6th International Conference on Computer and Knowledge Engineering (ICCKE 2016), October 20-21 2016, Ferdowsi University of Mashhad

Weighted Semi-Supervised Manifold Clustering via sparse representation

Amir Abedi
Department of Computer Engineering
Ferdowsi university of Mashhad
Mashhad, Iran
amir.abedi@stu.um.ac.ir

Reza Monsefi
Department of Computer Engineering
Ferdowsi university of Mashhad
Mashhad, Iran
monsefi@um.ac.ir

Davood Zabih zadeh
Department of Computer Engineering
Asrar Institute of Higher Education
Mashhad, Iran
d-zabihzadeh@asrar.ac.ir

***Abstract*— over the last few years, manifold clustering has attracted considerable interest in high-dimensional data clustering. However achieving accurate clustering results that match user desires and data structure is still an open problem. One way to do so is incorporating additional information that indicate relation between data objects. In this paper we propose a method for constrained clustering that take advantage of pairwise constraints. It first solves an optimization program to construct an affinity matrix according to pairwise constraints and manifold structure of data, then applies spectral clustering to find data clusters. Experiments demonstrated that our algorithm outperforms other related algorithms in face image datasets and has comparable results on hand-written digit datasets.**

Keywords—manifold clustering; semi-supervised clustering; high-dimensional clustering; sparse representation

I. INTRODUCTION

In many areas high-dimensional data is arising and traditional clustering approaches have substantial problems in term of efficiency and effectiveness [1] when working with these data. Quality of density estimates around each data point which are utilized to characterizing cluster structure are reduced because of sparseness of high-dimensional data [2]. On the other hand, several distance measures fail to show data distances correctly when dimensionality increases [3, 4]. Therefore, identifying close and distant data and recognizing suitable cluster boundaries becomes more difficult. So it is needed to use another approach for high-dimensional clustering, due to difficulties with using density-based and distance-based methods. The idea is to assume data are on lower dimensional manifolds and find a good projection of data [5]. Recently some studies have shown that we can improve clustering performance by using underlying manifold structure of the data [6-8]. Manifold learning [9, 10] aims to find low-dimensional non-linear manifolds from high dimensional data. Manifold methods is among methods that are used frequently for high-dimensional data clustering.

There is another method in addition to manifold clustering to cluster high-dimensional data. With the assumption that each cluster is combined of multiple components, Expectation Maximization (EM) can be adjusted to learn the clusters as a mixture model. In this approach, the probabilities in Gaussian Mixture Model (GMM) are estimated.

Spectral clustering has shown good performance as a tool for clustering high-dimensional data. Spectral clustering [11] which originates from spectral graph theory, is stable for high-dimensional data clustering [12] and outperformed traditional clustering algorithms because of its polynomial-time and deterministic solution [11]. Nevertheless, spectral clustering performance is heavily related to the affinity matrix it works on. Thus, it is necessary to construct an affinity matrix that reflects similarity information among each pair of data. Traditional weighting methods like ϵ-ball neighborhood, k-nearest neighbors, inverse Euclidean distance [13, 14] and Gaussian RBF[12], depend on Euclidean distance in the ambient data space, so it does not work properly on high-dimensional data. To overcome this difficulties, it is proposed to use sparse representation.

Sparse representation, originating from compressed sensing [15], is shown to be an effective tool for representing and compressing high-dimensional data. It represent each data object based on a sparse linear weight vector of other data objects. Sparse representation convert ambient space to a sparse space. [16] used individual sparse coefficients to construct an affinity matrix for spectral clustering.

In many real world problems that is tied with high-dimensional data, there is some side information in form of labels or constraints, which can help the process of clustering. So it is straightforward to apply Semi-Supervised Learning (SSL) that exploits both unlabeled and labeled data objects, on this problems [17-19]

Our concentration is on clustering with side information of type pairwise constraints (i.e. constrained clustering) that have two different typical form, named Must-links (ML) and cannot-links (CL). ML constraints show the pairs of data objects that must be specified into the same clusters and CL constraints show the pairs of data objects that cannot be specified into the same clusters. It has been shown that side information can improve clustering performance substantially [20, 21].

978-1-5090-3586-1/16/$31.00

One approach to integrate constraints to spectral clustering is to modify affinity matrix so that the must-links and cannot-links be enforced and then applying typical spectral clustering on this affinity matrix [22].

All the method mentioned above have one or several limitation listed below:

(1) Most of semi-supervised methods cannot handle high-dimensional data and capture their geometrical structure to have good clustering performance.

(2) GMM-based and manifold learning methods mentioned, don't consider how to use prior knowledge given by experts in the form of pairwise constraints

(3) Subspace clustering methods work only on high-dimensional data with linear structure.

(4) Some methods of constrained clustering are sensitive to ordering of constraints feed.

In order to address these limitations, we investigate the problem of semi-supervised multi-manifold clustering which is an appearing pattern recognition and machine learning topic and has variety of applications including face detection, document and image ranking [23], image and video annotation [24] web-scale image search [25, 26], protein classification [27], etc. We propose Weighted Semi-Supervised Manifold Clustering (WSSMC) which main goal is to achieve an affinity matrix that reflect the intrinsic manifold structure underlying data objects. To do so, we inherited one recent work [28] on sparse manifold clustering which is proven to be very efficient in finding clusters in high-dimension space and the ability to use side information given by user has been added to it. They suggested to rebuild each data object x_i using another objects in X in a least square manner. Their goal is to fit a subspace to point x_i that has shortest distance to point x_i and is spanned by as few as possible points in X. It causes the subspace to have the dimension as lowest as possible. They used a proximity inducing matrix that is based on Euclidean distance and solved an optimization program to achieve a matrix for applying spectral clustering on it. We have changed the method to create proximity matrix so that pairwise constraint information can be utilized.

The remaining of this paper is organized as follows: Section II briefly reviews some methods that are closely related to our method. Section III introduces the overall framework of WSSMC and some related discussions. Extensive experiments are conducted in Section IV. Finally, we conclude the paper in section V.

II. Related Works

Constrained clustering algorithms can be divided into two main categories: (1) search-based and (2) similarity-based. 1) **Search-based** methods adjust the solution space to be searched according to the constrained via modifying the objective function of clustering algorithms. One common modification is adding penalty terms to objective function for unsatisfied constraints. Another search-based approach is to use prior knowledge to initialize clusters. Seminal work on search-based constrained clustering is COP-k-means. It follows similar clustering process as k-means while respecting constraints not to be violated in the clustering process. This strict approach causes great performance degradation when constraints are noisy. In the case of adding penalization terms to objective function, the papers [29-33] work based on k-means and nonnegative matrix factorization clustering. COnstrained 1-Spectral Clustering (CO1SC) [34] is an extension of spectral clustering to the semi-supervised setting and iteratively attempts to solve exactly an NP-hard discrete optimization problem that captures 2-way constrained clustering. K-way partitions are computed via recursive calls to the 2-way partitioner. It aims to partition a similarity graph such that edges within clusters have high weights and edges between clusters have low weights. Naturally, iterative algorithms are expected to be somewhat slower. CO1SC adapt the optimization objective of spectral clustering to incorporate constraints and propose an alternative optimization procedure. CO1SC handles constraints in the Eigen space construction stage and tries to fulfill all of them by adding terms in spectral clustering to penalize the violation of constraints. Recently, [35] proposed a robust semi-supervised subspace clustering method based on Non-Negative Low Rank Representation (NNLRR). It combines low-rank representation framework and Gaussian Fields and Harmonic Function (GFHF) method in a unified optimization problem. Affinity matrix construction and subspace clustering are done simultaneously in NNLRR. As one of the latest work in constrained clustering domain, Related DP-means (RDP-means) [36] extend k-means for incorporating side information while it doesn't need number of clusters. It works based on DP-means algorithm [37] and uses relational side information.

Gaussian Mixture Model (GMM) is another tool that has been applied to cluster data. It estimates the parameters of multiple Gaussian components using Expectation-Maximization (EM). Locally Consistent GMM (LCGMM) [6] has been proposed to improve clustering performance via exploiting the local manifold structure of data. Gan et al [38] introduced a Semi-supervised LCGMM (Semi-LCGMM) to incorporate prior knowledge into maximum likelihood function of LCGMM. As another GMM-based method, Xing et al [39] proposed a multi-manifold regularized, semi supervised Gaussian mixture model (M2SGMM) to classify multiple manifolds. They use a similarity graph that preserve local and geometrical consistency. Geometrical similarity is measured by applying local tangent space.

2) **Similarity-based** methods try to learn a distance metric or modify similarity matrix of data objects in accordance with the constraints so that data objects with similar cluster labels become closer and data objects with different cluster labels become farther from each other. Relevant Components Analysis (RCA) [40] is an efficient algorithm to learn a Mahalanobis distance metric as a linear transformation of data features. Although RCA is unable to take advantage of cannot link constraints. So [41] proposed Discriminative Component Analysis (DCA) algorithm to address this drawback. It is shown that Mahalanobis distance metric cannot handle data with multi-modal distribution because it learns a fixed metric for entire input space [42].

Modifying similarity matrix is usually done in spectral based algorithms. Constrained Clustering via Spectral Regularization (CCSR) [43] is proposed to incorporate the constraints in

clustering multi-class data using SDP. Flexible constrained spectral clustering (CSP) [44] solves an eigenvalue problem to reduce the space of feasible solutions in order to satisfy certain amount of constraints. However, it require to compute a full-eigenvalue decomposition.

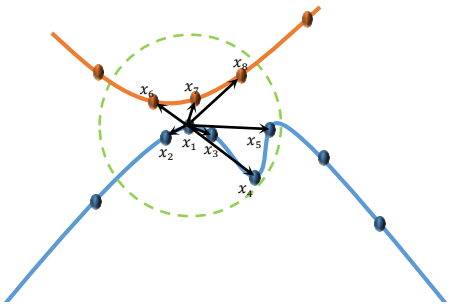


Fig. 1: It is possible for solution C_i to not be unique based on distance of neighbors of x_i

III. The Proposed Metod

As mentioned above, the most important step in spectral clustering is to build the similarity matrix especially in case of having many intersecting or non-intersecting manifolds. To construct a good affinity matrix for clustering, main challenge is to find a formulation to connect data points of the same manifolds to each other. Based on the method proposed in [28], Given a set of N data points $\{x_i \in R^D\}_{i=1}^N$lying in low-dimensional manifolds $\{\mathcal{M}_l\}_{l=1}^c$with intrinsic dimensions $\{d_l\}_{l=1}^c$, we construct the affinity matrix. For each data point x_i, first we use a formulation to find the data points which are in the same manifold with x_i, then choose a few nearest point between them to represent x_i. Let $\mathcal{B}_i \in R^D$be the smallest ball for each $x_i \in \mathcal{M}_l$, that contains $d_l + 1$ nearest neighbors of x_i from $\mathcal{M}_l$ and $\mathcal{N}_i \in R^D$ be the set of all data points in $\mathcal{B}_i$ excluding x_i that contains data points from different manifolds. Suppose for all i there exist an $\epsilon \geq 0$ such that the nonzero entries of the sparsest solution of

$$\left\|\Sigma_{j\in\mathcal{N}_i} c_{ij}(x_j - x_i)\right\|_2 \leq \epsilon \quad and \quad \Sigma_{j\in\mathcal{N}_i} c_{ij} = 1 \qquad (1)$$

Corresponds to the $d_l + 1$ neighbors of x_i from $\mathcal{M}_l$. In other word, the solution C_i is a sparse vector and non-zero elements in it correspond to data points which are neighbors of x_i and there are in the same manifold as x_i. These neighbors span an affine subspace that passes near x_i up to ϵ error and has the lowest dimension d_l.

It is possible for solution C_i to not be unique. For example, in *Fig. 1*, a solution of (1) for two non-zero element can correspond to an affine combination of x_2 and x_3 or an affine combination of x_2 and x_5. So to select between all possible $C_i s$, the solution C_i that contains nearest neighbors to x_i as nonzero elements, we use the below optimization program that tries to select a few data points from neighbors of x_i subject to constraint in (1) at the same time that it tries to satisfy the pairwise constraints that the user has provided.

$$\min \|\boldsymbol{Q_i C_i}\|_1 \quad subject\ to \quad \|\boldsymbol{X_i C_i}\|_2 \leq \epsilon, \quad \mathbf{1}^T \boldsymbol{C_i} = 1 \qquad (2)$$

This objective function penalize each data point according to its proximity to x_i and existence of must-link or cannot-link between the data point and x_i. Data points with smaller distance to x_i and data points for which there are a must-link with x_i, have lesser penalty compared with data points with larger distance to x_i and data points for which there are a cannot-link with x_i. We name data points those have small distance to x_i and data points that have a must-link with x_i as target neighbors (TNs). l_1-norm increases Sparsity of solution C_i and proximity matrix Q_i which is a positive-definite diagonal matrix tries to select TNs. So elements of Q_i should have small values in the case of TNs to allow assigning non-zero coefficients to them. Conversely, for non-target neighbors (NTNs), it should have large values in order to assigning zero coefficient to them. The way that we proposed to construct proximity matrix Q_i based on pairwise constraints is expressed in next section. The optimization program (2) can be rewritten using lagrangian method:

$$\min \lambda\|\boldsymbol{Q_i C_i}\|_1 + \frac{1}{2}\|\boldsymbol{X_i C_i}\|_2^2 \quad subject\ to \quad \mathbf{1}^T \boldsymbol{C_i} = 1 \qquad (3)$$

Where the parameter $\lambda > 0$ sets the trade-off between the sparse solution and the reconstruction error. X_i denotes the matrix of normalized vectors $\{x_j - x_i\}_{j\neq i}$ as:

$$X_i = \left[\frac{x_1 - x_i}{\|x_1 - x_i\|_2}, \dots, \frac{x_N - x_i}{\|x_N - x_i\|_2}\right] \in R^{D*(N-1)} \qquad (4)$$

A. *Proximity matrix construction*

We proposed a formulation to manipulate the proximity inducing matrix for the sake of taking advantage of pairwise constraints. As it is expressed, it is required for TNs of data point x_i to have smaller $\boldsymbol{Q_i}$ comparing with non-TN data points. So we suggest following assignment for q_{ij}:

$$q_{ij} = \begin{cases} \left(\frac{-b_{ij}}{2}\right) + 0.5 \in [0,0.5) & x_i, x_j \in M \\ \left(\frac{-b_{ij}}{2}\right) + 0.5 \in (0.5,1] & x_i, x_j \in C \\ \frac{\|x_j - x_i\|_2}{\Sigma_{t\neq i}\|x_j - x_i\|_2} \in (0,1] & etc \end{cases} \qquad (5)$$

Where M and C are the sets of must-links and cannot-links respectively. b_{ij} is also equal to the importance that user assign to each pairwise constraint. The importance value can be between a value near 0 (as the lowest importance) and 1 or -1 (as the highest importance) and is positive for must-links and negative for cannot-links:

$$b_{ij} \in \begin{cases} [-1..0) & x_i, x_j \in C \\ (0..1] & x_i, x_j \in M \end{cases} \qquad (6)$$

Most of constraint clustering algorithms represent each pairwise constraint as 0 or 1 for cannot-link and must-link constraints, respectively. Although in some real world applications it is not possible for the user to assign a crisp importance value to each constraint. So by applying this weighted approach, we expect to improve usability of our algorithm.

According to equations (5) and (6), data points with any side information about them, have been assigned a q_{ij} according to their distance from each other. So data points with short distance get small q_{ij} and data points those are farther from each other get a large q_{ij}. On the other hand, data points whose absolute importance value is large, have small q_{ij}. In the contrary, data points with small absolute importance value get a large q_{ij}. Although our algorithm can be utilized with crisp importance

values in which situation each constraint has been assigned 1(-1) as importance value, indicating that all constraints has equal importance. According to equation (5), in crisp approach, the value q_{ij} related to each cannot link is 1 and the value q_{ij} related to each must-link is 0. Following this method, TNs of a data point x_i, get a small q_{ij} and consequently non-zero c_{ij},because of having short distance to x_i or having a must-link with x_i and large q_{ij} causes to non-TN points get zero c_{ij}.

B. Clustering

By solving optimization program (3), we can use solution $\boldsymbol{C_i}$ to construct similarity graph to obtain clustering of the data. Considering each data point as a node of graph, we connect each node x_i to other nodes according to elements of $\boldsymbol{C_i}$. Since the non-zero elements of $\boldsymbol{C_i}$ are expected to correspond to TNs of x_i, the constructed graph ideally has several components in which nodes from same manifolds are connected to each other and are separated from the other nodes. Weights are defined as:

$$w_{ii} = 0\ , w_{ij} = \frac{\frac{c_{ij}}{\|x_j - x_i\|}}{\Sigma_{t \neq i}\left(\frac{c_{it}}{\|x_t - x_i\|}\right)}, j \neq i$$

The similarity matrix of the constructed graph is expected to have the ideal form of bellow:

$$\boldsymbol{W} \triangleq [|w_1|\ \ \ldots\ \ |w_N|\,] = \begin{bmatrix} \boldsymbol{W}[1] & 0 & \cdots & 0 \\ 0 & \boldsymbol{W}[2] & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{W}[n] \end{bmatrix} \boldsymbol{\Gamma}$$

Where $\boldsymbol{W}[1]$ is the similarity matrix of the data lying on manifold $\mathcal{M}_l$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{N*N}$ is an unknown permutation matrix. Clustering the data can be done by applying spectral clustering to $\boldsymbol{W}$.

IV. Experiments

Datasets: we have conducted a set of experiments to investigate the performance of our algorithm on 2 Synthetic dataset and 5 benchmark dataset that are listed in *TABLE 1*. All these datasets have been used in state of art articles and can be predicated as high-dimensional data. All of experiments are performed on a system configured with 2.13GHz CPU and 2GB of RAM memory under MATLAB 2012b. All hand-written digits datasets which their number of samples are greater than 1000, are reduced to datasets with 1000 samples via random selections.

To show impact of side-information, experiments are performed with four different side-information rate. Having true labels of each data set, we have chosen 0.01, 0.02, 0.03 and 0.04 of pairwise relations as constraints in each experiment, respectively.

Compared algorithms: As spectral clustering methods are based on graph theory and it is shown that they work properly on high-dimensional data, we concentrate on this methods and most of algorithms which are compared with our algorithm are in this category. On the other hand, prior knowledge in constrained clustering has two form of labeled data or pairwise constraints. Although achieving pairwise constraints is easier than expensive process of obtaining labels of samples, but some constrained clustering algorithms only work on data sets with labeled samples.

Two types of algorithms are compared with the proposed method. As it is possible to obtain pairwise constraints from labels, we can compare our algorithm with algorithms that work based on labels. We have compared our algorithm with NNLRR, M2SGMM and Semi-LCGMM as algorithms that work based on labels. On the other hand RDP-means, CCSR, CSP and CO1SC have been compared with WSSMC as algorithms that work based on pairwise constraints. CO1SC has been implemented in two forms that have different definition of laplacian, named ratio cut (CO1SC-RC) and normalized cut (CO1SC-NC.)

Evaluation metrics: We have used BCubed F-measure or pairwise F-measure to evaluate clustering results. Based on comparison performed between different extrinsic clustering evaluation metrics in [45], pairwise F-measure has better performance. F-measure definition is based on precision and recall metrics:

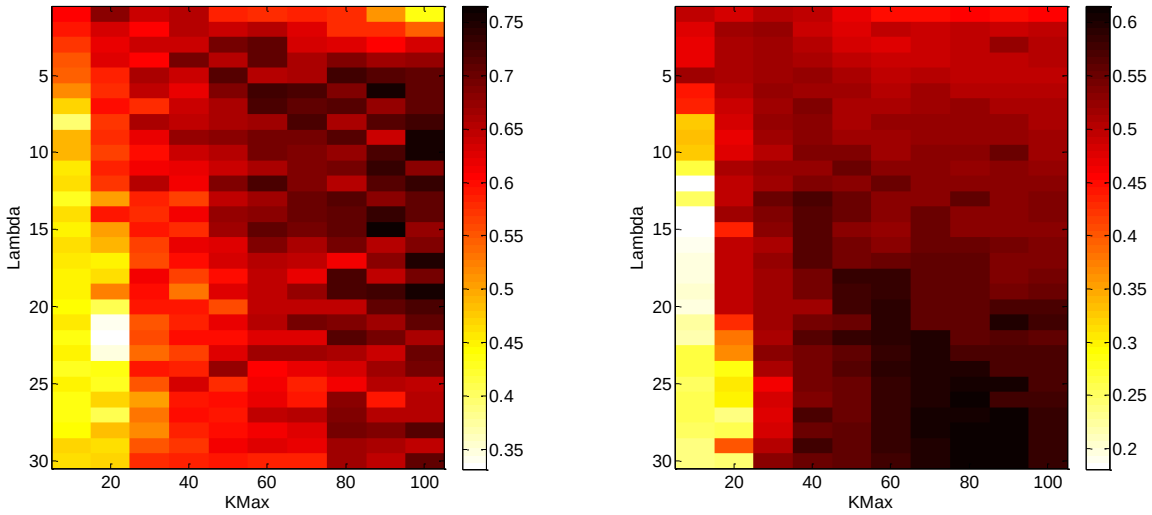
$$Precision = \frac{\#\ Pairs\ Correctly\ Predicted\ In\ Same\ Cluster}{\#\ Total\ Pairs\ Predicted\ In\ Same\ Cluster}$$

$$Recall = \frac{\#\ Pairs\ Correctly\ Predicted\ In\ Same\ Clsuter}{\#\ Total\ Pairs\ In\ Same\ Cluster}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

TABLE 1: Datasets

Name	Type	# of clusters	# samples per cluster	# of attributes	# of total samples
2trifoils	Synthetic	2	100	100	200
2semi_trifoils-plane_with_hole	Synthetic	3	Different	100	468
USPS	Hand-written digits	10	1100	256	11000
MNIST	Hand-written digits	10	Different	784	70000
ORL	Face	40	10	64*64	400
UMIST	Face	20	20 to 30	92*112	575
YALE	Face	15	11	64*64	165



(a) (b)

Fig. 2: The impact of changing λ and K_{Max} on F-measure value for WSSMC algorithm applied on (a) ORL and (b) USPS datasets respectively with side-information rate of 0.01

A. *Comparing with pairwise constraint-based algorithms*

In this subsection we study the performance of WSSMC … The parameters used in WSSMC are empirically set for each dataset. For example, in the following figure (Fig. *2* (a)) we can see different F-measure values obtained by applying WSSMC algorithm on ORL dataset and USPS dataset with 0.01 side information for $K_{Max} \in \{10,20,30,40,50,60,70,80,90,100\}$ and $\lambda \in [1,30]$. Points with a color near dark red has greater F-measure and points with lighter color has less F-measure value. It shows that greater F-measure values occur for K_{max} values more than 60 and the greatest value for F-measure is occurred having $K_{Max} = 90$ and $\lambda = 15$.

The resulting parameter assignments for all datasets for WSSMC algorithm are shown in *TABLE 2*. The parameters for other algorithms are set as the value that their authors configured. Each experiment has been run 5 times and the average F-measure calculated.

Figure 3 shows the F-measure value versus different side-information rates for different algorithms on some datasets listed in TABLE 1: Datasets. As it is clear, our proposed algorithm, WSSMC, outperforms in all datasets except USPS. Also it has comparable results on USPS dataset. Generally, adding more side information, improve clustering result in most of datasets

TABLE 2: WSSMC parameter assignment

Datasets	λ	K_{Max}
2trifoils	17	10
2semi_trifoils-plane_with_hole	1	10
USPS	29	90
MNIST	30	100
ORL	15	90
UMIST	27	30
YALE	10	80

and algorithms. Increasing side-information rate causes great growth in F-measure value for CO1SC-RC, CO1SC-NC and CSP. It can be seen that WSSMC is more robust.

Considering manifold structure information is one reason of this robustness. WSSMC has the best performance on two synthetic datasets with having F-measure equal to 1. Also, on the face image datasets, it is WSSMC that has the best F-measure value.

B. *Comparing with label-based algorithms*

In this subsection we compare clustering result of WSSMC with Semi-LCGMM, M2SGMM and NNLRR algorithms. This algorithms have good precision, but very weak recall causes to low F-measure value for this algorithms. As it is shown in *TABLE 3*, WSSMC algorithm has the rank 1,2 and 3 in precision value of clustering results on 2trifoils, MNIST and ORL respectively, but having the best recall value on all datasets, causes to WSSMC has the best F-measure value among all algorithms. The best values in each column become bolded. NNLRR and Semi-LCGMM don't respond on MNIST and ORL respectively. *Figure 4* shows clustering performance of label-based algorithms for different side-information rates. WSSMC outperform other algorithms on 2trifoils and ORL dataset and has comparable result on MNIST. Again we can see robustness of WSSMC on all datasets.

V. Conclusion

In this paper, we have introduced a novel semi-supervised method for clustering high-dimensional data based on sparse representation. It first construct an affinity matrix by solving an optimization program that considers side-information and manifold structure information, simultaneously. Then, it apply spectral clustering on constructed affinity matrix to cluster data. Extensive experiments show that our algorithm outperformed on face image datasets and has comparable results on hand-written digits datasets.

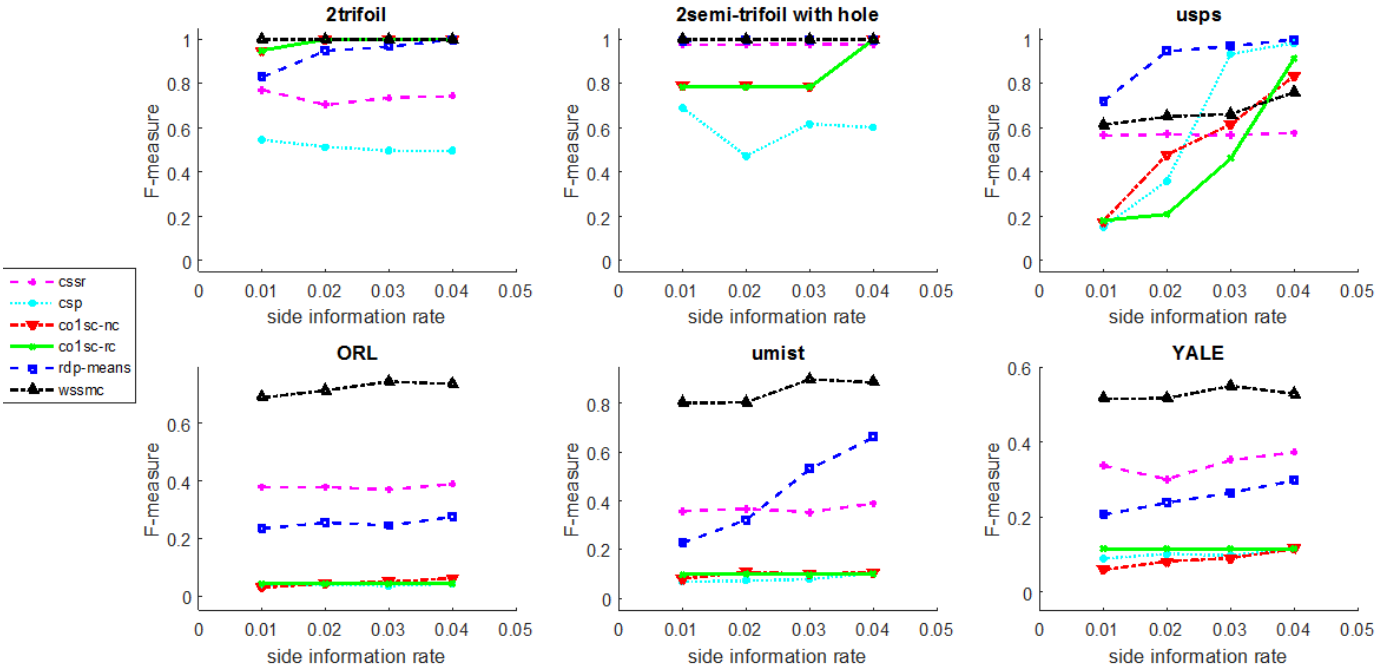


Figure 3: Clustering results of pairwise constraint based algorithms on six datasets

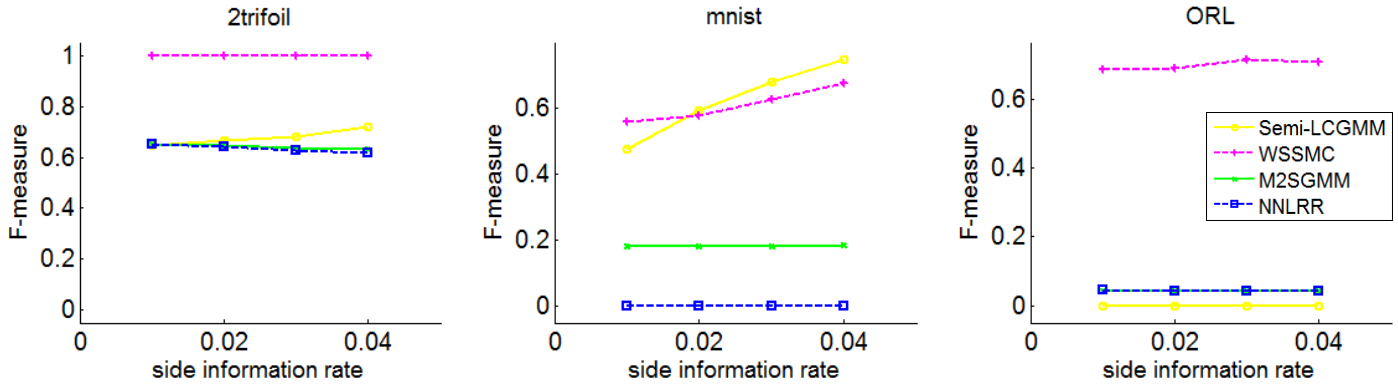


Figure 4: Clustering results of label-based algorithms on three datasets

TABLE 3: Clustering evaluation on different datasets for label-based algorithms

Method	2trifoils			MNIST			ORL		
metric	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Semi-LCGMM	0.705993	0.598019	0.644841	0.55666	0.417249	0.47557	0	0	0
WSSMC	**1**	**1**	**1**	0.59033	**0.527644**	**0.556997**	0.725926	**0.64734**	**0.684154**
M2SGMM	0.95202	0.498018	0.653946	**0.908669**	0.099936	0.180068	0.805	0.022299	0.043396
NNLRR	0.950808	0.497384	0.653114	0	0	0	**0.818519**	0.022674	0.044125

ACKNOWLEDGMENT

This work is supported by machine learning laboratory in engineering college of Ferdowsi university of Mashhad. Authors would like to thank the authors who help us by giving their algorithms code for experimental phase.

REFERENCES

[1] M. Steinbach, L. Ertöz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*, L. T. Wille, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 273-309.

[2] D. W. Scott and J. R. Thompson, "Probability density estimation in higher dimensions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, 1983, pp. 173-179.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International Conference on Database Theory*, 2001, pp. 420-434.

[4] D. Francois, V. Wertz, and M. Verleysen, "The Concentration of Fractional Distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 873-886, 2007.

[5] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, p. 1, 2009.

[6] J. Liu, D. Cai, and X. He, "Gaussian Mixture Model with Local Consistency," in *AAAI*, 2010, pp. 512-517.

[7] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized gaussian mixture model for data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1406-1418, 2011.

[8] J. Shen, J. Bu, B. Ju, T. Jiang, H. Wu, and L. Li, "Refining Gaussian mixture model based on enhanced manifold learning," *Neurocomputing*, vol. 87, pp. 19-25, 2012.

[9] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, pp. 2268-2269, 2000.

[10] P. Zhang, H. Qiao, and B. Zhang, "An improved local tangent space alignment method for manifold learning," *Pattern Recognition Letters*, vol. 32, pp. 181-189, 2011.

[11] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395-416, 2007.

[12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849-856, 2002.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, pp. 1373-1396, 2003.

[14] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, pp. 2319-2323, 2000.

[15] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, pp. 1289-1306, 2006.

[16] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, pp. 1031-1044, 2010.

[17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 39, pp. 103-134, 2000.

[18] R. Cruz-Barbosa and A. Vellido, "Semi-supervised geodesic generative topographic mapping," *Pattern Recognition Letters*, vol. 31, pp. 202-209, 2010.

[19] M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton, "Semi-supervised kernel canonical correlation analysis with application to human fMRI," *Pattern Recognition Letters*, vol. 32, pp. 1572-1583, 2011.

[20] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, and J. Han, "Reinforced similarity integration in image-rich information networks," *IEEE transactions on knowledge and data engineering*, vol. 25, pp. 448-460, 2013.

[21] A. Khoreva, F. Galasso, M. Hein, and B. Schiele, "Learning must-link constraints for video segmentation based on spectral clustering," in *German Conference on Pattern Recognition*, 2014, pp. 701-712.

[22] K. Kamvar, S. Sepandar, K. Klein, D. Dan, M. Manning, and C. Christopher, "Spectral learning," in *International Joint Conference of Artificial Intelligence*, 2003.

[23] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," *Advances in neural information processing systems*, vol. 16, pp. 169-176, 2004.

[24] Y.-G. Jiang, Q. Dai, J. Wang, C.-W. Ngo, X. Xue, and S.-F. Chang, "Fast semantic diffusion for large-scale context-based image and video annotation," *IEEE Transactions on Image Processing*, vol. 21, pp. 3080-3091, 2012.

[25] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1877-1890, 2008.

[26] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang, "Noise resistant graph ranking for improved web image search," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 849-856.

[27] J. Weston, D. Zhou, A. Elisseeff, W. S. Noble, and C. S. Leslie, "Semi-supervised protein classification using cluster kernels," in *Advances in neural information processing systems*, 2003, p. None.

[28] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Advances in neural information processing systems*, 2011, pp. 55-63.

[29] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 11.

[30] Y. Chen, M. Rege, M. Dong, and J. Hua, "Incorporating user provided constraints into document clustering," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 2007, pp. 103-112.

[31] F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," in *SDM*, 2008, pp. 1-12.

[32] T. Li, C. Ding, and M. I. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 2007, pp. 577-582.

[33] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowledge and Information Systems*, vol. 17, pp. 355-379, 2008.

[34] S. S. Rangapuram and M. Hein, "Constrained 1-Spectral Clustering," in *AISTATS*, 2012, p. 90.

[35] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Robust Semi-Supervised Subspace Clustering via Non-Negative Low-Rank Representation," 2015.

[36] D. Khashabi, J. Y. Liu, J. Wieting, and F. Liang, "Clustering With Side Information: From a Probabilistic Model to a Deterministic Algorithm," *arXiv preprint arXiv:1508.06235*, 2015.

[37] K. Jiang, B. Kulis, and M. I. Jordan, "Small-variance asymptotics for exponential family Dirichlet process mixture models," in *Advances in Neural Information Processing Systems*, 2012, pp. 3158-3166.

[38] H. Gan, N. Sang, R. Huang, and X. Chen, "Manifold regularized Gaussian mixture model for semi-supervised clustering," in *2013 2nd IAPR Asian Conference on Pattern Recognition*, 2013, pp. 361-365.

[39] X. Xing, Y. Yu, H. Jiang, and S. Du, "A multi-manifold semi-supervised Gaussian mixture model for pattern classification," *Pattern Recognition Letters*, vol. 34, pp. 2118-2125, 2013.

[40] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *ICML*, 2003, pp. 11-18.

[41] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, pp. 2072-2078.

[42] L. Wu, R. Jin, S. C. Hoi, J. Zhu, and N. Yu, "Learning Bregman distance functions and its application for semi-supervised clustering," in *Advances in neural information processing systems*, 2009, pp. 2089-2097.

[43] Z. Li, J. Liu, and X. Tang, "Constrained clustering via spectral regularization," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 421-428.

[44] X. Wang and I. Davidson, "Flexible constrained spectral clustering," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 563-572.

[45] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information retrieval*, vol. 12, pp. 461-486, 2009.