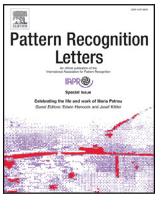
Pattern Recognition Letters 68 (2015) 15–21

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



A robust semi-supervised learning approach via mixture of label information[☆]

Yun Yang[a], Xingchen Liu[b,*]

[a] National Pilot School of Software, Yunnan University, Guandu, Kunming, Yunnan, China
[b] School of Computer Science and Technology, Tianjin University, No.92, Street Weijin, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 27 December 2014
Available online 24 August 2015

Keywords:
Semi-supervised learning
Clustering
Classification

ABSTRACT

Due to the fact that limited amounts of labeled data are normally available in real-world, semi-supervised learning has become a popular option, where we expect to use unlabeled data information to improve the learning performance. However, how to use such unlabeled information to make the predicted labels more reliable remains to be a key for any successful learning. In this paper, we propose a semi-supervised learning framework via combination of semi-supervised clustering and semi-supervised classification. In our approach, the predicted labels are selected by both the constrained *k*-means and safe semi-supervised SVM (S4VMs) to improve the reliability of the predicted labels. Extensive evaluations on collection of benchmarks and real-world action recognition datasets show that the proposed technique outperforms the others.

1. Introduction

Generally, machine learning can be classified into two categories: supervised learning and unsupervised learning. Supervised learning analyzes the labeled training data and produces an inferred function that predicts the relationship between the data and the class label, while unsupervised learning is focused on finding hidden structure/clusters inside unlabeled data sets. Supervised learning always requires sufficient labeled training data in order to establish a classifier with satisfactory capability in generalization. In real-world applications, however, most of the data are unlabeled and manually annotating such data is often infeasible. To make use of these data, semi-supervised learning has been developed to learn a classifier by both unlabeled and labeled data. Since semi-supervised learning requires less human effort yet has the potential to offer higher accuracy, exploiting unlabeled data to improve the learning performance has received enormous attentions from the machine learning community over recent years.

In semi-supervised learning, the hidden information is exploited from unlabeled data to support constructing a good classifier, and a number of approaches have been proposed before. Bennett and Demiriz [1] introduced a semi-supervised support vector machine (S^3VM), in which both training and testing datasets are exploited. To improve the efficiency, the so-called transductive SVM (TSVM) [2] tried to find a hyperplane, which is far away from the unlabeled points. While the Laplacian SVM [3] is developed to exploit the manifold structure of data via the Laplacian graph, both labeled and unlabeled data are mapped into the connected graph, where each example is represented as a vertex, and an edge connecting two vertices is created if they have a high similarity. In this approach, the ultimate goal is to find a class label for those unlabeled data as such that their contradiction between both the supervised data and the underlying graph structure are minimized. By using the knowledge of the means of the class labels, Yu-Feng Li et al. [4] proposed a new S^3VM, which is closely related to the supervised SVM with known labels on all the unlabeled data. After that, they proposed a safe semi-supervised support vector machines (S4VMs) [5]. Unlike S^3VMs, which typically focus on approaching an optimal low-density separator, S4VMs try to exploit multiple low-density separators in such a way that the risk of identifying the poor separator with unlabeled data is reduced. Recently, semi-supervised learning techniques are widely used in many real-world applications. Hoi et al. [6] shows a “Collaborative Image Retrieval” (CIR) by semi-supervised distance metric learning. Xiao Liu et al. [7] presented a semi-supervised splitting method to build up a Random Forest (RF). Yong Luo et al. [8] proposed a manifold regularized multi-task learning algorithm, which was reported to improve the performance of semi-supervised multi-label image classification, and Thorsten Joachims [2] reported a Transductive Support Vector Machines for text mining. Generalized classification has three problems of machine learning: recognition, taxonomy, and semi-supervised learning. Borisova and Zagoruiko proposed the FRiS-TDR algorithm, they solved all the three problems by examining them as special cases of the generalized classification problem. [9]

[☆] This paper has been recommended for acceptance by M.A. Girolami.
[*] Corresponding author. Tel.: 86 18322076887.
E-mail address: xcliu@tju.edu.cn (X. Liu).

http://dx.doi.org/10.1016/j.patrec.2015.08.009
0167-8655/© 2015 Elsevier B.V. All rights reserved.

In this work, we aim to develop a simple framework for semi-supervised learning that on one hand is easy to implement, and on the other it is guaranteed to improve the generalization performance of learning process. The main idea of the proposed approach is combining semi-supervised clustering and semi-supervised classification algorithms to improve the reliability of predicted labels. In other words, the predicted labels have to be confirmed by both the learning algorithms. Correspondingly, our contributions can be highlighted as follows:

- Via combining the strength of semi-supervised clustering and semi-supervised classification, a novel hybrid learning approach is proposed to reveal the intrinsic class structure of the input dataset.
- By considering the agreement of class structure obtained from different semi-supervised learning algorithms, an optimal selection of the predicted label is achieved to improve the learning performance.

The rest of this paper is organized as follows. We describe the semi-supervised clustering and classification algorithms related to our approach in Section 2. Section 3 presents our approach in detail. Section 4 reports the simulation test results on various datasets. Section 5 discusses issues relevant to our approach and finally the conclusions are drawn in Section 6.

2. Related works

In this section, we describe two existing semi-supervised learning algorithms, constrained k-means and S4VMs, to pave the way for introduction of our approach.

2.1. *Constrained* K-*means*

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). K-means [10] is a popular clustering algorithm that has been used in a variety of application domains. But in real-world applications, it is not enough that cluster samples have little structure knowledge, therefore background information is often needed at the same time. Constrained k-means algorithm [11] incorporates the background knowledge in the form of instance-level constraints, which is superior to the traditional clustering algorithm to obtain prior knowledge of which instances should be or should not be grouped together. There is a pair of constraints in this algorithm, *Must-link* and *Cannot-link*. Must-link requires that two instances have to be in the same cluster, and Cannot-link requires that two instances must not be placed in the same cluster.

Given a data set $V \in R^{m\times n}$, a labeled set $L \in R^{l\times n} \subseteq D$, m and l are the number of instances in V and L respectively. When the data set V is clustered, we produce a matrix, $\mathrm{Con} \in D^{l\times l}$, according to the set L, and the element $\mathrm{Con}(i, j)$ is defined as follows:

$$\mathrm{Con}(i,j) = \begin{cases} 1, & \mathrm{label}(i) = \mathrm{label}(j) \\ -1, & \mathrm{label}(i) \neq \mathrm{label}(j) \\ 0, & \mathrm{else} \end{cases} \tag{1}$$

where $1 \leq i, j \leq l$, $\mathrm{label}(i)$ and $\mathrm{label}(j)$ are the labels of the ith and the jth instance in L. Then cluster V, as k-means, ensures that two instances are put in the same cluster if the value of $\mathrm{Con}(i, j)$ is 1, and in different clusters if the value is –1.

2.2. *Safe semi-supervised SVM (S4VMs)*

In contrast to S^{3}VMs which find an optimal low-density separator, S4VMs try to exploit multiple low-density separators. As shown in Fig. 1, there are usually more than one large-margin low-density separators, while it is difficult to select the optimal one by giving a limited number of labeled data. Though these candidates all separate the labeled data well, not everyone suits the unlabeled data. Therefore, an inappropriate selection will badly influence the learning outcome. In this case, S4VMs considers all the low-density separators as the candidates.

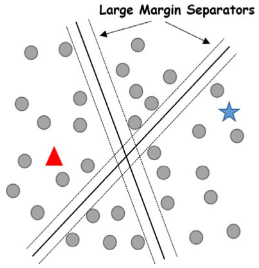


Fig. 1. Multiple large-margin low-density separators coincide well with labeled data.

Given the predictors of multiple low-density separators $\{\tilde{y}_t\}_{t=1}^{T}$, S4VMs intend to find $\boldsymbol{y} \in \{\pm 1\}$ that maximizes the performance improvement over the inductive SVM. Such optimization problem is formulated as follows:

$$\max_{\mathrm{y}\in\{\pm 1\}^{\mathrm{u}}} \mathrm{earn}\left(y, y', y^{\mathrm{svm}}\right) - \lambda \mathrm{lose}\left(y, y', y^{\mathrm{svm}}\right) \tag{2}$$

where λ is a trading-off parameter which decides how much risk is undertaken, $\boldsymbol{y}'$ is the ground-truth for unlabeled data, $\boldsymbol{y}^{\mathrm{svm}}$ denotes the predictions made by the inductive SVM on unlabeled data. Finally, $\mathrm{earn}(y,y',y^{\mathrm{svm}})$ and $\mathrm{lose}(y,y',y^{\mathrm{svm}})$ are defined as the increased and decreased accuracies in comparison with the inductive SVM, respectively.

As y' is unknown, it is quite difficult to solve (2). Assuming that the ground-truth boundary $\boldsymbol{y}'$ can be obtained by a low-density separator in $\{\tilde{y}_t\}_{t=1}^{T}$, i.e., $y' \in M = \{\tilde{y}_t\}_{t=1}^{T}$ optimization of the worst-case improvement over the inductive SVM can be formulated as:

$$\begin{aligned} \bar{y} &= \arg\max_{y\in\{\pm 1\}^u} \min_{\tilde{y}\in M} J(y, \tilde{y}, y^{\mathrm{svm}}) \\ J(y, \tilde{y}, y^{\mathrm{svm}}) &= \mathrm{earn}\left(y, y', y^{\mathrm{svm}}\right) - \lambda \mathrm{lose}\left(y, y', y^{\mathrm{svm}}\right) \end{aligned} \tag{3}$$

Without loss of generality, let $J(y, \tilde{y}, y^{\mathrm{svm}}) = c_t' y + d_t$. Eq. 3 can be re-defined as:

$$\max_{y\in\{\pm 1\}^u} \theta \ \mathrm{s.t.}\ \theta \leq c_t' y + d_t, \quad \forall t = 1, \cdots, T. \tag{4}$$

To solve Eq. 4, a convex linear programming has to be solved first by relaxing the integer constraint of $\mathbf{y}$ in Eq. 4 to $[-1,1]^u$, and then project it back to integer solution with minimum distance. By introducing dual variables α for constraints in Eq. 4, it can be re-defined as:

$$\max_{y\in\{\pm 1\}^u} \min_{\alpha\cdot 1=1, \alpha\geq 0} \sum_{t=1}^{T} \alpha_t (c_t' y + d_t) \tag{5}$$

Here α_t can be regarded as a probability that $\tilde{y}_t$ reveals the ground-truth solution. From this, if prior knowledge about the probabilities α is known, one can learn the optimal y according to the target in (Eq. 5) using the prior knowledge α.

3. Description of our approach

In this section, we illustrate the framework of our approach described in Algorithm 1. In fact, the proposed approach is designed to find out an optimal learner which not only minimizes classification errors on the labeled data, but also must be compatible with the input data space by predicting the class structure of unlabeled data. Initially we apply both S4VMs and constrained k-means on the target dataset which consists of labeled and unlabeled data. Then we can obtain labels of all unlabeled data via the two algorithms. For unlabeled set, if their predicted class and cluster labels are the same, then