Expert Systems with Applications 41 (2014) 6086–6097

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An active learning paradigm based on a priori data reduction and organization

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

[a] Institute of Computing, University of Campinas, SP, Brazil
[b] Dept. of Computer Engineering, Federal Technological University of Parana, PR, Brazil
[c] IMMUNOCAMP Research and Development of Technology, SP, Brazil
[d] Institute of Biology, University of Campinas, SP, Brazil

ARTICLE INFO

Keywords:
Pattern recognition
Active learning
Image annotation
Machine learning
Data mining

ABSTRACT

In the past few years, active learning has been reasonably successful and it has drawn a lot of attention. However, recent active learning methods have focused on strategies in which a large unlabeled dataset has to be reprocessed at each learning iteration. As the datasets grow, these strategies become inefficient or even a tremendous computational challenge. In order to address these issues, we propose an effective and efficient active learning paradigm which attains 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. Experimental results showed that the proposed paradigm allows to achieve high accuracy quickly with minimum user interaction, further improving its efficiency.

1. Introduction

The advances in computing and multimedia technologies enable the low-cost acquisition and storage of large datasets containing millions of samples. An increasingly urgent research topic in the information retrieval and machine learning communities is to determine effective and efficient ways of handling and analyzing these large datasets.

Manually annotating large datasets becomes impractical. Moreover, human annotations are usually too subjective and ambiguous. Therefore, most of the research efforts has employed automatic or semi-automatic annotation techniques.

Since human knowledge is indispensable for the success of the learning phase and user's time and effort are precious resources, active learning techniques have aimed to iteratively select unlabeled samples for manual annotation that are most difficult for classification, so that human effort is focused on annotating the informative ones.

A key challenge in the selection of the samples is the consideration of aspects such as:

- How many samples (iterations) should be used in the learning process?
- How to ensure the selection of at least one representative sample from each semantic class?
- How to ensure that these samples will be informative to speed up the learning process?

In this paper, we present a novel and efficient active learning paradigm (Data Reduction and Organization Paradigm, DROP for short) that answers these questions in a satisfactory manner. To do so, we developed a preprocessing algorithm that reduces and organizes the dataset.

The proposed paradigm 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 leave us with essentially the informative samples.

By reducing and organizing the learning set a priori, DROP is verifiably effective and more suitable to be used in practical applications with large datasets. Furthermore, a remarkably faster selection process is completed by interlacing the choice of samples and their classification in a joint process that reduces their number

* Corresponding author at: Institute of Computing, University of Campinas, SP, Brazil. Tel.: +55 01935215881.

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

http://dx.doi.org/10.1016/j.eswa.2014.04.007
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

while selecting the most relevant ones. This strategy sets DROP apart from most recent learning methods since, in these, *all* samples in the database have to be classified and/or reorganized at each iteration.

Being a paradigm, DROP can be implemented using various strategies. We developed boundary-based reduction and sorting strategies. The boundary reduction strategy relies on an effective clustering approach. The reduced learning set is comprised of the representative samples of each cluster in order to ensure samples from all classes, as well as the boundary samples from distinct clusters, increasing the possibility that they represent more informative samples. The proposed sorting strategy is based on the Minimum Spanning Tree edges of boundary samples (obtained in the reduction process). The selected samples correspond to pairs farthest apart and whose labels (assigned by the current classifier) are different within each pair.

We have evaluated the proposed paradigm with distinct classifiers and clustering algorithms, using the traditional random selection process as baseline. Experiments performed on real datasets show that the proposed paradigm is able to iteratively generate classifiers that improve quickly, requires few iterations, and attains high accuracy while keeping user interaction to a minimum.

1.1. Our contributions

The main contributions of this paper are: (1) a new active learning paradigm which is shown to be more efficient and feasible in practice for large datasets; (2) an active learning method that: (i) selects the most informative samples for the learning process; (ii) provides high classification accuracy, (iii) is computationally, interactively and iteratively efficient. That is to say, the number of learning iterations is significantly reduced, while requiring the annotation and classification of only a small number of samples.

The remainder of this paper is structured as follows. Section 2 summarizes the main works and concepts in the field of image annotation and active learning. Section 3 details the proposed paradigm. Sections 4 and 5 discusses our experiments and results, respectively. Finally, Section 6 presents conclusions and ideas for future improvements.

2. Background

Image annotation is a process by which labels are associated with images, either manually, automatically or semi-automatically (Chiang, 2012; Lughofer et al., 2009; Tang, Yan, Zhao, Chua, & Jain, 2012; Zhang, Li, & Xue, 2010). The manual annotation approach presents some drawbacks such as being time consuming and laborious. Hence, the new trend towards automatic image annotation seems promising (Carneiro, Chan, Moreno, & Vasconcelos, 2007; Escalante, Montes, & Sucar, 2012; Dimitrovski, Kocev, Loskovska, & Deroski, 2011; Liu, Li, Liu, Lu, & Ma, 2009; Zhang, Islam, & Lu, 2012).

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 classifiers 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 (Zhang et al., 2012) explore the label annotation using conventional classification methods, such as Bayesian (Li & Wang, 2008; Carneiro et al., 2007; Jeon, Lavrenko, & Manmatha, 2003), Support Vector Machines (SVM) (Goh, Chang, & Li, 2005; Qi & Han, 2007), Artificial Neural Network (ANN) (Del Frate, Pacifici, Schiavon, & Solimini, 2007; Park, Lee, & Kim, 2004), *k*-Nearest Neighbor (*k*-NN) (Jain & Kapoor, 2009; Tang et al., 2011), Decision Tree (DT) (Liu, Zhang, & Lu, 2008; Vens, Struyf, Schietgat, Džeroski, & Blockeel, 2008; Wong & Leung, 2008) and Optimum-Path Forest (OPF) (da Silva, Falcão, & Magalhães, 2010, 2011).

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 user 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 (Bugatti, Ribeiro, Traina, & Traina, 2011; Rodrigues et al., 2014) and learning techniques (Bengio, 2009; Wang, Xia, & Chang, 2010; Zhong, Liu, & Liu, 2011).

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 user.

The idea of using active learning to assist in image labeling has received a lot of research attention (da Silva, Falcão, & Magalhães, 2011; Jain & Kapoor, 2009; Joshi, Porikli, & Papanikolopoulos, 2012; Lughofer, 2012; Rebbapragada & Wagstaff, 2011; Shen et al., 2011; Sychay, Chang, & Goh, 2002; Tong & Chang, 2001, 2002; Wu, Kozintsev, Bouguet, & Dulong, 2006). Fig. 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 user is asked to annotate/correct a non-annotated/classified sample set chosen by the selector. As the samples are annotated/corrected by the user, 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 user. 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 (Ertekin, Huang, Bottou, & Giles, 2007; Fu, Li, Zhu, & Zhang, 2011; Garg & Sundararajan, 2009; Tian & Lease, 2011). Relatively few works have been devoted for multi-class active learning and these are typically based on ensemble or committee classifiers (as in Rebbapragada & Wagstaff (2011), Fu & Zhu (2011), Muslea, Minton, & Knoblock (2006), Dagan & Engelson (1995) and Melville & Mooney (2004)) or extensions of predominantly binary active learning methods to the multi-class scenario (as in Kapoor, Grauman, Urtasun, & Darrell (2010) and Jain & Kapoor (2009)).

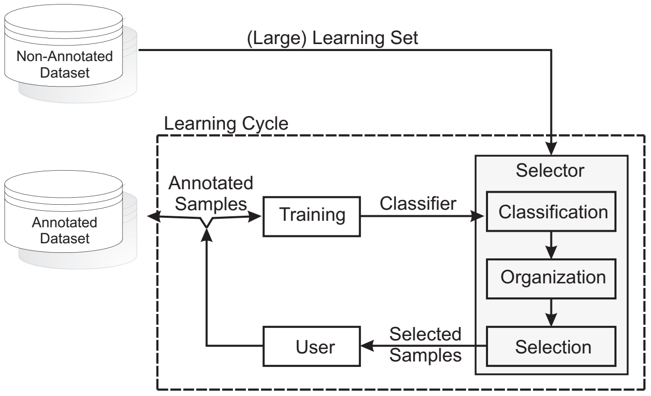

Fig. 1. Pipeline of the traditional active learning paradigm.