Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised classification in stratified spaces by considering non-interior points using Laplacian behavior

Zohre Karimi, Saeed Shiry Ghidary*

Department of Computer Engineering & IT, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 24 January 2016
Revised 31 July 2016
Accepted 7 February 2017
Available online xxx

Communicated by Dr. H. Yu

Keywords:
Manifold
Semi-supervised
Laplacian
Stratified space

ABSTRACT

Manifold-based semi-supervised classifiers have attracted increasing interest in recent years. However, they suffer from over learning of locality and cannot be applied to the point cloud sampled from a stratified space. This problem is resolved in this paper by using the fact that the smoothness assumption must be satisfied with the interior points of the manifolds and may be violated in the non-interior points. Distinction of interior and non-interior points is based on the behavior of graph Laplacian in the ϵ-neighborhood of the intersection points. First, this property was generalized to ***K***NN graph representing the stratified space and then a new algorithm was proposed that penalizes the smoothness on the non-interior points of the manifolds by modifying the edge weights of the graph. Compared to some recent multi-manifold semi-supervised classifiers, the proposed method does not require neither knowing the dimensions of the manifolds nor large amount of unlabeled points to estimate the underling manifolds and does not assume similar properties for neighbors of all data points. Some experiments have been conducted in order to show that it improves the classification accuracy on a number of artificial and real benchmark data sets.

1. Introduction

The manifold-based semi-supervised classification has achieved promising success in many applications in recent years [8,9,10,14,37,40]. These algorithms assume that data resides on a single manifold [38,39] and impose the smoothness assumption along the neighborhood graph, which represents the manifold. It is common to approximate the geodesic distance by the local Euclidean distance in the neighborhood graph. This approximation is not accurate in the point clouds sampled from the intersecting multi manifolds, arising from the face that at the intersection points of manifolds the near points of the ambient space might be faraway in the intrinsic space. This will violate the smoothness assumption, which states that near points with high probability have the same label. So, label propagation across the points of the intersection regions will generate large errors. In many real applications, data lie on some intersecting manifolds having different dimensionality [24,32]. Intersecting manifolds are created when two classes representing different structures give rise to similar objects. For instance, in handwritten digit recognition "1" and "7" are similar objects. In face recognition, if we consider all patches of an image as a manifold, the similarity of the patches of two eyes from two different subjects is usually higher than that of an eye and a cheek of the same person [12]. As a consequence, in such applications the semi-supervised classification would be suffered by data points of intersection regions.

In the past few years, some methods have been proposed for dealing with high dimensional data lying on the intersecting manifolds. However, they have limitations imposed by some improper prior knowledge: (1) The assumption of knowing the number and dimensions of manifolds [34], (2) existence of similar neighborhood properties in all data points [15] and (3) The assumption of corresponding manifolds with the connected components of the graph [12].

In this paper, we propose a new semi-supervised classification method for classifying stratified space, i.e., some manifolds with different dimensionalities which may intersect together [17], by identifying interior points of manifolds using Laplacian behavior of data points. In the interior points of manifolds smoothness assumption is true and might be violated in the other points. Recent studies show that graph Laplacian, often used for applying the smoothness assumption and converge to Laplace–Beltrami operator, has different behavior in the ϵ-neighborhood of intersecting regions and tends to a first-order differential operator with different scaling. We exploit this property for modifying the edge weights of neighborhood graph. The main contributions in this paper are: (1) It is proved that the ***K***NN graph of representing

* Corresponding author.
E-mail addresses: z_karimi@aut.ac.ir (Z. Karimi), shiry@aut.ac.ir (S.S. Ghidary).

http://dx.doi.org/10.1016/j.neucom.2017.02.019
0925-2312/

stratified space can approximate ϵNN graph, and thus, different behavior of points near the non-interior points is also established in the $\boldsymbol{K}$NN graph, (2) a new algorithm is proposed which penalizes the high weights in the non-interior regions and (3) experimental evaluations confirm our claims by reducing errors of classification in the comparison with the state-of-the art methods. The proposed algorithm modifies the weight function and can be applied to any graph based learning algorithm assuming data lies on the stratified space.

Section 2 of this paper reviews related works. Section 3 exhibits our proposed method. In Section 4, the experimental results are presented followed by the conclusion in Section 5; the experiments show the effectiveness of the proposed method.

2. Related works

Manifold-based semi-supervised classification (MBSSC) methods apply smoothness assumption along the manifold, which expresses that two near points along the manifold have the same label with high probability. Application of this assumption requires definition of the data closeness model, which defines which points are near to one another and the label coupling model which defines how labels propagate on the near points [13]. Early works on the MBSSC focus on the label coupling models and for data closeness model simply is either $\boldsymbol{K}$NN or ϵNN graph, where data points are nodes of the graph and edges connect the nearest points regarding to Euclidean distance. They assume that data lie on a single smooth manifold where local Euclidean distance represents the geodesic distance along the manifold. The manifold regularization (MR) framework, one of the most representative works on semi-supervised classification, assumes that the support of the intrinsic data probability distribution is a compact manifold [2,25]. MR incorporates a regularization term to minimize the functional complexity of classifier function along the manifold. Since MR imposes the smoothness assumption along the neighborhood graph which is locally constructed based on Euclidean distance, MR cannot directly handle intersecting manifolds. Many studies verify the importance of the data closeness model and show that the success of the label propagation method mainly depends on how well the constructed neighborhood graph follows the underlying data manifold [11,15,35]. Some semi-supervised neighborhood graph construction methods have been proposed which use the discriminative power of the labeled data in addition to unlabeled data. For instance, some researches propose methods for kernel learning by transforming the eigenvalues of the Laplacian of the initial graph which formalize non-parametric kernel learning by maximizing kernel alignment to the labeled data [19,41]. In another work presented by Rohban et al. a supervised neighborhood graph construction method has been proposed which constructs a $\boldsymbol{K}$NN graph with a large enough value of $\boldsymbol{K}$ and then eliminates some additional edges using supervised SVM which classifies the graph's edges [26]. The SVM uses estimated labels of the Tikhonov regularization which are obtained from Laplacian matrix of initial graph, however it is at the risk of wrong label propagation, when the data lie on some intersecting manifolds.

The above mentioned methods are based on the assumption that the high dimensional data lie on a single low dimensional manifold. Recent studies show that the data can be better modeled by considering data unlabeled data points to lie in a stratified space that contains some intersecting manifolds with possibly different intrinsic dimensions which are properly glued together [16,17,20,22]. MBSSC methods are not efficient in the stratified space, since they suffer from locality over-learning where near points in the Euclidean distance appear to be far (from each other) in the intrinsic distance. Recently, some supervised, semi-supervised and unsupervised approaches have been proposed for learning from high dimensional data lying on intersecting multi manifolds. A multi-manifold discriminant analysis (MMDA) method under the Fisher discriminant framework has been proposed [36]. In MMDA the within-class graph can represent the sub manifold information while the between-class graph can represent the multi-manifold information. A multi-manifold sparse graph embedding algorithm (MSGE) for handling multi modal data has been proposed [21]. A non-parametric discriminant multi-manifold learning (NDMM) has been proposed by Li. et al. which reduces dimensionality of data by maximizing manifold-to-manifold distance and preserving manifold locality [20]. MMDA, MSGE and NDMM are supervised algorithms that use label information for resolving possible intersections. They all suffer from overfitting once labeled data are rare.

There are some multi manifold clustering algorithms which assume a linear structure for the intrinsic manifolds in the intersection points [32]. To handle the nonlinearity of real world data, Souvenir et al. have introduced K-manifolds which are an extension of ISOMAP [28,30]. It applies an EM algorithm and handles the nonlinear structure of data, however, it fails when intersecting manifolds are classified because the estimation of the face that geodesic distance is limited to the separated clusters. Li. et al introduced Spectral Clustering via Composite manifold and Local discriminant learning to estimate the intrinsic structure of data by assuming that the data lies in a convex combination of some pre given manifolds [22]. As unsupervised methods do not consider the labeled data, only related semi-supervised methods are discussed in the following paragraphs.

Goldberg et al. focused on theoretical analysis and proposes an algorithm, named Multi-Manifold Semi-Supervised Algorithm (MMSSA), which uses Hellinger distance for constructing the graph and then applies size-constraint spectral clustering to the graph in order to address multi manifold assumption [16]. A greedy procedure is used to select a subset of unlabeled data. Hellinger distance is sensitive to density of data and requires large unlabeled data to represent real distances on intersecting regions [34]. In the work performed by Xing et al. a geometrical similarity function based on local tangent space and principal angles has been introduced for multi-manifold semi-supervised Gaussian mixture model (M2SGMM) which models nonlinear manifolds by a Gaussian mixture model [34]. This method assumes that the number of manifolds is known and all have the same dimension, which is not a real assumption. Moreover, the computation of them is an open problem [5,6] and estimation of intrinsic dimension is very challenging, also according to statistical learning theory [31], the capacity and generalization capability of a given classifier may depend on the intrinsic dimensions [5].

Fang et al. [12] proposed a semi-supervised classification algorithm which presents a semi-supervised graph construction method and considers the geodesic distance on the graph as a kernel. This algorithm, then, gives a regularized regression model based on this kernel using both local and label information in order to find the low dimensional representation of data, and finally it applies a nearest neighbor classifier. However, over-learning of locality has not been considered and manifolds are specified by the connected components of the graph. Ensemble manifold regularization (EMR) is designed to automatically target the intrinsic manifold structure of data [15]. It assumes that the optimal manifold lies in the convex hull of some initial manifolds and tries to find their suitable combination. It is worthy to mention that the optimal solution of EMR reaches the similar neighborhood properties for all data points. It means that the locality of data points in the intersection points and other points have the same impact on the label propagation, however it is not a correct assumption.

Please cite this article as: Z. Karimi, S.S. Ghidary, Semi-supervised classification in stratified spaces by considering non-interior points using Laplacian behavior, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.02.019