Pattern Recognition 48 (2015) 3572–3583

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Robust active learning for the diagnosis of parasites

Priscila T.M. Saito [a,b,*], Celso T.N. Suzuki [a,c], Jancarlo F. Gomes [a,d], Pedro J. de Rezende [a], Alexandre X. Falcão [a]

[a] Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, 13083-852 - Campinas, SP, Brazil
[b] Department of Computing, Federal University of Technology - Parana, Brazil
[c] IMMUNOCAMP Research and Development of Technology, Brazil
[d] Biology Institute, University of Campinas, Brazil

ARTICLE INFO

Article history:
Received 9 April 2014
Received in revised form
9 March 2015
Accepted 18 May 2015
Available online 28 May 2015

Keywords:
Active learning
Pattern recognition
Automated diagnosis of intestinal parasites
Microscopy image analysis
Optimum-path forest classifiers

ABSTRACT

We have developed an automated system for the diagnosis of intestinal parasites from optical microscopy images. The objects (species of parasites and impurities) segmented from these images form a large dataset. We are interested in the active learning problem of selecting a reasonably small number of objects to be labeled under an expert's supervision for use in training a pattern classifier. However, impurities are very numerous, constitute several clusters in the feature space, and can be quite similar to some species of parasites, leading to a significant challenge for active learning methods. We propose a technique that pre-organizes the data and then properly balances the selection of samples from all classes and uncertain samples for training. Early data organization avoids reprocessing of the large dataset at each learning iteration, enabling the halting of sample selection after a desired number of samples per iteration, yielding interactive response time. We validate our method by comparing it with state-of-the-art approaches, using a previously labeled dataset of almost 6000 objects. Moreover, we report results from experiments on a very realistic scenario, consisting of a dataset with over 140,000 unlabeled objects, under unbalanced classes, the absence of some classes, and the presence of a very large set of impurities.

1. Introduction

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 [1] — an exhaustive process whose abbreviation may seriously compromise the quality of the diagnosis. We describe in this paper an automated system we have developed for this application, which can considerably improve the diagnosis sensitivity and reliability [2,3].

In our system, each lab exam produces about 2700 images of 4M pixels each for analysis, and each image may contain from tens to thousands of objects to be labeled either as impurity or as some specie 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 [4,5], 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, impurities are exceptionally abundant, form several clusters in the feature space, and are quite similar to some species of parasites, resulting in a major challenge for existing methods. We propose to pre-organize the data and then properly balance the selection of samples from all classes and uncertain samples for training a classifier, resulting in a robust active learning approach. Although, *robust active learning* often refers to processing

* Corresponding author at: Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, 13083-852 - Campinas, SP, Brazil. Tel.: +551935215881; fax: +551935215847

E-mail addresses: maeda@ic.unicamp.br (P.T.M. Saito), celso.suzuki@ic.unicamp.br (C.T.N. Suzuki), jgomes@ic.unicamp.br (J.F. Gomes), rezende@ic.unicamp.br (P.J. de Rezende), afalcao@ic.unicamp.br (A.X. Falcão).

http://dx.doi.org/10.1016/j.patcog.2015.05.020
0031-3203/

misannotated samples [6,7], in this work, it is employed in relation to the performance variation of the method with respect to the absence or presence of a "diverse class" (a class with samples that might appear everywhere in the feature space), such as the fecal impurity class in the diagnosis of parasites.

In our approach, data organization is based on clustering, followed by sorting of the samples within each cluster, according to their distance to their representative (root) sample in the cluster. The distance criterion can be based on the values of the optimum paths from the roots to their most strongly connected samples in the cluster. The definition of distance between sample and cluster root is better described in Section 2. 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 cluster according to the corresponding ordered list so long as their classification does not match the class of the corresponding root. This strategy allows us to explore class diversity by covering all classes faster and, at the same time, to select uncertain samples, which are more informative for training a classifier for the next iteration. Different 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.

Traditional active learning methods usually focus on binary classification for information retrieval, which is not the aim of this work. Although some effort has been made to extend these methods to multi-class classification and also to exploit clustering in the selection of training samples, they usually classify the *entire* dataset at each iteration of the learning process (see Section 2). Moreover, regrouping samples at each learning iteration often takes place as well.

To the best of our knowledge, the method proposed here is unique in the sense that it performs data organization only once, *a priori*, and takes into account, for sample selection, class diversity and sample uncertainty. We validate our technique by using distinct clustering and classification methods in comparison to two other state-of-the-art active learning approaches. The experiments we performed involved a dataset with almost 6000 objects, carefully labeled by an experienced expert in parasitology. For accuracy evaluation, this dataset was divided into learning and test sets several times, such that active learning was performed from training samples selected from the first set, while the final classifier was evaluated on the second set of unseen samples. We have also evaluated the proposed method in a realistic scenario, consisting of a dataset with over 140,000 unlabeled objects, under unbalanced classes, the absence of some classes, and the presence of a very large set of impurities. In this case, the expert participated in the active learning process that involved label verification of only about 7% of the samples. Subsequently, the classifier annotated the remaining samples and 6% of them were randomly selected for accuracy evaluation by the expert.

Our contributions: The primary contribution of this paper is an active learning technique that

1. performs data organization only once (*a priori*), as a preprocessing.
2. properly balances class diversity and sample uncertainty for useful sample selection during the learning process of a classifier;
3. provides high classification accuracy for the automated diagnosis of parasites (much higher than in visual analysis);
4. is computationally and iteratively efficient, providing interactive response times and requiring verification of only a small part of the dataset;
5. is validated by an expert in a realistic scenario;
6. is shown to be more robust to the presence of impurities than two other state-of-the-art methods; and
7. is general enough to be further investigated for other applications where a diverse class (such as the impurity class) is present.

This paper describes background material and related works in Section 2, introduces the new active learning technique in Section 3, discusses the experiments and results in Section 4, while Section 5 presents our conclusions and future work.

2. Background

Image annotation methods have relied on several types of supervised classifiers [8]: Bayesian [9–11], Support Vector Machines (SVM) [12,13], Artificial Neural Network (ANN) [14,15], *k*-Nearest Neighbor (*k*-NN) [16,17], Decision Tree (DT) [18–20] and Optimum-Path Forest (OPF) [21,22]. In conventional supervised learning, the algorithm passively accepts randomly selected 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 for automatically annotating the remaining or future samples. Active learning techniques can accomplish both goals. Theoretical results show that they can significantly reduce the number of required training samples as compared to random selection for achieving similar classification accuracy [4,5].

A common approach for collecting data in active learning is to select the most uncertain samples as the closest ones to the classification boundary. An active learning strategy using Support Vector Machines (Al-SVM) has been used in [23,24], under the assumption that the most useful samples are the closest to the separating hyperplanes.

The closest-to-boundary criterion performs well in typical active learning applications [25,26]. However, it presents deficiencies in the presence of a diverse class, such as the fecal impurity class in the diagnosis of parasites. Given that the classifier actively participates in its own learning process, it is crucial that it has knowledge of samples from all classes as early as possible, so as to become aware of the diversity of the problem. Clustering techniques can provide that information when representative cluster samples (roots) are selected first for an expert verification (initially for manual annotation and, subsequently, for confirmation or correction). Furthermore, samples in the same cluster are likely to have the same label. This assumption can be explored to accelerate active learning, by reducing the number of annotating samples from the same cluster, and to identify uncertain samples whenever the classifier assigns a label different from the one of the corresponding root.

Several works have explored clustering techniques for active learning [27–31]. However, a common aspect in all these methods is that they classify and/or sort the entire learning set at each iteration, as illustrated in Fig. 1. One should realize that this procedure is not feasible for practical applications that use large datasets, as it compromises interactive response time during the active learning process.

In contrast, in [32], the authors developed data reduction and organization strategies (MST-BE) that are carried out only once (*a priori*), as shown in Fig. 2. There, data reduction relies on an effective clustering approach. The cluster roots are initially selected for manual annotation, so as to acquire knowledge about the diversity of the problem by ensuring samples from all classes at the first iteration. Uncertain samples are obtained from the boundary between distinct clusters, as the most difficult ones for