Expert Systems with Applications 41 (2014) 2372–2378

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Semi-supervised learning combining co-training with active learning

Yihao Zhang [a], Junhao Wen [b,*], Xibin Wang [a], Zhuo Jiang [a]

[a] College of Computer Science, Chongqing University, Chongqing 400030, China
[b] College of Software Engineering, Chongqing University, Chongqing 400030, China

ARTICLE INFO

Keywords:
Semi-supervised learning
Co-training
Confidence estimation
Active learning
Informative instances

ABSTRACT

Co-training is a good paradigm of semi-supervised, which requires the data set to be described by two views of features. There are a notable characteristic shared by many co-training algorithm: the selected unlabeled instances should be predicted with high confidence, since a high confidence score usually implies that the corresponding prediction is correct. Unfortunately, it is not always able to improve the classification performance with these high confidence unlabeled instances. In this paper, a new semi-supervised learning algorithm was proposed combining the benefits of both co-training and active learning. The algorithm applies co-training to select the most reliable instances according to the two criterions of high confidence and nearest neighbor for boosting the classifier, also exploit the most informative instances with human annotation for improve the classification performance. Experiments on several UCI data sets and natural language processing task, which demonstrate our method achieves more significant improvement for sacrificing the same amount of human effort.

1. Introduction

Semi-supervised learning is very useful in many practical applications, which learn from both labeled data and unlabeled data and automatically exploit unlabeled data for improving the learning performance without human intervention (Chapelle, Scholkopf, & Zien, 2006; Zhu, 2008). Co-training is a well-known Semi-supervised learning paradigm started from Blum and Mitchell's seminal work (Blum and Mitchell, 1998), which is proposed for binary classification problems in which two different views are available. The standard co-training algorithm requires two sufficient and redundant views (Blum and Mitchell, 1998), that is, the attributes can be naturally partitioned into two sets, each of which is sufficient for learning and conditionally independent to the other given the class label. Co-training works in an iterative manner that two classifiers are trained separately on the different views and the predictions of either classifier on unlabeled instances are used to augment the training set of the other (Zhang & Zhou, 2011; Zhu, 2008).

There are a notable characteristic shared by many co-training algorithm: the selected unlabeled instances should be predicted with high confidence, since a high confidence score usually implies that the corresponding prediction is correct (Blum and Mitchell, 1998; Mihalcea, 2004). Unfortunately, it is not always able to improve the classification performance with these high confidence unlabeled instances. Tang et al. (2007) proposed a new strategy that updating the classifiers through co-training, which add negative instances that are close to the classifier hyper-plane such that the classifier will learn to better distinguish these instances. Following the work on standard co-training, a number of relevant methods have been developed. Wang and Zhou (2010) analyzed the co-training process and viewed it as combinative label propagation over two views. Yu, Krishnapuram, Rosales, and Rao (2011) proposed a Bayesian undirected graphical model for co-training, which can elegantly handle data samples with missing views. Sun et al. (2011) proposed an entity-based co-training algorithm, which requires no prior knowledge about the underlying class distribution which is crucial in standard co-training algorithms. Unfortunately, co-training is called for precisely when the labeled training set is small, and it is uncertain whether the standard co-training would work or not on small labeled training sets (Du et al., 2011).

In this paper, a new semi-supervised classification algorithm was proposed which combines the benefits of both co-training and active learning, and the major contributions are two-fold:

(1) Firstly, in each co-training round, a few of most reliable instances were picked out from the unlabeled data for the next round of learning, and the most reliable instances were chosen according to the two criterions of high confidence and nearest neighbor. Specifically, the contribution degree was defined as the criteria of select informative instances, which not only considering the most uncertain of instances but also considering the uncertainty difference between the instance and its nearest neighbor.

* Corresponding author. Tel.: +86 13983146919.
E-mail addresses: yihaozhang@cqu.edu.cn (Y. Zhang), jhwen@cqu.edu.cn (J. Wen), binxiwang@cqu.edu.cn (X. Wang), jiangzhuo1986@gmail.com (Z. Jiang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.035

(2) Secondly, the active learning uses query framework that an active learner queries the instances about which it is least certain how to label, our algorithm defined contribution degree as the selection criteria of informative instances, which achieved more significant improvement for sacrificing the same amount of human effort and worked well on small labeled training sets.

The rest of this paper is organized as follows. Section 2 reviews some issues in the co-training and active learning. After that Section 3 introduces the sketch of the algorithm and presents the algorithms details. Section 4 reports experimental results on a number of real-world datasets and further analyzes the underlying reasons for the algorithm. Finally, Section 5 concludes and indicates several issues for future work.

2. Issues in co-training and active learning

Co-training is a semi-supervised, multi-view algorithm that uses the initial labeled data set to learn a weak classifier in each view (Blum and Mitchell, 1998). Then each classifier is applied to the rest of unlabeled instances, and co-training detects the instances on which each classifier makes the most confident predictions. These high-confidence instances are labeled with the estimated labels and added into the labeled data set. Based on the new training set, a new classifier is repeated for several iterations. At the end, a final hypothesis is created by a voting scheme that combines the predictions of the classifiers learned in each view.

Co-training, a good paradigm of semi-supervised learning, has drawn considerable attentions and interests recently (Zhou & Li, 2010). The standard co-training assumes that the data can be described by two disjoint sets of features or views, and it works well when the two views satisfy the sufficiency and independence assumptions (Blum and Mitchell, 1998). However, these two assumptions are often not known or ensured in practice, and view splitting is unreliable under the given small labeled training sets. More commonly, most supervised data sets are described by one set of attributes (one view). To exploit the advantage of co-training, Goldman and Zhou (Goldman and Zhou, 2000) proposed an algorithm which does not exploit feature partition; the algorithm uses two different supervised learning algorithms to train the two classifiers. Zhou and Li (2005) proposed the tri-training approach, which uses three classifiers generated from bootstrap samples of the original training set. Du and Ling (2011) got the conclusions that co-training's effectiveness are mixed. That is, if two views are given, and known to satisfy the two assumptions, co-training works well; Otherwise, based on small labeled training sets, verifying the assumptions or splitting single view into two views are unreliable; thus, it is uncertain whether the standard co-training would work or not.

Active learning, a subfield of machine learning, can perform better with less training by choosing the data form which it learns. It attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle (Settles, 2010) (e.g., a human annotator). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. The key idea behind most active learning algorithms is to select the instances that are most uncertain to classify. Therefore, a key aspect of active learning is to measure the classification uncertainty of unlabeled instances. Zhu (2003) proposed a new semi-supervised learning strategy, which combines active learning and semi-supervised learning under a Gaussian random field model. Yang et al. (2009) proposed Bayesian framework to active distance metric learning by selecting those unlabeled example pairs with the greatest uncertainty in relative distance. Lughofer (2012) proposed a novel active learning strategy for data-driven classifiers, which is essential for reducing the annotation and supervision effort of operators in off-line and on-line classification systems, as operators only have to label an exquisite subset of the off-line training data. Li, Shi, and Liu (2012) proposed a joint active learning approach which combines a novel generative query strategy and the existing discriminative one, which adaptively fits the distribution difference and shows higher robustness than the ones using single strategy.

From the above analysis, co-training is an important technique for improving the predictive accuracy when labeled data are scarce. However, this algorithm is often not ensured work well in real world application. Firstly, co-training requires the data set can be splits two views, and satisfy the sufficiency and independence assumptions. In practice, those conditions are not easy to achieve. Secondly, although co-training usually selects high confidence instances that are labeled with the estimated class labels and add them to the training sets, which does not ensured these selected high confidence instances are more valuable for improving the predictive accuracy. In the paper, a semi-supervised algorithm combining co-training with active learning was proposed, which can utilize the benefits of the two algorithms and reduce the annotation effort of operators.

3. Combining co-training with active learning

In this section we provide a high-level description of the semi-supervised learning algorithm, and its framework can be described as Fig. 1.The algorithm (SSLCA) which combines co-training with active learning can be divided into three steps: firstly, the labeled data was split into two views in order to apply the standard two-view co-training, which learns the classifier h_1 and classifier h_2 based solely on the two views of labeled data; secondly, the unlabeled data also was split into two views for estimating their confidence using separate classifier; thirdly, the most reliable instances or informative instances were selected based on some strategy. The most informative instances were chosen according to two criterions of high confidence and nearest neighbor, and then were put into another pool for further annotation.

3.1. Confidence estimation methods

3.1.1. Naive Bayes method

Naïve Bayes forms maximum a posteriori estimates for the class conditional probabilities for each feature from the labeled training data D. The prior probabilities of each class are calculated in a similar fashion by counting over instances. Define $P(c_j)$ denotes the probabilities of class c_j, and $|D|$ denotes the number of instances in training data:

$$P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j)}{|D|}$$

Then the posteriori probabilities estimates for each instance are calculated according to the independence assumption, define a_i denotes the each feature in each instance, n denotes the number of the features:

$$P(c_j|a_i) \propto P(c_j)P(a_i|c_j) = P(c_j)\prod_{i=1}^{n} P(a_i|c_j)$$

3.1.2. Expectation Maximization method

Expectation Maximization (EM) is an iterative statistical technique for maximum likelihood estimation in problems with