higher accuracies and decreasing the number of annotated images earlier, as well as presenting shorter learning times) in the presence of impurities. Therefore, we selected OPF_RDS_OPF method for evaluation by the specialist on the chosen realistic dataset d_3. Table 4.11 presents the total size of the learning set $\mathcal{Z}_2$, total annotated images, mean accuracies with standard deviations and computation time in seconds for selection and classification on each dataset for OPF_RDS_OPF.

In this case, an expert interface was developed to allow sample verification (manual annotation first and subsequently label correction/confirmation) by a parasitologist (see Figure 4.16). In order to fit the images of the returned samples in a single display for verification, we set the system to select 24 samples per iteration. The expert could also remove samples that did not have enough visual and morphological information to indicate whether it was a parasite or an impurity. In addition, a 10-fold cross-validation was calculated in the training set to predict the accuracy per iteration and guide the expert when to stop the learning process. To evaluate the final accuracy, a random subset of the remaining unlabeled samples was selected and automatically annotated by the final

Table 4.11: Total size of the learning set $\mathcal{Z}_2$, total number of annotated images, mean accuracies ± standard deviations and computational time in seconds for selection and classification on three Parasites datasets for OPF_RDS_OPF.

datasets	$\lvert\mathcal{Z}_2\rvert$	**annotated**	**accuracy ± std**	**time**
d_1	1,455	4.47%	96.99% ± 0.99	0.18
d_2	4,458	2.94%	91.58% ± 0.90	0.27
d_3	141,059	6.9%	88.00%	52.03

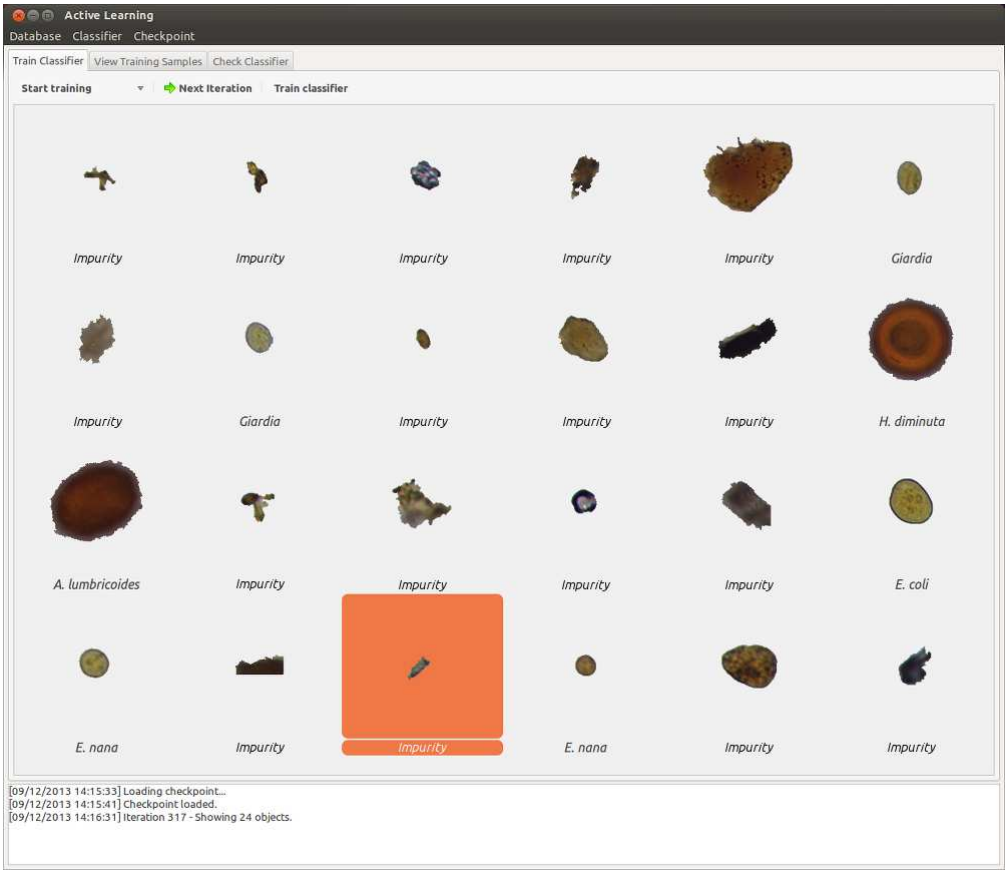
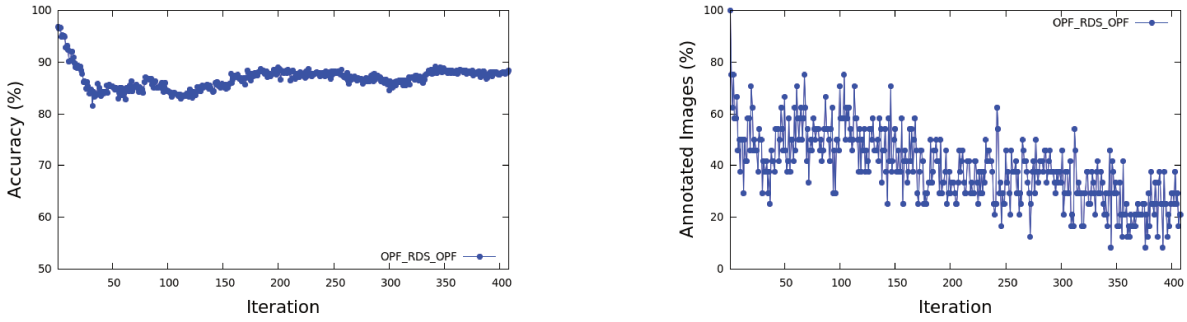


Figure 4.16: Expert interface of software used by the parasitologist to verify the label of the selected samples.

classifier. These samples were evaluated by the expert, who indicated the classification errors to compute the final accuracy.

Figure 4.17 shows the 10-fold cross-validation average accuracy and the percentage of annotated images in each iteration. We can see that the predicted accuracy started high and decreased when new species were detected by the expert, until it stabilized within a



(a) (b)

Figure 4.17: Results of the practical experiment performed by the parasitologist using OPF_RDS_OPF on the Parasites dataset with impurities. (a) Mean accuracy of the method on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

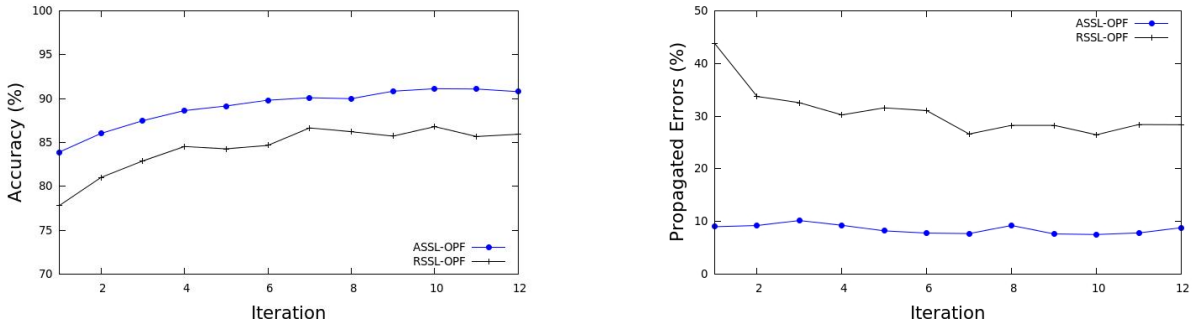
small range. After 408 iterations, the expert verified 9,792 samples (6.9% of the dataset) and corrected the label of 3,796 samples (38.7% of the selected samples). After removing the dubious samples, the final training set remained with 7,821 samples (5.5% of the dataset). The expert decided to stop the learning process when the mean accuracy by cross-validation on the labeled samples stabilized between 87% and 88%.

In order to evaluate the real accuracy of the final classifier, we created a random subset with 6% of the remaining unlabeled samples (7,870 samples). The classifier achieved 87.2% of accuracy on the random subset, matching the accuracy predicted during the training process. This is an impressive result, specially when we take into account that the expert verified only 6.9% of 141,059 samples. Furthermore, considering the low sensitivity rates from the traditional diagnosis procedure, based on visual analysis (48.3% up to 75.9%) [40], we may conclude that our solution is very relevant for the area of clinical parasitology.

4.3.5 ASSL-OPF Method

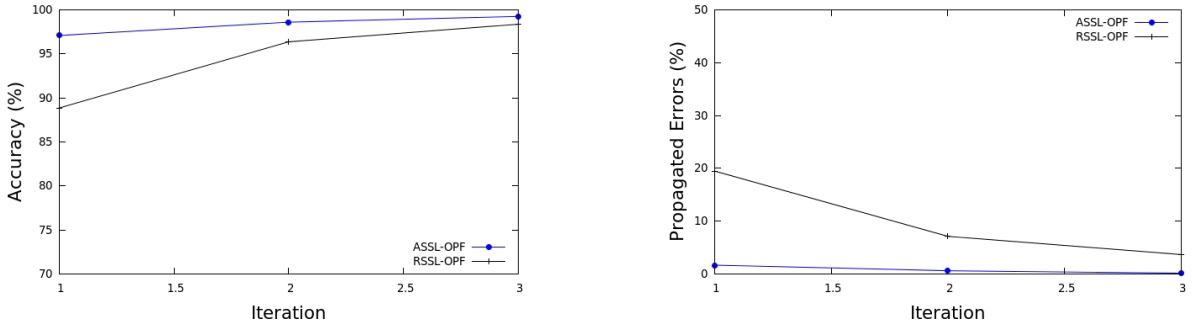
A first instantiation, ASSL-OPF, of the active semi-supervised learning strategy was developed in order to illustrate its effectiveness. ASSL-OPF is based on the OPF methodology, while relying on clustering and classification for the learning process. We compared ASSL-OPF with RSSL-OPF, a semi-supervised learning method [1] in which the labeled and unlabeled samples were randomly selected. To compare the effectiveness of each method, throughout the learning iterations, we considered the accuracy measured (on an unseen test set obtained from each dataset), the percentage of propagated errors on the unlabeled set, as well as the number of known classes.

Figures 4.18-4.22 show the results using ASSL-OPF and RSSL-OPF for the datasets



(a) (b)

Figure 4.18: Comparison for ASSL-OPF strategy on the Statlog dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Propagated errors, as a percentage of the unlabeled samples selected.

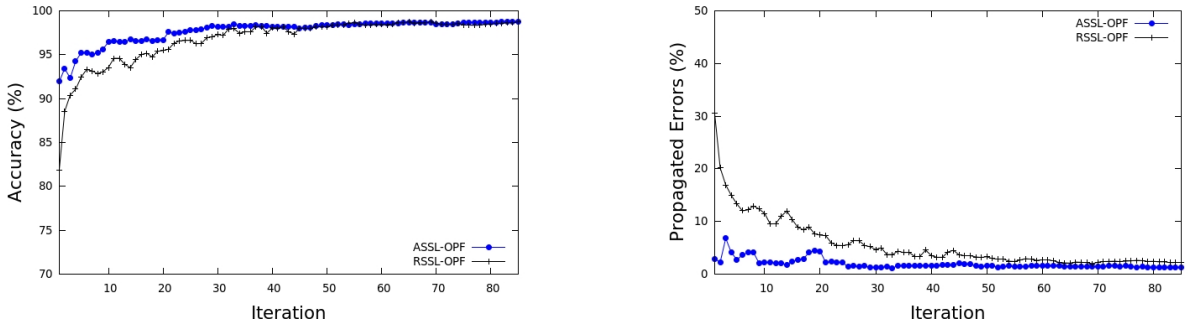


(a) (b)

Figure 4.19: Comparison for ASSL-OPF strategy on the Faces dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Propagated errors, as a percentage of the unlabeled samples selected.

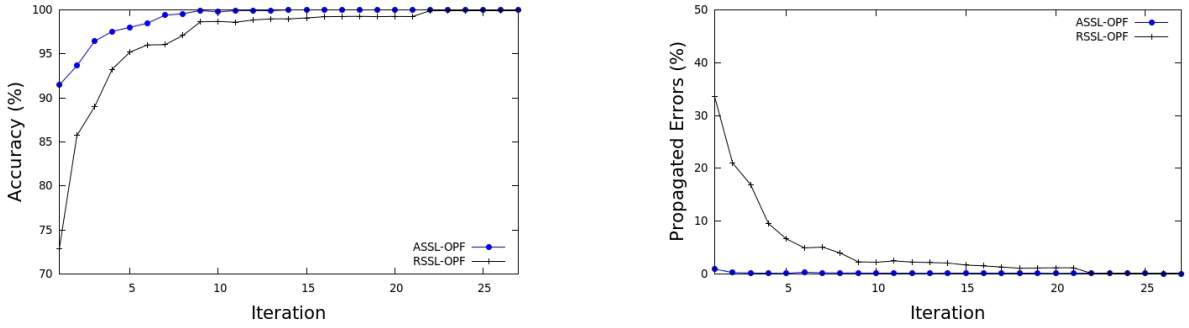
Statlog, Faces, Pendigits, Cowhide and Parasites, respectively. As we can observe, the ASSL-OPF approach had a better performance than the RSSL-OPF approach, for all datasets analyzed. ASSL-OPF achieved high accuracies earlier (see Figures 4.18a, 4.19a, 4.20a, 4.21a and 4.22a). Moreover, the ASSL-OPF approach presented fewer propagated errors on unlabeled samples (4.18b, 4.19b, 4.20b, 4.21b and 4.22b).

Considering the Statlog dataset, as early as on the second iteration, the ASSL-OPF reached accuracies over 85%, while the randomized approach achieved similar accuracies only on the sixth iteration (Figure 4.18a). Besides, the RSSL-OPF approach propagated many more errors on the unlabeled set, namely, up to 40% of that set (see Figure 4.18b). The reason for this behavior is the absence of any strategy for reduction or prioritization of samples and simply the reliance on randomness, on a large dataset.



(a) (b)

Figure 4.20: Comparison for ASSL-OPF strategy on the Pendigits dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Propagated errors, as a percentage of the unlabeled samples selected.

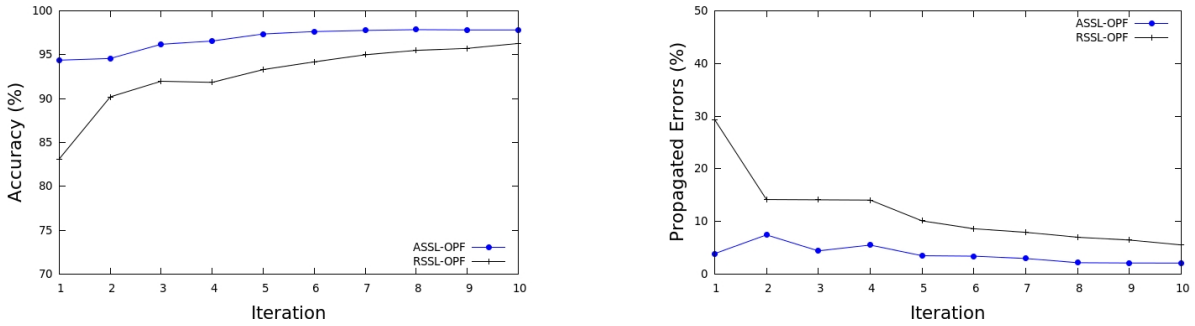


(a) (b)

Figure 4.21: Comparison for ASSL-OPF strategy on the Cowhide dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Propagated errors, as a percentage of the unlabeled samples selected.

Overall, these results are similar to those observed on the Faces, Pendigits, Cowhide and Parasites datasets (Figures 4.19, 4.20, 4.21 and 4.22). While the ASSL-OPF approach behave with a broadly superior performance, for the RSSL-OPF approach the learning process is fairly slow in reaching equally high accuracies and requires more iterations to identify samples from all classes (see Table 4.12). This shows that the proposed approach outperform RSSL-OPF in effectiveness.

In order to appreciate the quality of the results obtained on each dataset by ASSL-OPF, we also present the total number of samples annotated/corrected by the expert, mean accuracies with standard deviations and computational time for selecting the most representative samples at the end of the learning cycle (Table 4.13). This highlights the superior efficacy and efficiency advantage of the ASSL-OPF on practical applications.



(a) (b)

Figure 4.22: Comparison for ASSL-OPF strategy on the Parasites dataset. (a) Mean accuracy of the methods on unseen test sets. (b) Propagated errors, as a percentage of the unlabeled samples selected.

Table 4.12: Total number of known classes in the first three iterations for ASSL-OPF and RSSL-OPF approaches.

	Statlog		**Faces**		**Pendigits**		**Cowhide**		**Parasites**	
Iteration	**ASSL-OPF**	**RSSL-OPF**	**ASSL-OPF**	**RSSL-OPF**	**ASSL-OPF**	**RSSL-OPF**	**ASSL-OPF**	**RSSL-OPF**	**ASSL-OPF**	**RSSL-OPF**
1	7.30	7.30	54.00	46.80	10.00	8.67	5.00	4.10	14.80	12.00
2	8.00	8.00	54.00	53.00	10.00	9.83	5.00	4.90	15.00	14.40
3	8.00	8.00	54.00	53.80	10.00	10.00	5.00	5.00	15.00	14.80

Table 4.13: Total size of the learning set $\mathcal{Z}_2$, total number of annotated samples, mean accuracies ± standard deviations and selection computational times (in minutes) for ASSL-OPF.

datasets	$\lvert\mathcal{Z}_2\rvert$	**annotated**	**accuracy ± std**	**time**
Statlog	1,761	8.74%	90.79% ± 0.91	0.14
Faces	1,469	9.12%	99.26% ± 0.25	0.13
Pendigits	8,791	4.72%	98.76% ± 0.48	25.22
Cowhide	1,351	2.27%	99.66% ± 0.10	0.56
Parasites	1,455	6.45%	97.82% ± 1.69	0.85

Our approach increases the probability of the most representative samples being selected earlier and allows for a more efficient and effective training of the classifiers, leading to a considerable reduction in classifier errors after just a few iterations. This amounts to a major advantage of our approach, since it requires very few iterations in the learning

process to reach a high accuracy, while decreasing the propagated errors on the unlabeled set. Moreover, unlike the traditional active learning approaches, once the learning set has been organized, ASSL-OPF does not require classification and reorganization of all samples in the dataset at each iteration. For this reason, the selection process turns out to be very fast even for large datasets.

Chapter 5

Conclusions and Extensions

Advances in data acquisition and storage technologies have provided large datasets to support research, technological development, entertainment, medical diagnosis, among others. Annotation is the most effective way to organize data and retrieve the desired information. However, as the datasets grow large, manual annotation becomes impractical, and despite the efforts in automatic annotation, their success usually depend on a much smaller training set. Active learning strategies, aim to select a considerably lower number of the most informative samples to train a pattern classifier with expert's supervision. However, the majority of them cannot cope with large datasets, once they fall in a single paradigm which requires, at each learning iteration, the classification and/or organization of the entire dataset.

This PhD research addressed these issues by proposing active learning methods that are effective and efficient, in the number of iterations and response time for each iteration. It presents a priori data reduction and organization strategies, as well as strategies for selecting samples from the reduced and/or organized dataset. It describes a learning framework gathering the paradigm and strategies proposed as well as enabling posterior adoption and incorporation of new active learning methods. As far as we know, the proposed paradigm differs from the existing approaches in the following aspects: it previously organizes the data and then properly performs, both phases, classification and selection, alternately until the number of samples to be displayed to the expert at each iteration is reached. A priori data organization avoids to reprocess the large dataset at each iteration while classification and selection of one sample (at a time) on the ordered set avoids to label all the samples in the dataset, so providing interactive response time.

We also analyzed and developed new active learning strategies in order to select the most informative samples for training classifiers more effectively and with minimal human intervention, as well as apply them in different applications of pattern recognition. Our active learning strategies iteratively seek to select the most informative samples based on

the current knowledge of the classifier and on a priori reduction and/or organization of samples. Basically, the data reduction and organization processes rely on graph clustering. From the clustering, we gained valuable information, since representative samples located at the center of the clusters (root samples) are more likely to cover all classes and are good candidates to be selected first for manual annotation. Furthermore, samples in the same cluster are likely to have the same label. This assumption could be used to accelerate active learning by reducing the number of annotating samples from the same cluster.

The previous reduction process, presented in Section 3.2.1, was performed for some of the proposed sorting strategies, such as: DBE, MST-BE and ASSL-OPF (presented in Sections 3.3.1, 3.3.2 and 3.4, respectively). Cluster roots and boundary between distinct clusters samples that form the reduced learning set, allow to select the most informative samples earlier for the training of the classifier. The MST-BE strategy presents a better organization way, selecting the more difficult samples than by the DBE one. Therefore, in a semi-supervised learning setting, the ASSL-OPF used MST-BE strategy. The proposed strategies were extensively assessed with different types of unsupervised and supervised classifiers using datasets from distinct applications, such as: image segmentation [26], forest cover type [21], handwritten digits [25], faces [23], cowhide [22] and parasites [24] recognition. The experiments performed on these datasets show that the proposed approach requires only a few iterations to achieve high accuracy and with less expert involvement than the baseline approaches.

In a real problem of diagnosis of parasites, under a scenario with the presence of a diverse class (as the fecal impurity class), these strategies with a reduction process showed considerably less effective. The data reduction can discard crucial samples to the learning process. In this case, it is important to be careful because some parasite species and/or impurities can be out of the cluster border. Therefore, we also searched for a more robust solution, the RDS approach (presented in Section 3.3.3), which previously organizes the data but without discarding any of them and then properly balance the selection of diverse and uncertain samples for training. Selecting samples from the ordered list of each cluster, give us a greater diversity. Selecting samples classified into a different class of their root's class, offered us the most difficult (most uncertain) samples for classification.

We validated our approach by an experienced expert in parasitology whithin a realistic scenario concerning a laboratory routine. RDS provides high classification accuracy for the automated diagnosis of parasites (much higher than in the traditional visual analysis, which can reach from 48.3% up to 75.9%). Moreover, it is computationally and iteratively efficient, providing interactive response times and requiring verification of a considerably smaller part of the dataset. It is worth noting that the lack of human knowledge or the exhaustive work can cause expert's misannotation. The presence of label-noise has an adverse effect on the performance of the classifier's learning. The presented approach is

not designed with label-noise in mind. This problem was avoided, because the expert was very careful in reviewing his annotations. However, extensions related to label-noise should be considered in future works.

Further, this developed approach was demonstrated to be useful in the context of medicine, more specifically focusing on the diagnosis of intestinal parasites. It is important to highlight that the presented approach is general enough to be further investigated and adaptabled for other application domains, such as demonstrated by several examples, which were presented, as well as for other ones with a diverse class (see future works).

There are many possible extensions to this PhD research. Examples of some of these extensions are: (i) development of new ways to explore the reduction and organization of data; (ii) application and/or adaptation of the proposed strategies in different research areas, such as: computer networks, remote sensing, faces recognition, among others; (iii) comparison of the proposed strategies with different pattern classifiers, clustering techniques and state-of-the-art's approaches using different application domains; (iv) use of active learning strategies in the selection and fusion of pattern classifiers. We can select the most informative training set for each classifier; (v) development of new methods related to active semi-supervised learning (ASSL) approach. In the ASSL approach, in general, the unlabeled set is randomly chosen. One possible extension is explore the reduction, organization and selection of the most representative samples to form (besides the labeled set) the unlabeled set too; (vi) use of the RDS approach for other applications with a diverse class. Since RDS presents a general solution for active learning, we believe that it can be successfully used for other applications with a diverse class such as the impurity class; (vii) investigation of techniques to make the RDS method more robust to possible expert's mislabeling during active learning. One extension could be consider multiple experts throughout the annotation process. Another one could be develop a mechanism that identifies possible mislabeling according to the previous learning iterations; (vii) application of our paradigm to Content-Based Image Retrieval (CBIR) problems. Although our paradigm concentrates on pattern recognition problems, it can be easily extended to CBIR problems. This extension to CBIR applications requires an additional sorting of samples in the training set based on the knowledge of queries and on relevant and irrelevant samples from previous iterations; (viii) use of active learning in superpixel-based interactive classification of very high resolution images [96]; (ix) another direction is towards active learning in multi-label problems, wherein each image can belong to multiple categories simultaneously.

Bibliography

[1] Willian P. Amorim, Alexandre X. Falcão, and Marcelo H. de Carvalho. Semi-supervised pattern classification using optimum-path forest. In *XXVII SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–7, 2014.

[2] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[3] Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th Internation Conference on Machine Learning (ICML)*, pages 59–66. AAAI Press, 2003.

[4] Pedro H. Bugatti, Marcela X. Ribeiro, Agma J. M. Traina, and Caetano Traina Jr. Feature selection guided by perception in medical CBIR systems. In *IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*, pages 323–330. IEEE Computer Society, 2011.

[5] Colin Campbell, Nello Cristianini, and Alex J. Smola. Query learning with large margin classifiers. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 111–118, 2000.

[6] Fábio A. M. Cappabianco, Alexandre X. Falcão, Clarissa L. Yasuda, and Jayaram K. Udupa. Brain tissue mr-image segmentation via optimum-path forest clustering. *Computer Vision and Image Understanding*, 116(10):1047–1059, 2012.

[7] Ana Cardoso-Cachopo and Arlindo L. Oliveira. Semi-supervised single-label text categorization using centroid-based classifiers. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC)*, pages 844–851, 2007.

[8] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[9] Jian Cheng and Kongqiao Wang. Active learning for image retrieval with Co-SVM. *Pattern Recognition*, 40(1):330 – 334, 2007.

[10] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[11] Giovani Chiachia, Aparecido N. Marana, João P. Papa, and Alexandre X. Falcão. Infrared face recognition by optimum-path forest. In *16th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–4, 2009.

[12] Cheng-Chieh Chiang. Interactive tool for image annotation using a semi-supervised and hierarchical approach. *Computer Standards & Interfaces*, 35(1):50–58, 2012.

[13] Li-Jen Chien, Chien-Chung Chang, and Yuh-Jye Lee. Variant methods of reduced set selection for reduced support vector machines. *Journal of Information Science and Engineering*, 26(1):182–196, 2010.

[14] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[15] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145, 1996.

[16] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Large scale transductive SVMs. *The Journal of Machine Learning Research*, 7:1687–1712, 2006.

[17] André T. da Silva, Alexandre X. Falcão, and Léo P. Magalhães. A new CBIR approach based on relevance feedback and optimum-path forest classification. *Journal of WSCG*, 18(1-3):73–80, 2010.

[18] André T. da Silva, Alexandre X. Falcão, and Léo P. Magalhães. Active learning paradigms for CBIR systems based on optimum-path forest classification. *Pattern Recognition*, 44(12):2971–2978, 2011.

[19] Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 150–157. Morgan Kaufmann, 1995.

[20] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 235–242. MIT Press, 2005.

[21] Covertype Dataset. Forest Covertype Dataset. UCI - Machine Learning Repository, 2014. `http://archive.ics.uci.edu/ml/datasets/Covertype`.

[22] Cowhide Dataset. Cowhide Defect Dataset. Institute of Computing, Federal University of Mato Grosso do Sul, 2014.

[23] Face Dataset. Biometrics Datasets. The Computer Vision Laboratory, University of Notre Dame, 2014. `http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html`.

[24] Parasite Dataset. Parasite Diagnosis Dataset. IC-UNICAMP - Institute of Computing, University of Campinas, 2014.

[25] Pendigit Dataset. Pen-Based Recognition of Handwritten Digits Dataset. UCI - Machine Learning Repository, 2014. `http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits`.

[26] Statlog Dataset. Statlog (Landsat Satellite) Dataset. UCI - Machine Learning Repository, 2014. `https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)`.

[27] F. Del Frate, F. Pacifici, G. Schiavon, and C. Solimini. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):800–809, 2007.

[28] Li Deng and Dong Yu. *Deep Learning: Methods and Applications*. Now Publishers Inc, 1 edition, 2014.

[29] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Deroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10–11):2436–2449, 2011.

[30] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 127–136, 2007.

[31] Hugo J. Escalante, Manuel Montes, and Luis E. Sucar. Multi-class particle swarm model selection for automatic image annotation. *Expert Systems with Applications*, 39(12):11011–11021, 2012.

[32] Alexandre X. Falcão, Jorge Stolfi, and Roberto de A. Lotufo. The Image Foresting Transformation: Theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[33] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[34] Yifan Fu, Bin Li, Xingquan Zhu, and Chengqi Zhang. Do they belong to the same class: active learning by querying pairwise label homogeneity. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2161–2164, 2011.

[35] Yifan Fu and Xingquan Zhu. Optimal subset selection for active learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 1776–1777. AAAI Press, 2011.

[36] K. Fukunaga and Patrenahalli M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, 24(7):750–753, 1975.

[37] Lynne S. Garcia. *Practical Guide to Diagnostic Parasitology*. ASM Press, 2 edition, 4 2009.

[38] Priyanka Garg and Sellamanickam Sundararajan. Active learning in partially supervised classification. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1783–1786, 2009.

[39] Kingshy S. Goh, Edward Y. Chang, and Beitao Li. Using one-class and two-class SVMs for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering*, 17(10):1333–1346, 2005.

[40] Jancarlo F. Gomes, Sumie Hoshino-Shimizu, Luiz C. S. Dias, Ana J. S. A. Araujo, Vera L. P. Castilho, and Fátima A. M. A. Neves. Evaluation of a novel kit (tf-test) for the diagnosis of intestinal parasitic infections. *Journal of Clinical Laboratory Analysis*, 18(2):132–138, 2004.

[41] Ivan R. Guilherme, Aparecido N. Marana, João P. Papa, Giovani Chiachia, Luis C. S. Afonso, Kazuo Miura, Marcus V. D. Ferreira, and Francisco Torres. Petroleum well drilling monitoring through cutting image analysis and artificial intelligence techniques. *Engineering Applications of Artificial Intelligence*, 24(1):201–207, 2011.

[42] Richang Hong, Meng Wang, Yue Gao, Dacheng Tao, Xuelong Li, and Xindong Wu. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Transactions on Cybernetics*, 44(5):669–680, 2014.

[43] Te Ming Huang and Vojislav Kecman. SemiL - software for solving semi-supervised learning problems, 2009. `http://www.support-vector.ws/html/semil.html`.

[44] Alexander I. Iliev, Michael S. Scordilis, João P. Papa, and Alexandre X. Falcão. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*, 24(3):445–460, 2010.

[45] Adriana S. Iwashita, João P. Papa, André N. de Souza, Alexandre X. Falcão, Roberto de A. Lotufo, Victor M. de A. Oliveira, Victor Hugo C. de Albuquerque, and João Manuel R. S. Tavares. A path- and label-cost propagation approach to speedup the training of the optimum-path forest classifier. *Pattern Recognition Letters*, 40(0):121–127, 2014.

[46] Prateek Jain and Ashish Kapoor. Active learning for large multi-class problems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–769, 2009.

[47] Jiwoon Jeon, Victor P. Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings the 26th Annual International ACM Conference on Research and Development in Informaion Retrieval (SIGIR)*, pages 119–126, 2003.

[48] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 200–209, 1999.

[49] Ajay J. Joshi, Fatih Porikli, and Nikolaos P. Papanikolopoulos. Scalable active learning for multi-class image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2259–2273, 2012.

[50] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian Processes for Object Categorization. *International Journal of Computer Vision*, 88(2):169–188, 2010.

[51] Christine Körner and Stefan Wrobel. Multi-class ensemble-based active learning. In *European Conference on Machine Learning (ECML)*, volume 4212 of *Lecture Notes in Computer Science*, pages 687–694. Springer-Verlag, 2006.

[52] Yuh-Jye Lee and Olvi L. Mangasarian. RSVM: Reduced support vector machines. In *IEEE International Conference on Data Mining (ICDM)*, pages 00–07, 2001.

[53] Jia Li and James Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[54] Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[55] Ying Liu, Dengsheng Zhang, and Guojun Lu. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570, 2008.

[56] Edwin Lughofer. Hybrid active learning for reducing the annotation effort of operators in classification systems. *Pattern Recognition*, 45(2):884–896, 2012.

[57] Edwin Lughofer. Single-pass active learning with conflict and ignorance. *Evolving Systems*, 3(4):251–271, 2012.

[58] Edwin Lughofer, James E. Smith, Muhammad A. Tahir, Praminda Caleb-Solly, Christian Eitzinger, Davy Sannen, and Marnix Nuttin. Human-machine interaction issues in quality control based on online image classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 39(5):960–971, 2009.

[59] James B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[60] Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 584–591, 2004.

[61] Rodrigo Minetto, João P. Papa, Thiago V. Spina, Alexandre X. Falcão, Neucimar J. Leite, and Jorge Stolfi. Fast and robust object tracking using image foresting transform. In *16th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–4, 2009.

[62] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27(1):203–233, 2006.

[63] Rodrigo Y. M. Nakamura, Priscila T. M. Saito, João P. Papa, Pedro J. de Rezende, and Alexandre X. Falcão. Choosing the most effective pattern classification model under learning-time constraint. In *submitted to Pattern Recognition Letters*, pages 1–33, 2014.

[64] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 79–86, 2004.

[65] Yabo Ni, Miao Zheng, Jiajun Bu, Chun Chen, and Dazhou Wang. Personalized automatic image annotation based on reinforcement learning. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2013.

[66] Bruno M. Nogueira, Alípio M. Jorge, and Solange O. Rezende. Hierarchical confidence-based active clustering. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC)*, pages 216–219, 2012.

[67] Mustapha Oujaoura, Brahim Minaoui, and Mohammed Fakir. A semantic approach for automatic image annotation. In *8th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–8, 2013.

[68] João P. Papa, Alexandre X. Falcão, Victor Hugo C. de Albuquerque, and João Manuel R. S. Tavares. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, 45(1):512–520, 2012.

[69] João P. Papa, Alexandre X. Falcão, Alexandre L. M. Levada, Débora C. Corrêa, Denis H. P. Salvadeo, and Nelson D. d'Ávila. Mascarenhas. Fast and accurate holistic face recognition using optimum-path forest. In *16th International Conference on Digital Signal Processing*, pages 1–6, 2009.

[70] João P. Papa, Alexandre X. Falcão, and Celso T. N. Suzuki. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, 19(2):120–131, 2009.

[71] João P. Papa, Aparecido N. Marana, André A. Spadotto, Rodrigo C. Guido, and Alexandre X. Falcão. Robust and fast vowel recognition using optimum-path forest. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2190–2193, 2010.

[72] João P. Papa, Alexandre X. Falcão, Greice M. de Freitas, and Ana Maria H. de Avila. Robust pruning of training patterns for optimum-path forest classification applied to satellite-based rainfall occurrence estimation. *IEEE Geoscience and Remote Sensing Letters*, 7(2):396–400, 2010.

[73] Soo Beom Park, Jae Won Lee, and Sang-Kyoon Kim. Content-based image classification using a neural network. *Pattern Recognition Letters*, 25(3):287–300, 2004.

[74] Xiaojun Qi and Yutao Han. Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition*, 40(2):728–741, 2007.

[75] Yanjun Qi, Sujatha G. Das, Ronan Collobort, and Jason Weston. Deep learning for character-based information extraction. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, 2014.

[76] Umaa Rebbapragada and Kiri L. Wagstaff. Using ensemble decisions and active selection to improve low-cost labeling for multi-view data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1–6, 2011.

[77] Leonardo M. Rocha, Fábio A. M. Cappabianco, and Alexandre X. Falcão. Data clustering as an optimum-path forest problem with applications in image analysis. *International Journal of Imaging Systems and Technology*, 19(2):50–68, 2009.

[78] Douglas Rodrigues, Luís A. M. Pereira, Rodrigo Y. M. Nakamura, Kelton A. P. Costa, Xin-She Yang, André N. Souza, and João P. Papa. A wrapper approach for feature selection based on bat algorithm and optimum-path forest. *Expert Systems with Applications*, 41(5):2250–2258, 2014.

[79] Enrique Vidal Ruiz. An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters*, 4(3):145–157, 1986.

[80] Priscila T. M. Saito, Willian P. Amorim, Alexandre X. Falcão, Pedro J. de Rezende, Celso T. N. Suzuki, Jancarlo F. Gomes, and Marcelo H. de Carvalho. Active semi-supervised learning using optimum-path forest. In *22nd International Conference on Pattern Recognition (ICPR)*, pages 1–6, 2014.

[81] Priscila T. M. Saito, Pedro J. de Rezende, Alexandre X. Falcão, Celso T. N. Suzuki, and Jancarlo F. Gomes. Improving active learning with sharp data reduction. In *WSCG Communication Proceedings of 20th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, pages 27–34, 2012.

[82] Priscila T. M. Saito, Pedro J. de Rezende, Alexandre X. Falcão, Celso T. N. Suzuki, and Jancarlo F. Gomes. An active learning paradigm based on a priori data reduction and organization. *Expert Systems with Applications*, 41(14):6086–6097, 2013.

[83] Priscila T. M. Saito, Pedro J. de Rezende, Alexandre X. Falcão, Celso T. N. Suzuki, and Jancarlo F. Gomes. A data reduction and organization approach for efficient image annotation. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*, pages 53–57, 2013.

[84] Priscila T. M. Saito, Celso T. N. Suzuki, Jancarlo F. Gomes, Alexandre X. Falcão, and Pedro J. de Rezende. Robust active learning for the diagnosis of parasites. *submitted to Pattern Recognition*, 2014.

[85] Jianfeng Shen, Bin Ju, Tao Jiang, Jingjing Ren, Miao Zheng, Chengwei Yao, and Lanjuan Li. Column subset selection for active learning in image classification. *Neurocomputing*, 74(18):3785–3792, 2011.

[86] Xuehua Shen and ChengXiang Zhai. Active feedback - UIUC TREC-2003 HARD experiments. In *Text REtrieval Conf. (TREC)*, pages 662–666, 2003.

[87] Celso T. N. Suzuki, Jancarlo F. Gomes, Alexandre X. Falcão, João P. Papa, and Sumie Hoshino-Shimizu. Automatic segmentation and classification of human intestinal parasites from microscopy images. *IEEE Transactions on Biomedical Engineering*, 60(3):803–812, 2013.

[88] Celso T. N. Suzuki, Jancarlo F. Gomes, Alexandre X. Falcão, Sumie H. Shimizu, and João P. Papa. Automated diagnosis of human intestinal parasites using optical microscopy images. In *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 460–463, 2013.

[89] Gerard Sychay, Edward Y. Chang, and Kingshy Goh. Effective image annotation via active learning. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, volume 1, pages 209–212, 2002.

[90] Jinhui Tang, Richang Hong, Shuicheng Yan, Tat-Seng Chua, Guo-Jun Qi, and Ramesh Jain. Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology*, 2(2):14:1–14:15, 2011.

[91] Jinhui Tang, Shuicheng Yan, Chunxia Zhao, Tat-Seng Chua, and Ramesh Jain. Label-specific training set construction from web resource for image annotation. *Signal Processing*, 93(8):2199–2204, 2012.

[92] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 406–414, 1999.

[93] Aibo Tian and Matthew Lease. Active learning to maximize accuracy vs. effort in interactive information retrieval. In *Proceedings of the 34th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 145–154, 2011.

[94] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the 9th ACM International Conference on Multimedia (MULTIMEDIA)*, pages 107–118, 2001.

[95] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[96] John E. Vargas, Priscila T. M. Saito, Alexandre X. Falcão, Pedro J. de Rezende, and Jefersson A. dos Santos. Superpixels-based interactive classification of very high resolution images. In *XXVII SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–7, 2014.

[97] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[98] Zhiyu Wang, Dingyin Xia, and Edward Y. Chang. A deep-learning model-based and data-driven hybrid architecture for image annotation. In *Proceedings of the International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval*, pages 13–18, 2010.

[99] Roger C. F. Wong and Clement H. C. Leung. Automatic semantic annotation of real-world web images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1933–1944, 2008.

[100] Yi Wu, Igor Kozintsev, Jean-Yves Bouguet, and Carole Dulong. Sampling strategies for active learning in personal photo retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 529–532, 2006.

[101] Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. Representative sampling for text classification using support vector machines. In *Proceedings of the 25th European Conference on IR Research (ECIR)*, pages 393–407, 2003.

[102] Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[103] Shile Zhang, Bin Li, and Xiangyang Xue. Semi-automatic dynamic auxiliary-tag-aided image annotation. *Pattern Recognition*, 43(2):470–477, 2010.

[104] Sheng-hua Zhong, Yan Liu, and Yang Liu. Bilinear deep learning for image classification. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 343–352, 2011.

[105] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919, 2003.