Information Sciences 465 (2018) 265–284

Contents lists available at ScienceDirect

Information Sciences

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

Consensus rate-based label propagation for semi-supervised classification

Jaehong Yu[a], Seoung Bum Kim[b,*]

[a] Department of Population Health, School of Medicine, New York University 650 First Avenue, New York 10016, USA
[b] Department of Industrial Management Engineering, Korea University 145 Anam-Ro, Seoungbuk-Gu, Anam-dong, Seoul 02841, South Korea

ARTICLE INFO

Article history:
Received 19 September 2017
Revised 15 June 2018
Accepted 24 June 2018
Available online 10 July 2018

Keywords:
Consensus rate
Label propagation
Semi-supervised classification
Smoothness assumption

ABSTRACT

Label propagation is one of the most widely used semi-supervised classification methods. It utilizes neighborhood structures of observations to apply the smoothness assumption, which describes that observations close to each other are more likely to share a label. However, a single neighborhood structure cannot appropriately reflect intrinsic data structures, and hence, existing label propagation methods can fail to achieve superior performance. To overcome these limitations, we propose a label propagation algorithm based on consensus rates that are calculated by summarizing multiple clustering solutions to incorporate various properties of the data. Thus, the proposed algorithm can effectively reflect the intrinsic data structures, and yield accurate classification results. Experiments are conducted on various benchmark datasets to examine the properties of the proposed algorithm, and to compare it with the existing label propagation methods. The experimental results confirm that the proposed label propagation algorithm demonstrated superior performance compared to the existing methods.

1. Introduction

Traditional data analysis techniques can be generally categorized into unsupervised and supervised learning [33]. Unsupervised learning methods extract the implicit patterns and elicit natural cluster structures within the data without relying on any information from the output feature [38]. Conversely, supervised learning methods use both the input features and label information to classify or predict the output values of new observations. Regression methods can be used when the labels include real values, and classification methods can be used if the labels are in a set of categorical values [35]. In this study, we focus on the classification problems that classify the categorical labels in the response variable.

To secure superior performance of the classification models, a sufficient number of labeled observations should be used [42]. However, in many real situations, data analyses frequently suffer from an insufficiency of labeled observations because the annotation of labels requires expensive human labor and considerable time [41]. To address the lack of labeled observations, semi-supervised classification methods have been proposed [5,6,43]. A semi-supervised classification method entails a modeling process that uses both labeled and unlabeled observations [16,17]. A significant number of unlabeled training observations can be beneficial for establishing accurate classification models, because they facilitate the description

[*] Corresponding author.
E-mail address: sbkim1@korea.ac.kr (S.B. Kim).

https://doi.org/10.1016/j.ins.2018.06.074
0020-0255/© 2018 Elsevier Inc. All rights reserved.

of the intrinsic data structure [16,17]. Hence, semi-supervised classification methods have been widely used in many fields, including text classification [27], image segmentation [36], and bioinformatics [37].

To consider the unlabeled training observations for the classification model construction, semi-supervised classification models are designed under appropriate assumptions on the intrinsic data structures [6]. Among several assumptions on the data structure, the smoothness assumption is the most important one for semi-supervised classification methods. The smoothness assumption indicates that closer points tend to have similar class labels. Hence, the labels can smoothly change over the same clusters, and the decision boundary between clusters can be constructed by reflecting the geometric properties of the datasets [6,35].

To apply the smoothness assumption that describes that the points close to each other are more likely to share a label, existing semi-supervised classification methods utilize the *k*-nearest neighborhoods of observations. The *k*-nearest neighborhood-based methods can be categorized into label propagation [35,41,42,43], manifold regularization [5], and graph-based kernel design methods [44]. Among these, this study focuses on label propagation, which is one of the most widely used semi-supervised classification methods. Label propagation iteratively propagates the label information of labeled observations to their neighboring observations until the propagated labels are converged [42,43].

Gaussian field harmonic function (GFHF; [43]) and local and global consistency (LGC; [41]) are the most well known graph-based label propagation methods. These methods utilize the *k*-nearest neighborhood graph where the weights on the edges between observations are defined as the radial basis function (RBF) kernel. The GFHF computes a weight matrix from the graph structure, and transforms it into a transition probability matrix. The label information is then iteratively propagated to the unlabeled training observations using the transition probability matrix. During the iteration process, the GFHF maintains the initial label information. Conversely, the LGC iteratively propagates the label information using the normalized weight matrix. Unlike the GFHF algorithm, the LGC allows the initial label information to be changed during the iteration process. Hence, the LGC can effectively manage noisy labeled observations [6,41]. However, graph-based label propagation methods can only be used for classifying the labels of unlabeled observations in a training dataset, and thus they cannot classify the labels of new observations that do not belong to the training dataset [35].

For classifying the labels of new observations, Wang and Zhang [35] proposed linear neighborhood propagation (LNP). In the training phase, the LNP calculates locally linear reconstruction (LLR) weights for each observation that achieve the best reconstruction as a linear combination of its *k*-nearest neighborhoods. Then, the label information is iteratively propagated to the unlabeled training observations using the LLR weights. Finally, the LLR weights for new observations are computed, and the labels of the new observation are determined using the weights and label propagation results for the training observations.

Although these existing label propagation methods attain satisfactory results within the situations for which they were designed, no consensus exists regarding the best all-around performer in real-life situations. First, the classification accuracy of the existing methods is significantly influenced by hyperparameters, such as the neighborhood hyperparameter *k* [7,28], and the bandwidth parameter of the RBF kernel function [35]. Further, the single neighborhood hyperparameter *k* cannot appropriately describe the intrinsic data structures [2]. A small value of *k* cannot accommodate the global properties in datasets, whereas a large value of *k* frequently ignores the local patterns of the datasets [4,32].

To address these limitations of the existing label propagation methods, we propose a consensus rate-based label propagation (CRLP) algorithm. The proposed CRLP algorithm utilizes a consensus matrix that summarizes multiple clustering solutions as a similarity matrix [14,39,40]. To generate multiple clustering solutions, we use a random subspace and random-*K* selection ensemble scheme. The random subspace method generates multiple subspaces with randomly sampled features, and builds individual models in each subspace [21,23]. In each random subspace, cluster structures are constructed by a *K*-means clustering algorithm with randomly selected *K* values (*K* denotes the number of clusters). Using various numbers of clusters yields improved performances compared to a single number of clusters because individual random subspaces can have a different number of clusters [24–26]. The consensus rate is an element of the consensus matrix, which indicates the number of times a pair of observations belongs to the same cluster. Using the consensus rates as a weight for label propagation, the CRLP algorithm can achieve the following advantages:

- The CRLP algorithm is robust in that the values of the consensus rates are not significantly changed in spite of changes to the hyperparameters.
- The consensus rates help to reveal the intrinsic data structure because they are calculated from various properties of the data, such as multiple subspaces and different numbers of clusters.

The label information is iteratively propagated to other observations using the consensus rates. After classifying the labels of the unlabeled observations in a training set, the CRLP classifies the labels of new observations using the consensus rates between the new observations and training observations. Unlike the GFHF and LGC algorithms, the proposed CRLP algorithm can correctly classify the labels of new observations, because the consensus rates for the new observations can be easily computed.

The remainder of this paper is organized as follows. Section 2 briefly reviews the existing label propagation methods. Section 3 describes the proposed algorithm. Section 4 details the computational complexity analysis of the proposed CRLP algorithm. Section 5 presents the experimental results for various benchmark datasets to demonstrate the properties of the proposed algorithm and to compare its performance with those of the existing label propagation algorithms. Section 6 provides the concluding remarks.