Semi-Supervised Classification in Stratified Spaces by considering non-interior points using Laplacian Behavior

Zohre Karimi
Department of Computer Engineering & IT
Amirkabir University of Technology
Tehran, Iran
z_karimi@aut.ac.ir

Saeed Shiry Ghidary
Department of Computer Engineering & IT
Amirkabir University of Technology
Tehran, Iran
shiry@aut.ac.ir

***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 this fact that the smoothness assumption must be satisfied with the interior points of the manifolds and may be violated in the non-interior points. This fact is based on the property of graph Laplacian in the ϵ-neighborhood of the intersection points. We first generalize this property to KNN graph representing the stratified space and then propose a new algorithm 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.

Keywords- manifold, semi-supervised, Laplacian, Stratified Space.

1. INTRODUCTION

The manifold-based semi-supervised classification has achieved promising success in many applications in recent years [8, 21, 32]. These algorithms assume that data resides on a single manifold [30, 31] and impose the smoothness assumption along the neighborhood graph, which represents the manifold. The geodesic distance has been approximated by the local Euclidean distance in the neighborhood graph. This approximation is not accurate in the point clouds sampled from the intersecting multi manifolds. That's because near points in the ambient space may be far in the intrinsic space, which occurs at the intersection points of manifolds. The smoothness assumption, which expresses near points have the same label with high probability, has been violated in these points. So, label propagation across these points in the intersection regions propagates large errors. In many real applications, data lie on some intersecting manifolds with different dimensionality [19, 26]**.** Intersecting manifolds are created when two classes representing different structures give rise to similar objects. For instance, in handwritten digit recognition "2" and "3" are similar objects, in face recognition 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, when we consider all the patches of an image as a data manifold [11]. As a consequence, semi-supervised classification by considering these points is invaluable and increases accuracy.

In the last years, some methods have been proposed for dealing with high dimensional data lying on the intersecting manifolds. However, they have limited by some improper prior knowledge: (1) The assumption of knowing the number and dimensions of manifolds [27], (2) similar neighborhood properties in all data points [13] and (3) applying regularization on the KNN graph without regard to non-interior points of the manifolds [11] are the main challenges.

In this paper, we propose a new semi-supervised classification method for classifying stratified space, roughly speaking some manifolds with different dimensionalities [15], by considering non-smooth points in the construction of graph. Smoothness assumption is true in the interior points of manifold and may be violated in the other points. Recent studies show that graph Laplacian, which is always 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. Our main contribution is: (1) we prove that the different behavior of points near the non-interior points is also established in the KNN graph, (2) a new algorithm is proposed which penalizes the high weights in the non-interior regions and (3) experimental evaluation confirm our claims by decreasing the error 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 base learning algorithm which assumes data lies on the stratified space. We evaluate it on the semi-supervised classification problem by applying the manifold regularization framework.

Section 2 reviews related works. Section 3 details our proposed method. In section 4 the experimental results are presented and conclusions are made in section 5. The experiments show the effectiveness of the proposed method.

2. Related Work

The manifold-based semi-supervised classification (MBSSC) methods apply the smoothness assumption along the manifold, which expresses that two near points along the manifold have the same label with high probability. Imposing this assumption require the definition of the data closeness model, which defines what points are near and the label coupling model, which defines how labels propagate on the near points [12]. Early works on the MBSSC focus on the label coupling models and data closeness model simply is either KNN or ϵNN graph, where data points are nodes of the graph and edges connect the nearest points w.r.t. 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 [2], one of the most representative works on semi-supervised classification, assumes that the support of the intrinsic data probability distribution is a compact manifold. MR incorporates a regularization term, to minimize the functional complexity along the manifold. Since MR impose the smoothness assumption along the neighborhood graph which locally constructed based on Euclidean distance, MR cannot handle directly intersecting manifolds. However, many studies verify that the data closeness model is very important and the success of the label propagation method mainly depends on how well the constructed neighborhood graph follows the underlying data manifold [10, 13, 28]. Some semi-supervised neighborhood graph construction methods by using the discriminative power of labeled data in addition to unlabeled data have been proposed. Methods of kernel learning have been proposed to build a kernel by transforming the eigenvalues of the Laplacian of the initial graph. The non parametric kernel learning has been formalized by maximizing the alignment to labeled data [17, 33]. In [22] a supervised neighborhood graph construction method has been proposed which constructs a KNN graph with large enough K and then delete some additional edges using supervised SVM which classifies the graph's edges. The SVM uses the estimated labels of the Tikhonov regularization using the Laplacian matrix of initial graph which is at the risk of wrong label propagation, where the data lie on the 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 lie in the stratified space that contains some intersecting manifolds with possibly different intrinsic dimensions which nicely glued together [14, 15]. MBSSC methods are not efficient in the stratified space, since they suffer from over-learning locality, where near points in the Euclidean distance may be far in the intrinsic distance. Recently, some supervised, semi-supervised and unsupervised approaches have been proposed that aim to learn from high diminution data when it lies on multi intersecting manifolds. A multi-manifold discriminant analysis (MMDA) method under the fisher discriminant framework has been proposed [29]. In MMDA, the within-class graph can represent the sub manifold information, while the between-class graph can represent the multi-manifold information. It is a supervised algorithm that overfits when labeled data are rare.

There are some multi manifold clustering algorithms which assume the intrinsic manifolds in the intersection points have a linear structure [26], however, the real world data often have nonlinear structures. K-manifolds [24]is an extension of ISOMAP which applies an EM algorithm and handles the nonlinear structure, however, it is failing when classifying intersecting manifolds because the estimation of geodesic distance is limited to separated clusters. Unsupervised methods didn't consider the labeled data. In what follows, we discuss about related semi-supervised methods.

To address the multi manifold, Goldberg et al. [14] focus on the theoretical analysis and proposes an algorithm, namely MMSSA (Multi-Manifold Semi-Supervised Algorithm), which uses the Hellinger distance for constructing the graph and then applies size-constraint spectral clustering to the graph. 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 [27]. A geometrical similarity function based on local tangent space and principal angles has been introduced in multi-manifold semi-supervised Gaussian mixture model (M2SGMM) and nonlinear manifolds have been modeled by a Gaussian mixture model [27]. This method assumes that the number of manifolds is known and has the same dimension, which is not a real assumption and the computation of them is an open problem [5, 6]. Besides, intrinsic dimension reduction is very important since, according to statistical learning theory [25], the capacity and generalization capability of a given classifier may depend on the intrinsic dimension [5].

Fan et al. [11] 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, then gives the regularized regression model based on this kernel using both the local and label information to find the low dimensional representation of data. Finally, it applies 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 [13]. It assumes that the optimal manifold lies in the convex hull of some initial manifolds and tries to find the suitable combination of them. It is noticeable 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, which is not a correct assumption.

3. Proposed algorithm

3.1. Problem Setup

Let $\chi = \{\mathrm{x_i} \epsilon \mathrm{R^d}, \dots \mathrm{i} = 1, \cdots, \mathrm{n}\}$ be the whole data set consisting two subsets $\chi = \{\chi_\mathrm{l}, \chi_\mathrm{u}\}$, where without loss of generality, $\chi_\mathrm{l} = \{\mathrm{x_1}, \cdots, \mathrm{x_l}\}$ is the labeled subset and $\chi_\mathrm{u} = \{\mathrm{x_{l+1}}, \cdots, \mathrm{x_{l+u}}\}$ is the unlabeled subset. Given a binary classification task, $\mathrm{c_i} \epsilon \{+1, -1\}$ for $\mathrm{i} = 1, \cdots, \mathrm{l}$ are the partial available labels and $l \ll n$. $\mathrm{C_l} = \{\mathrm{c_i}\}_{\mathrm{i}=1}^{\mathrm{l}}$ is the label set of labeled data. The object of semi-supervised classification algorithms is to find the classification function, f, with the assumption that this data is sampled from the stratified space, where the labels vary smoothly on each manifold.

The underlying manifold of data is represented based on the local neighborhood relations, which often are constructed using the K nearest neighbor graph (the graph with one vertex per observed example, and arcs between K nearest neighbors). Matrix L =D-W, is the Laplacian matrix, where W is the data adjacency weight matrix, wherein each element $\mathrm{w_{i,j}}$ is the edge between two samples $\mathrm{x_i}$ and $\mathrm{x_{j,}}$ and D is a diagonal matrix where $\mathrm{D(i,i)} = \sum_j \mathrm{W}(x_i, x_j)$. One typical weight function is Gaussian, defined as $\mathrm{W_t(x_i, x_j)} = \mathrm{e}^{-\frac{\|\mathrm{x_i}-\mathrm{x_j}\|^2}{2\mathrm{t}}}$, where t is bandwidth.

If points are uniformly randomly sampled from a smooth manifold, then in the interior of the manifold, in the limit as n goes to infinity and t tends to 0 at an appropriate rate, the Laplacian operator, $\mathcal{L}_t$, would converge to the Laplace-Beltrami operator [1]:

$$\mathcal{L}_t f(\mathrm{x}) = \Delta f(\mathrm{x}) + \mathrm{O}(1) \qquad (1)$$

Where Δ is the Laplace-Beltrami operator of the manifold. The graph Laplacian applied to function f is computed as:

$$\mathrm{L}_t f(x) = \sum_{j=1}^{n} \mathrm{W}_t(\mathrm{x}, \mathrm{x}_j)[\mathrm{f}(\mathrm{x}) - \mathrm{f}(\mathrm{x}_j)] \qquad (2)$$

, where $\mathrm{L}_t f$ is the discontinuous form of $\mathcal{L}_t$.

3.2. The Proposed Semi-Supervised Algorithm

The traditional methods [2, 8] have been focused on well-separated manifolds while in many applications data can be better modeled by considering intersecting manifolds. Theses algorithms cannot effectively deal with such conditions because in the intersection of manifolds, the graph connects points belonging to different manifolds strongly, therefore the manifold regularization on the graph assigns similar labels to the intersection points with high probability leading to wrong propagation of labels. For example, the weights of the KNN graph of dollar sign data set are shown in Figure 1. The weights of edges connecting two manifolds (sign and S curve) have large values. Consequently, manifold regularization on this graph assigns similar labels to the points from different manifolds with high probability.

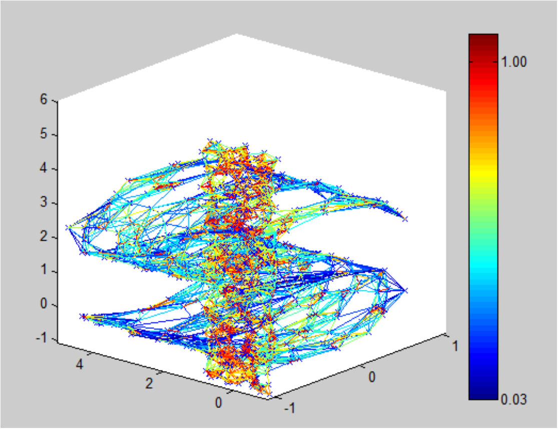


Figure 1. The edge weights of the *K*NN graph on dollar sign data set

To address the above issue, we propose a multi manifold regularization framework which decreases the importance of the non-interior connections in the label propagation. The proposed optimization problem can be formulated as the bivariate optimization problem which simultaneously estimates the weight edge values and classifier. This objective takes the following form:

$$min_{f\epsilon H_k, W' \geq 0} \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \|f\|_K^2 + \gamma_{I'} \sum_{i,j=1}^{n} w'_{ij} \left(\mathrm{f}(x_i) - \mathrm{f}(x_j)\right)^2 \qquad (3)$$

$$s.t \begin{cases} 0 \leq w'_{ij} < w_{ij} & x_i \ is\ not\ an\ interior\ point \\ w'_{ij} = w_{ij} & otherwise \end{cases}$$

where $H_k$is the *reproducing kernel Hilbert space* (RKHS); and V is a general loss function such as the hinge loss or the square loss, $\|f\|_K^2$ is a smooth penalty term in ambient space. Parameters γ_A and$\gamma_{I'}$ are used for trade off between the loss function and the ambient regularization term and intrinsic regularization term along the stratified space. The classifier smoothness along the stratified space estimated from the unlabeled data and is approximated by the last term. $\mathrm{w}'_{\mathrm{ij}}$ is the element of ith row and jth column of W'. To distinguish the interior points of the manifold from the rest, [14] proposed to use the Hellinger distance which detect changes in support, density, dimensionality or orientation of data and [27] proposed to compute the geometric similarity of points using the principal angle between the local tangent spaces. The first method needs large amount of data and the second assumes that the number of manifolds and the intrinsic dimension of it are known from the prior knowledge. To address those mentioned issues, we propose an approach for computing W' by exploiting the behavior of the Laplacian near the non-interior points [3]. We modify the graph such that non-interior points are also being considered.

Theoretical studies in recent years show that non-interior points containing intersection points, boundary and edges are important aspects of data. For a point x lying in the ϵ-neighborhood of non-interior points, we have

$$\mathcal{L}_t f(x) = \frac{1}{\sqrt{t}}\frac{\pi}{2}\partial_{\vec{n}} f(x) + O\left(\frac{1}{\sqrt{t}}\right) \quad (4)$$

where $\partial_{\vec{n}}$ is the unit outward normal to the point x and $\partial_{\vec{n}}\mathrm{f(x)}$ is the directional derivative of f in direction $\partial_{\vec{n}}$ [3].

According to the (4), $\mathcal{L}_\mathrm{t}\mathrm{f(x)}$ is of the order $\mathrm{O}\left(\frac{1}{\sqrt{\mathrm{t}}}\right)$, which is much larger than $\mathcal{L}_\mathrm{t}\mathrm{f(x)}$ of the interior points for the small number t, which is of the order $\mathrm{O}(1)$. Therefore, large values of $\mathcal{L}_\mathrm{t}\mathrm{f(x)}$ represent the points near to the non-interior points. We exploit the above property for modifying the neighborhood in the nearest neighborhood graph.

As mentioned, large values of $\mathcal{L}_\mathrm{t}\mathrm{f(x)}$ represent the points in the ϵ-neighborhood of non-interior points. So, this property could specify the interior points in the ϵNN graph. However, ϵNN is prone to generate disconnected or almost complete graph w.r.t. its threshold. KNN graph is robust to this problem and frequently is used in the manifold based methods [2, 8, 18, 22, 31]. Therefore, it is desirable to find a method to specify interior points in the stratified space, when it is presented by KNN graph.

If we prove there exist $\varepsilon > 0$ ($\varepsilon_{\min} < \varepsilon < \varepsilon_{\max}$) in KNN graph representing the stratified space, such that the neighborhood of $\mathrm{B_x}(\varepsilon_{\min})$ contains fewer than K points and the neighborhood of $\mathrm{B_x}(\varepsilon_{\max})$ contains more than K points with high probability, then it is shown that KNN graph contains all points in the neighborhood of $\mathrm{B_x}(\varepsilon)$. Asymptotic analysis proves the existence of ε. We have provided details of such proof in Appendix A.

Therefore, we conclude that with high probability larger values of $\mathcal{L}_\mathrm{t}\mathrm{f(x)}$ in KNN graph specify the non-interior points. Where smoothness assumption may be violated in these points, we assign a confidence to weights as

$$w'_{ij} = w_{ij} \times c_{cof}\,, \quad (5)$$

$$c_{cof} = \begin{cases} c_{dec} & x_i \ or \ x_j \ is \ not \ an \ interior \ point \\ 1 & otherwise \end{cases}$$

, where $0 < c_{dec} < 1$ is a confidence coefficient.

For a fixed W', (3) simplified to:

$$min_{f\epsilon H_k}\frac{1}{l}\sum_{i=1}^{l} V\left(f, x_i, y_i\right) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \quad (6)$$

, where $\mathrm{W}' = \mathrm{W_t}$. The minimizer of (6) is

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \quad (7)$$

, where K is the kernel of associated RKHS [2].

On the other hand, with a fixed f, we obtain W' by (5), where non-interior points are the top n_{dec} data points with largest values of Lf computed by (2); where n_{dec} is a predefined number. So, for optimizing (3), we compute iteratively (7) by modifying the weight matrix using (5). The proposed Algorithm is shown in Figure 2.

The proposed algorithm learns the function using Expectation Maximization (EM). It alternates between the assignments of data points to either interior or non interior points in E step and the function re estimation in M step. We propose three suitable candidates for initializing: (1) the available labels for labeled data and zero for unlabeled data, (2) the classification result of the manifold regularization and (3) random labeling. *We* initialize f by assigning all three candidates and then repeat step 2 of algorithm independently for each of them.

```
Input: $\chi$, $C_l$, $n_{dec}$, $c_{dec}$
Output: $f$
Step1: Initial $f$.
Step2: Repeat following steps until the objective function (5) is not decrement or reach predefined max number of iterations
    1. Compute Laplacian of $f$ by (4).
    2. Determine top $n_{dec}$ data points with largest value of $Lf$ as the non-interior points.
    3. Compute $W'$ by (7).
    4. Compute $f_{new}$ by (2)
    5. Set $f = f_{new}$
```

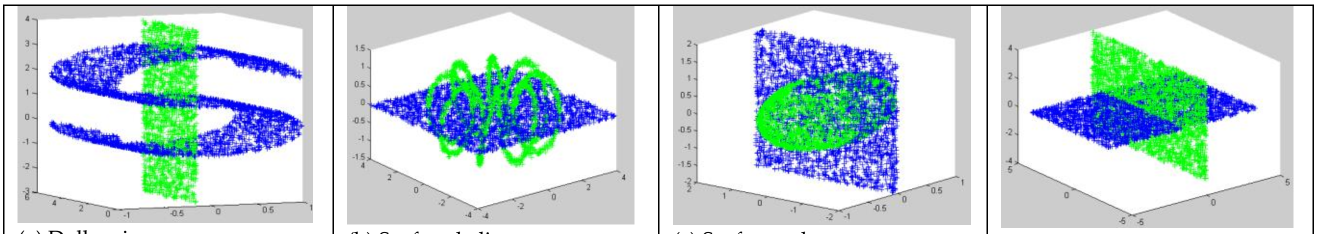
Figure 2. ***The proposed algorithm***

4. Experiments

We have conducted some experiments on artificial and real world data sets to evaluate our proposed algorithm. All experiments were repeated 10 times with random labeled data. The mean of the error rate and standard deviation have been reported for each experiment. We compare the proposed method with three graphs based semi-supervised classifiers: MR [2], M2SGMM [27] and EMR [13]. MR is based on a single manifold assumption; M2SGMM is based on multi-manifold assumption and EMR assumes that data lie in the linear combination of some pre given manifolds. We apply squared loss function to MR, EMR and the proposed algorithm. In all experiments c_{dec} is set to 0.0001. For fair comparison, we select λ_A and λ_I from ranges $\{.0001,0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8\}$ and $\{0.9,1,5,10,15,20\}$ respectively. The maximum number of iterations of proposed algorithm is set to 5.

4.1. Simulation on Synthetic Data

We have experimented with four synthetic datasets (Figure 3): (1) Dollar sign has the "S" manifold intersecting an "|" manifold, (2) Surface-helix containing two intersecting manifolds: one 1D toroidal helix and a surface, (3) Surface-sphere involving a sphere intersecting a surface, (4) Two intersecting planes containing two overlapping planes.



(a) Dollar sign

(b) Surface-helix

(c) Surface-sphere

(d) Two intersecting planes

Figure 3. Artificial datasets

We set $K = 5$, $t = 100$ and kernel width to 0.6 in all data sets. There are $n = 5000$ data in all artificial data sets, with equal number of samples in each class and n_{dec} set to 300. The mean error rate (%) and standard deviation on artificial datasets when $l = 100$ are presented in Table 1. Our algorithm clearly outperforms MR on all four data sets.

For more detailed analysis, we compute both the absolute of Laplacian of correct labels of all data, Lf, (Figure 4(a)) and the ratio of weight of neighbors which belong to different classes to the weight of all neighbors in KNN graph, W_d (Figure 4 (b)). As we see, in all data sets, the data points which have strong connections to the data points of other classes have the largest values of Lf as we expected from theoretical analysis.

Table 1. Error rate (%) and standard deviation of different distance functions on artificial data sets

Dataset	Error rate (Standard deviation)	
	MR	**The proposed algorithm**
Dollar Sign	13.05 (6.19)	**8.26 (2.09)**
Surface-helix	13.62 (3.5)	**10.68 (2.36)**
Surface-sphere	6.97 (1.65)	**5.8 (0.98)**
Two intersecting planes	2.99 (0.99)	**2.66 (0.89)**

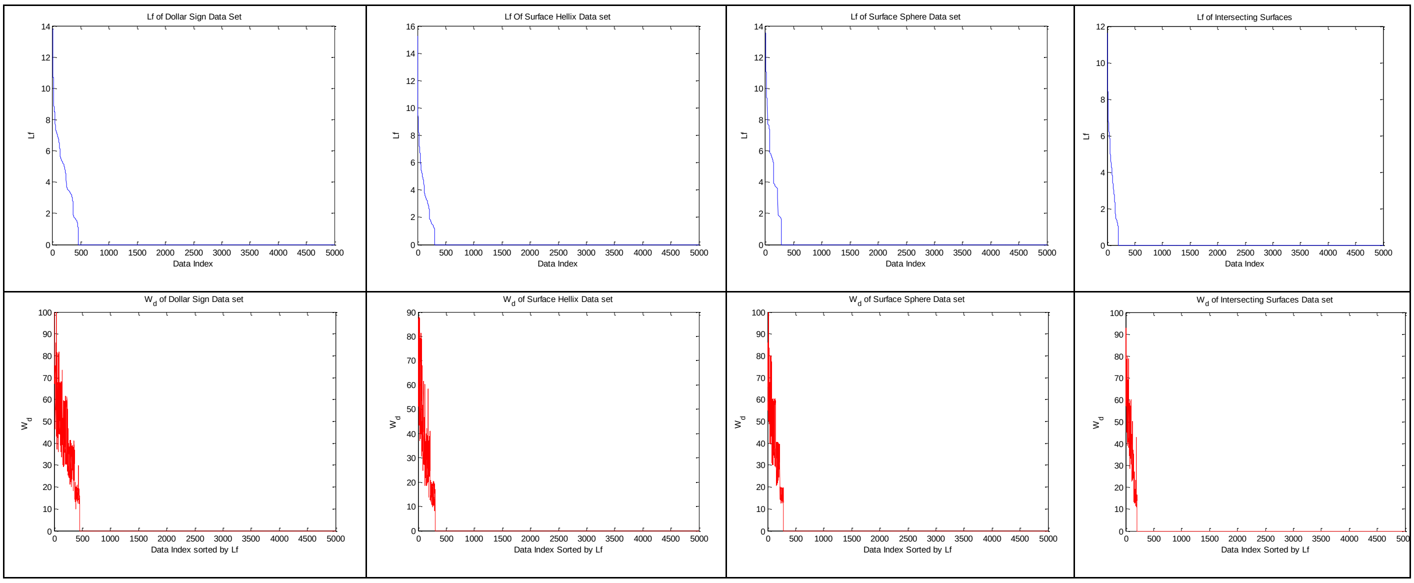


Figure4. (a) Top row presents the sorted absolute of the Laplacian of correct labels of the data, (b) Bottom row presents the ratio of weights of nearest neighbors which belongs to different classes to the weight of all neighbors (%) in KNN graph

4.2. Experiment on real data sets

In this section, we present the result of our proposed method on four datasets including a well known face data set, CBCL, and an object category recognition data set, COIL, a text categorization data set, WebKB and BCI data set [16]. These data sets are frequently used for the evaluation of semi-supervised methods [8, 11, 13]. The specification of them is given in Table 2.

Table 2. The specification of real world data sets

Dataset	***n*** (# of data)	***d*** (# of dimensions)	***l*** (# of labeled data)
COIL	*1500*	*241*	*150*
CBCL	*3000*	*361*	*300*
WebKB (Page)	*1051*	*3000*	*12*
BCI	*400*	*117*	*40*

COIL, the Columbia object image library contains a set of color images of different objects taken from different angles in steps of 5 degrees [20]. We downloaded the binary version of it [8]. CBCL data set is a set of 2429 face images and 4548 non-face images [7]. Each image has 19 × 19 pixels and is transformed into a 361-dimensional vector. We use a subset of it, which contains 1500 face and 1500 non-face images.

WebKB[1] is a subset of web documents of four universities, which belong to two categories of course and non-course. We consider the textual content of documents (page representation) and join it with link to other web pages pointing to documents to make a new representation (Page+Link). BCI (brain computer interface) data set contains 400 trails of imaginary movements with either the right hand or the left hand of a single person. Each trail is represented by 170 parameters about its EEG (electroencephalography) [16].

For WebKB, we use the same kernel parameters, graph weight function and graph weight parameter as [23]. For COIL and CBCL, the number of nearest neighbors in the KNN graph is selected from the smallest value which keeps the graph connected, K_{con}, and 2 K_{con} . The graph weight parameter is set to 100. Gaussian base kernels and Euclidean nearest neighbor graphs with Gaussian weights are used. The kernel width is set to 0.6.

The parameters of M2SGMM are set according to [27]: the number of components and the number of nearest neighbors of graph are set to $\left\lceil \frac{\mathrm{n}}{10\mathrm{N}} \right\rceil$ and $\lceil 1.5 \lceil \mathrm{Logn} \rceil \rceil$, respectively, and it is assumed that all underlying manifolds have the same dimension $0 < \mathrm{d_{in}} < 10$, where the original data are reduced to the 10-dimensional subspace by PCA. We tried with all possible values of $\mathrm{d_{in}}$ and the best result has been reported here. For fair comparison, graph weight parameters are set equal to the proposed method.

The parameters of candidate manifolds of EMR are set to $K = \{5,10,15,k'\}$, where $k'$is set to the value of K of the proposed algorithm to ensure fair comparison, $t = \{\left(\frac{\tau}{15}\right)^2, \left(\frac{\tau}{10}\right)^2, \left(\frac{\tau}{5}\right)^2, \tau^2, (5\tau)^2, (10\tau)^2, (15\tau)^2, (20\tau)^2, t'\}$, where $\tau = (\frac{1}{n^2}\sum_{i,j=1}^{n}\left\|x_i - x_j\right\|^2)^{-1}$ and $t'$is set to the value of t of the proposed algorithm and $\lambda_\mathrm{R} = 0.01\lambda_\mathrm{I}$ [13].

In the proposed method, $\mathrm{n_{dec}}$ is set to $\{\frac{n}{12}, \frac{n}{10}, \frac{n}{8}, \frac{n}{6}\}$. Table 3 presents the average accuracy of the best parameter configuration. As we see, the proposed method clearly outperforms MR, EMR and M2SGMM.

Table 3. Error RATE (%) AND Standard Deviation of different distance functions on real Datasets

Dataset	Error rate (Standard deviation)			
	MR	M2SGMM	EMR	Proposed Method
COIL20	8.09 (0.71)	9.07 (0.01)	3.56 (2.77)	**2.76 (2.13)**
CBCL	7.75 (0.76)	9.71 (0.02)	7.48 (0.47)	**6.88 (0.57)**
WebKB (Page)	13.87 (4.72)	7.10 (3 .56)	9.43(5.24)	**5.87 (1.40)**
WebKB(Page+Link)	10.14 (9.25)	5.12 (3.3)	7.43(4.94)	**3.4∧ (0.42)**
BCI	42.15 (2.49)	47.74 (3.06)	43.27 (2.59)	**41.42 (2.26)**

5. Conclusion

In this study, we have introduced a novel framework for classifying data residing on intersecting manifolds. Instead of forcing the impact of locality on label propagation, that is the source of large errors in the label prediction, we introduce a confidence coefficient for the connections between the points; the locality then need not be strictly enforced during the manifold regularization and may be violated in the non interior points of the manifolds. Our empirical studies reveal that the proposed method works better than M2SGMM, which uses the knowledge of the number of manifolds and their intrinsic dimensions and EMR, which represents the structure of data by the convex hull of some predefined manifolds. The last is because we do not constrain all connections to equally exploit the locality information and therefore avoid over learning of locality by giving more confidence to internal connections.

Acknowledgements

We would like to appreciate Xianglei Xing for offering the code of M2SGMM.

[1] WebKB is downloaded from http://vikas.sindhwani.org/manifoldregularization.html.

REFERENCES

[1] Belkin, Mikhail, and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural computation 15, no. 6*, 2003: 1373–1396.

[2] Belkin, Mikhail, Partha Niyogi, and Vikas Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *Journal of Machine Learning Research 7*, 2006: 2399-2434.

[3] Belkin, Mikhail, Qichao Que, Yusu Wang, and Xueyuan Zhou. "Toward Understanding Complex Spaces: Graph Laplacians on Manifolds with Singularities and Boundaries." *The 25th Annual Conference on Learning Theory (COLT 2012).* Edinburgh, Scotland, June 25-27, 2012. 1-36.

[4] Bernstein, Mira, Vin De Silva, John C. Langford, and Joshua B. Tenenbaum. *Graph approximations to geodesics on embedded manifolds.* Technical report, Department of Psychology, Stanford University, 2000.

[5] Campadelli, P., E. Casiraghi, C. Ceruti, and A. Rozza. "Intrinsic Dimension Estimation: Relevant techniques and a Benchmark Framework." *Mathematical Problems in Engineering*, 2015.

[6] Carter, Kevin M, Raich Raviv, and Alfred O Hero. "On local intrinsic dimension estimation and its applications." *IEEE Transactions on Signal Processing, vol 58, no. 2*, 2010: 650-663.

[7] CBCL Face Database, Philadelphia PA, USA. MIT Center For Biological and Computation Learning. 2001. http://www.ai.mit.edu/projects/cbcl.

[8] Chapelle, Olivier, Schölkopf,Bernhard, and Alexander Zien. *Semi-supervised learning.* Cambridge: MIT press, 2006.

[9] —. *The Semi-Supervised Learning Book.* 2006. http://www.kyb.tuebingen.mpg.de/ssl-book.

[10] de Sousa, Celso André R., Solange O. Rezende, and Gustavo EAPA Batista. "Influence of Graph Construction on Semi-supervised Learning." In *Machine Learning and Knowledge Discovery in Databases*, 160-175. Springer Berlin Heidelberg, 2013.

[11] Fan, M., Zhang, X., Lin, Z., Zhang, Z., Bao, H.. "A Regularized Approach for Geodesic Based Semi-Supervised Multi-Manifold Learning." *IEEE Transactions on Image Processing, Vol. 23, no. 5*, 2014: 2133-2147.

[12] Fang, Yuan, Kevin Chen-Chuan Chang, and Hady Wirawan Lauw. "Graph-based semi-supervised learning: Realizing pointwise smoothness probabilistically." *International Conference on Machine Learning (ICML).* 2014.

[13] Geng, Bo, Dacheng Tao, Chao Xu, Yichen Yang, and Xian-Sheng Hua. "Ensemble manifold regularization." *IEEE Transactions on Pattern Analysis and Machine Intelligence 34, no. 6*, 2012: 1227-1233.

[14] Goldberg, Andrew B., Xiaojin Zhu, Aarti Singh, Zhiting Xu, and Robert Nowak. "Multi-manifold semi-supervised learning." *International Conference on Artificial Intelligence and Statistics.* 2009. 169-176.

[15] Haro, Gloria, Gregory Randall, and Guillermo Sapiro. "Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds." *In Advances in Neural Information Processing Systems*, 2006: 553-560.

[16] Lal, Thomas Navin, et al. "Support vector channel selection in BCI." *IEEE Transactions on Biomedical Engineering, 51, no. 6*, 2004: 1003-1010.

[17] Lanckriet, Gert R.G., Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. "Learning the kernel matrix with semidefinite programming." *Journal of Machine Learning Research 5*, 2004: 27-72.

[18] Melacci, Stefano, and Mikhail Belkin. "Laplacian support vector machines trained in the primal." *The Journal of Machine Learning Research 12*, 2011: 1149-1184.

[19] Meng, Deyu, Yee Leung, Tung Fung, and Zongben Xu. "Nonlinear dimensionality reduction of data lying on the multicluster manifold." *IEEE Transactions on Systems, Man, and Cybernetics, Part B, vol 38, issue 4*, 2008: 1111-1122.

[20] Nene, Sameer A., Shree K. Nayar, and Hiroshi Murase. *Columbia object image library (COIL-100). Technical Report CUCS-006-96.* New York: Columbia University, 1996.

[21] Niyogi, Partha. "Manifold regularization and semi-supervised learning: Some theoretical analyses." *The Journal of Machine Learning Research 14.1*, 2013: 1229-1250.

[22] Rohban, Mohammad Hossein, and Hamid R. Rabiee. "Supervised neighborhood graph construction for semi-supervised classification." *Pattern Recognition 45, no. 4*, 2012: 1363-1372.

[23] Sindhwani, Vikas, Partha Niyogi, and Mikhail Belkin. "Beyond the point cloud: from transductive to semi-supervised learning." *In Proceedings of the 22nd international conference on Machine learning.* ACM, 2005. 824-831.

[24] Souvenir, Richard, and Robert Pless. "Manifold clustering." *Tenth IEEE International Conference on Computer Vision (ICCV 200), vol. 1.* IEEE, 2005. 648-653.

[25] Vapnik, V. N. *Statistical Learning Theory.* New York: John Wiley & Sons, 1998.

[26] Vidal, Rene, Ma, Yi, and Shankar Sastry. "Generalized principal component analysis (GPCA)." *IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 12* (1945-1959), 2005: 1945-1959.

[27] Xing, Xianglei, Yao Yu, Hua Jiang, and Sidan Du. "A multi-manifold semi-supervised Gaussian mixture model for pattern classification." *Pattern Recognition Letters 34, no.16*, 2013: 2118-2125.

[28] Yang, Li. "Building k-connected neighborhood graphs for isometric data embedding." *IEEE Transactions on Pattern Analysis and Machine Intelligence 28, no.5*, 2006: 827-831.

[29] Yang, Wankou, Changyin Sun, and Lei Zhang. "A multi-manifold discriminant analysis method for image feature extraction." *Pattern Recognition 44, no. 8*, 2011: 1649-1657.

[30] Zhou, Dengyong, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. "Learning with local and global consistency." *Advances in neural information processing systems 16, no. 16*, 2004: 321-328.

[31] Zhu, X., Ghahramani, Z., Lafferty, J. "Semisupervised learning using gaussian fields and harmonic functions." *Proceeding of 20th International Conference on Machine Learning.* 2003. 912–919.

[32] Zhu, Xiaojin. *Semi-supervised learning literature survey.* Madison: Technical Report 1530, Department of Computer Sciences, University of Wisconsin, 2005.

[33] Zhu, Xiaojin, Jaz Kandola, Zoubin Ghahramani, and John D. Lafferty. "Nonparametric transforms of graph kernels for semi-supervised learning." *Advances in neural information processing systems*, 2004: 1641-1648.

Appendix A

In the following, we provide a proof for the following theorem:

Main Theorem: Let G is is KNN graph on $\{x_i\}$ sampled from the stratified space, then there exists $\varepsilon > 0$ such that the neighborhood of $B_x(\varepsilon)$ of each point contains about K points.

Proof: If we sample a d_j dimensional manifold with n points, the proportion of points falling in the $B_{x_i}(\varepsilon)$ is $\alpha_j(i)\eta_{d_j}\varepsilon^{d_j}$. Consider the inhomogeneous process $P(r, x_i)$ which counts the points inside the $B_{x_i}(\varepsilon)$ of the manifold $\overline{\Omega}_j$. This process is a binomial process. If $n \longrightarrow \infty$, $k \longrightarrow \infty$ and $\frac{k}{n} \longrightarrow 0$ and, this process is approximated by the Poisson process and if we assume that $\alpha_j(i)$ is constant in the small neighborhood of ε , its expected number is $\alpha_j\eta_{d_j}\varepsilon^j$ [4]. The main notations used in our proof are given in Table 4.If we sample $\{x_i\}$ from intersecting manifold, the number of points falling in the $B_x(\varepsilon)$ is modeled by the mixture of poisons distribution, specified by λ_j, the expectation of each component and π_i, the weight of each component [15]. If we consider $K + 1 = \alpha_{min}\eta_d\varepsilon^d$ and construct KNN graph, the maximum of edge length at any given node would be about ε. The details are given in the ε_{min} theorem and ε_{max} theorem, which are the extension of l_{min} and l_{max} theorems in [4], respectively.

Table 4. Notations

$\overline{\Omega}$	The stratified Space
$\mathbf{x_i}$	ith input data point
D	The intrinsic dimension of stratified space
d_i	The intrinsic dimension of ith manifold
n	The total number of data points
$\overline{\Omega}_\mathbf{j}$	jth manifold
$\alpha_j : \overline{\Omega_j} \rightarrow R_+$	Sampling function of $\overline{\Omega}_j$
$\alpha : \overline{\Omega} \rightarrow R_+$	Sampling function of $\overline{\Omega}$
α_{min}	The minimum value of α_j on $\overline{\Omega}$
α_{max}	The maximum value of α_j on $\overline{\Omega}$
η_d	The volume of unit sphere in R^d
$V_{max}(\varepsilon)$	The volume of largest metric ball in $\overline{\Omega}$ of radius ε
$V_{min}(\varepsilon)$	The volume of smallest metric ball in $\overline{\Omega}$ of radius ε

The $\varepsilon_{\mathbf{min}}$ Theorem. Consider ε_{min} such that $\alpha_{max}V_{max}(2\varepsilon_{min}) < \frac{(K+1)}{2}$, then, with probability of at least $1 - \mu$, no ball with radius ε_{min} contains more than $K + 1$ data points, where $\mu = (\frac{e}{4})^{(\frac{K+1}{2})} \frac{V}{V_{min}(\frac{\varepsilon_{min}}{2})}$

Proof. According to the l_{min} theorem [4],

$$\Pr\left(\text{The number of data points of } B_x(\varepsilon_{min}) \text{ of } \overline{\Omega}_j\ ) \quad > K + 1\right) \leq (\frac{e}{4})^{(\frac{K+1}{2})} \frac{V_j}{V_{min,j}(\frac{\varepsilon_{min}}{2})}$$

Then we have for $\overline{\Omega}$,

$$\Pr(\text{ The number of data points of } B_x(\varepsilon_{min}) \text{ of } \overline{\Omega} \ > K + 1) \leq \sum_j \pi_j \left(\frac{e}{4}\right)^{\left(\frac{K+1}{2}\right)} \frac{V_j}{V_{min,j}\left(\frac{\varepsilon_{min}}{2}\right)}$$

$$\leq \sum_j \pi_j (\frac{e}{4})^{(\frac{K+1}{2})} \frac{V}{V_{min}(\frac{\varepsilon_{min}}{2})} = (\frac{e}{4})^{(\frac{K+1}{2})} \frac{V}{V_{min}(\frac{\varepsilon_{min}}{2})}$$

The $\varepsilon_{\mathbf{max}}$ Theorem. Consider $\varepsilon_{\max}$ such that $\alpha_{\min} V_{\min}\left(\frac{\varepsilon_{max}}{2}\right) > 2\,(K+1)$, then, with probability of at least $1-\mu$, no ball with radius $\varepsilon_{\max}$ contains fewer than $\mathrm{K}+1$ data points, where $\mu = e^{-(\frac{K+1}{4})}\frac{V}{V_{min}(\frac{\varepsilon_{max}}{4})}$

Proof. According to the $l_{\max}$ theorem [4],

$$\Pr\big(\text{The number of data points of } \mathrm{B_x}(\varepsilon_{\max}) \text{ of } \bar{\Omega}_{\mathrm{j}}\ ) \quad < K+1\big) \le e^{-(\frac{K+1}{4})}\frac{V_j}{V_{min,j}(\frac{\varepsilon_{max}}{4})}$$

Then we have for $\bar{\Omega}$,

$$\Pr(\ \textit{The number of data points of } B_{\mathrm{x}}(\varepsilon_{\max})\ \textit{of}\ \bar{\Omega}\ < K+1) \le e^{-(\frac{K+1}{4})}\frac{V_j}{V_{min,j}(\frac{\varepsilon_{max}}{4})} \le e^{-(\frac{K+1}{4})}\frac{V}{V_{min}(\frac{\varepsilon_{max}}{4})}$$

According to above theorems, kNN graph specifies the ϵ neighborhood of each data point sampled from stratified space, under suitable conditions.

From first parts of above theorems, we have

$$2\eta_d(2\varepsilon_{\min})^d \le \frac{(K+1)}{\alpha_{\max}}$$

$$\frac{\alpha_{\min}\eta_d\left(\frac{\varepsilon_{max}}{2}\right)^d}{2} \ge K+1$$

If we let equality in both above conditions, we have

$$\varepsilon_{\min} = \frac{\varepsilon_{max}}{4*(2A)^{\frac{1}{d}}}$$

, where $A = \frac{\alpha_{\max}}{\alpha_{\min}}$.

If K and α tends to infinity, the probability of $\mathrm{B_x}(l_{\min})$ containing too many points and $\mathrm{B_x}(l_{\max})$ containing too few points, becomes exponentially small.

	Zohre Karimi received her B.Sc. degree in Computer engineering from Faculty of Electrical and Computer Engineering at Shahid Beheshti University in 2006, and M.Sc. degree in Computer Engineering from Sharif University of Technology, in 2010. She is currently working toward the PhD degree in the Engineering and Information Technology department at the at Amirkabir University of Technology. Her research interests include Machine learning, Robotics and Data mining.
	Saeed Shiry Ghidary was born in Zanjan, Iran. He received his B.Sc degree in Electronic engineering and Msc in Computer architecture from Amirkabir University of Technology in 1990 and 1994 respectively. He studied robotics and artificial intelligent systems at Kobe University and received his PhD in 2002. He has been an assistant Prof. at Amirkabir University of Technology since 2004. His research interests include Robotics, Machine learning, Machine vision, Cognitive science and Brain modeling.