Knowledge-Based Systems 83 (2015) 58–65

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Prior class dissimilarity based linear neighborhood propagation

Chun Zhang [a,b,*], Shafei Wang [c], Dongsheng Li [a,b], Junan Yang [a,b], Hao Chen [d]

[a] Electronic Engineering Institute, Hefei, China
[b] Key Laboratory of Electronic Restriction, Hefei, China
[c] The Northern Institute of Electronic Equipment of China, Beijing, China
[d] The Space Information Transmission Research Centre, Beijing, China

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form 28 February 2015
Accepted 13 March 2015
Available online 23 March 2015

Keywords:
Semi-supervised learning
Classification
Graph-based method
Linear neighborhood propagation
Prior information
Dissimilarity

ABSTRACT

The insufficiency of labeled training data for representing the distribution of entire dataset is a major obstacle in various practical data mining applications. Semi-supervised learning algorithms, which attempt to learn from both labeled and unlabeled data, provide possibilities to solve this problem. Graph-based semi-supervised learning has recently become one of the most active research areas. In this paper, a novel graph-based semi-supervised learning approach entitled Class Dissimilarity based Linear Neighborhood Propagation (CD-LNP) is proposed, which assumes that each data point can be linearly reconstructed from its neighborhood. The neighborhood graph of the input data is constructed according to a certain kind of dissimilarity between data points, which is specially designed to integrate the class information. Our algorithm can propagate the labels from the labeled points to entire data set using these linear neighborhoods with sufficient smoothness. Experiment results demonstrate that our approach outperforms other popular graph-based semi-supervised learning methods.

1. Introduction

During the last years, learning from both labeled and unlabeled samples, known as semi-supervised learning (SSL), has emerged as a booming direction in machine learning research. Detailed survey of its related literatures presented in [1].

As a major family of semi-supervised learning, graph-based methods have attracted more and more research and have been widely applied in many areas, such as text categorization [2], image retrieval [3], and image annotation [4,5].

Encouraging results have been reported when samples have clearly intrinsic structure and the test data are well sampled, Nevertheless, as can be seen in following sections of this paper, these algorithms are not so powerful when confronted with different class overlapping and data distributed imbalance. They may cause the choice of the neighbors to be unreasonable and destroy label smoothness when constructing the graph in these approaches.

In this paper, we exploited the prior class information in the framework of graph-based semi-supervised learning and proposed a novel method named Class Dissimilarity based Linear Neighborhood Propagation (CD-LNP). Unfamiliar with traditional graph-based semi-supervised learning schemes which mentioned above, CD-LNP utilizes the class labels of the input data to guide the learning process. Thus, the interclass dissimilarity is definitely larger than intraclass dissimilarity, which is a superior property for classification.

The rest of this paper is organized as follows. In Section 2, we briefly introduced traditional graph-based semi-supervised learning schemes and analyzed their limitations; and the proposed CD-LNP strategy was detailed in Section 3. In Section 4, experiments are reported. Finally, in Section 5, conclusions are drawn and several issues for future work are indicated.

2. Related works

Graph-based schemes are typical approaches of semi-supervised learning [6], such as FAS (Frequent Approximate Subgraph) in [7] and DLP (Dynamic Label Propagation) in [8]. In these methods, labeled and unlabeled sample points are first organized as the nodes of a graph, of which the edge connecting two nodes directly has a weight proportional to the proximity of these two sample points. Then, labels are "propagated" along the weighted edges from labeled nodes to unlabeled ones, in order to get predictions of unlabeled data.

In Table 1, we briefly reviewed four recent graph-based semi-supervised learning algorithms. Further research about these

* Corresponding author at: Electronic Engineering Institute, Hefei, China. Tel.: +86 13956049657.
E-mail address: zhangchuncwins@163.com (C. Zhang).

http://dx.doi.org/10.1016/j.knosys.2015.03.011
0950-7051/

Table 1
A summary of resent graph-based semi-supervised learning algorithms.

Method	Publication	Superiority	Drawback
Minimum Cut (Mincut [9])	IEEE Transactions on Pattern Analysis and Machine Intelligence	A fast new fully dynamic algorithm for the mincut problem. It can be used to efficiently compute MAP solutions for certain dynamically changing MRF models	A graph may have many minimum cuts and the mincut algorithm produces just one, typically the "leftmost" one using standard network flow algorithms
Local and Global Consistency (LGC [10])	ICML 2009	A probabilistic framework for modeling both the topical and geometrical structure of the dyadic data that explicitly takes into account the local manifold structure	It fails take into account the geometry of the spaces where the objects (either column or row objects) reside. The learned probability distributions are simply supported on the ambient spaces
Gaussian Random Field (GRF [11])	Pattern Recognition and Image Analysis	Combining Gaussian processes with randomized decision forests to enable fast learning. An important advantage is its simplicity and ability to directly control the tradeoff between classification performance and computation speed	The underlying random field gives a coherent probabilistic semantics to this approach, but this paper has concentrated on the use of only the mean of the field, which is characterized in terms of harmonic functions and spectral graph theory
Linear Neighborhood Propagation (LNP [12])	Neurocomputing	This method highlights the role of those samples in that sparse neighborhood, meanwhile; eliminates the role of those samples out of that sparse neighborhood. The adapting graph was constructed and each edge was weighed	The strategy was helpless when face data regions with overlapping class and imbalance distribution. They may cause the choice of the neighbors to be unreasonable and destroy label smoothness when constructing the graph

schemes can be found in [9–12] and references therein. We will discuss all these four methods later and focused on LNP strategy.

LNP utilizes the basic assumption of local linear embedding that each sample can be reconstructed by its neighboring samples linearly, and further assumes that label of the sample can be reconstructed by labels of its neighboring samples using same coefficients. The objective function for minimization is

$$\varepsilon = \sum_i \left\| \mathbf{x}_i - \sum_{i_j : \mathbf{x}_{i_j} \in N(\mathbf{x}_i)} w_{ii_j} \mathbf{x}_{i_j} \right\|^2 \tag{1}$$

where $N(\mathbf{x}_i)$ represents the neighborhood of sample $\mathbf{x}_i, \mathbf{x}_{i_j}$ is the jth neighbor of $\mathbf{x}_i$, and w_{ii_j} is the contribution of $\mathbf{x}_{i_j}$ to $\mathbf{x}_i$, satisfying the constraint $\sum_{i_j \in N(\mathbf{x}_i)} w_{ii_j} = 1, w_{ij} \geqslant 0$.

Let $\mathbf{F}$ denote the set of classifying functions defined on $\mathbf{X}, \forall f \in \mathbf{F}$ can assign a real value f_i to every point $\mathbf{x}_i$. The label of unlabeled data point $\mathbf{X}_u$ is determined by the sign of $f_u = f(\mathbf{X}_u)$. Iteration equation is constructed as

$$\mathbf{f}^{m+1} = \alpha \mathbf{W} \mathbf{f}^m + (1-\alpha)\mathbf{Y} \tag{2}$$

where $\mathbf{f}^m = \left(f_1^m, f_2^m, \cdots, f_n^m\right)^T$ is the prediction label vector at mth iteration. $0 < \alpha < 1$ is the fraction of label information which $\mathbf{x}_i$ receives from its neighbors.

3. The algorithm

Graph-based semi-supervised learning starts by constructing a graph from the training data. These algorithms often resort to KNN method when specifying the edge weights. Each vertex defines its k nearest neighbor vertices in Euclidean distance. Therefore, selecting precise neighbor is of great importance. However, in real application, there are always existing data regions with overlapping class and imbalance distribution. They may cause unreasonable choice of the neighbors and destroy label smoothness when constructing graph. For better comprehension of the limitation on KNN graph, we present a negative example in Fig. 1.

In Fig. 1, each triangle or circle represents one class. Faced with overlapping class described in Fig. 1(a), KNN method (set $k = 5$) will select a collection as the neighborhood of unlabeled instance x (represented by square). As shown in Fig. 1a, the neighbor collection $N(x) = \{a1, a2, a3, a4, a5\}$ and $\{a1, a2\} \subseteq Class1, \{a3, a4, a5\} \subseteq Class2$. Therefore, unlabeled instance x will be labeled as $Class2$. However, based on our visualized perception, x should be labeled as $Class1$. It's the limitation of KNN method which will easily trap into distance function (e.g., Euclidean distance) and ignore density distribution in feature space. Fig. 1(b) also exposes the limitation of KNN method when data region hold imbalance distribution.

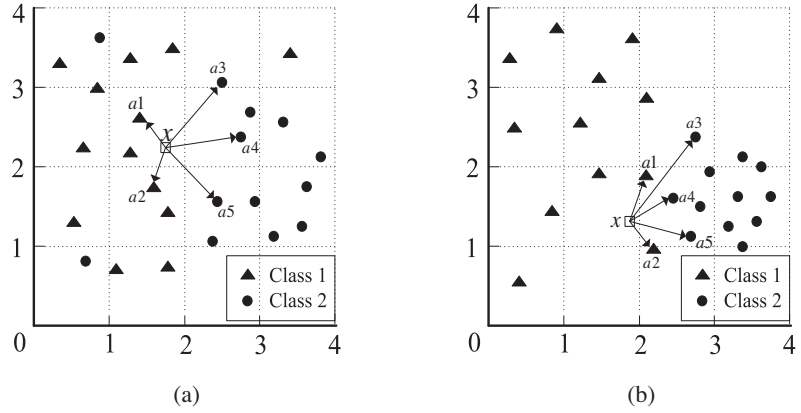

Fig. 1. Neighbor selection of KNN graph in different data region. Data region with overlapping class. (b) Data region with imbalance distribution.