International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.213-224
http://dx.doi.org/10.14257/ijdta.2017.10.1.20

Data Clustering Analysis on Grassmann Manifold Metric

Yinghong Xie*[1,2], Yuqing He[1], Xiaosheng Yu[3], Xindong You[4], Qiang Guo[5]

[1]School of Electronic Information Engineering,
Tianjin University, Tianjin 300072, China, xieyinghong@163.com
[2]School of Information Engineering, Shenyang University,
Shenyang , China, heyuqing@tju.edu.cn
[3]College of Information Science and Engineering, Northeastern University,
Shenyang, China, yuxiaosheng7@163.com
[4]Beijing Institute Of Graphic Communication, Beijing China
youxindong@bigc.edu.cn
[5]Library, National Police University of China, Shenyang, 110854, China,
royinchina@163.com

Abstract

In the standard spectrum clustering algorithm, the metric based on Euclidean space can not represent the complicate space distribution feature of some data set, which might lead to the clustering result inaccuracy. While the geometric relationship between data can be describe more precise by manifold space. Considering Grassmann manifold is a entropy of Lie group, which not only has the smooth curved surface but also has the feature more fit for measuring the distance between data. All these can make the clustering result more accurate. The improved spectrum clustering analysis algorithm based on the distance metric under Graasmann manifold is proposed by this paper. The similarity between data is analyzed under manifold space. Experimental results show that the proposed algorithm can cluster data set either belonging the same or different subspace more accurately, further more, it can cluster data set with more complicate geometric structure under manifold space efficiently.

Keywords*: clustering analysis; Lie group; Grassmann manifold; spectrum clustering; distance metr*ic

1. Introduction

Clustering analysis is the basis of modern data analysis. Clustering is to group data objects into multiple clusters, which makes the objects in the same cluster have high degree of similarity, while objects in different clusters are of great difference. This kind of algorithm firstly defines a matrix describing the similarity between the date points according to the given sample data set, and calculates the eigenvalues and eigenvectors of the matrix, then clusters different data points by selecting the appropriate feature vectors [1-2]. Classification based clustering method has been widely used in pattern recognition, data mining and other fields, and it is still the source of many research work. *K* means clustering and FCM(Fuzzy C means clustering) clustering are typical representatives of this kind of algorithm. In recent years, the research results mainly include: the density weighted fuzzy clustering algorithm [3], the double exponential fuzzy C mean algorithm [4] based on the hybrid distance learning and so on. The advantages of this kind of algorithm can be attributed to the fast convergence speed and easy to extend [5]. These algorithms can obtain more accurate classification results when the data set to be analyzed is consistent with the assumption of the model structure [6-7].But they usually need to specify the number of clusters. In addition, the selection of the initial cluster center, the

ISSN: 2005-4270 IJDTA

existence of noise data and the number of clusters set will have great influence on the clustering results. Especially when the data structure is complicated, the classification results of these algorithm are not accurate.

Usually spectral clustering analysis method are used to deal with the complicated problem of data clustering [8]. Essentially, spectral clustering algorithm educes the new characteristics of the objects to be clustered by the theory on matrix spectral analysis, and clusters the original data using the new data features [1]. And spectral clustering algorithm based on normal Laplasse matrix is one of the typical algorithms, which is also called NJW algorithm. But the topological structure of manifold is constructed based on Euclidean distance in all the classical spectral clustering analysis algorithms [9-15], which may lead to the chaotic topology.

Manifold study algorithm is dimension reduction algorithm developed in recent years with the aim of finding the more important lower dimension structure in the higher dimension data. and the algorithm is widely used in the recognition of face, traffic logo *etc.* [16-21]. Considering the Grassmann manifold is a manifold entropy in Lie group manifold, which not only has a smooth surface expression of space, but also has the characteristics being more suitable to measure the distance between data points. Based on the study of NJW algorithm, this paper proposed an analysis method for data clustering based on Grassmann manifold, which compares the similarity of data points in Grassmann manifold space that can cluster the data points either in dependent or independent subspace effectively. At the same time it can effectively cluster the manifold space data set.

2 Grassmann Manifold and its Metric

The points on Grassmann manifold $Gr(k,n)$ are the set of equivalence classes of $n\times k$ dimensional orthogonal matrix, that is:

$$Gr(k,n)=\lceil Y\rceil_{O_k}=\{YV:V\in O_k\} \tag{1}$$

Where, Y denotes $n\times k$ dimensional orthogonal matrix, $\lceil Y\rceil$ represents the relationship of equivalence classes, and V is $k\times k$ dimensional orthogonal matrix.

Grassmann manifold $Gr(k,n)$ can also represent the set of all the k-dimensional subspaces in n-dimensional vector space R^n. Grassmann manifold has the representation form of quotient space $Gr(n,k)=O(n)/(O(n-k)\times O(k))$, which is the remaining portion in the orthogonal Lie group removing the swirling of coplanar and non-coplanar.

$$T_pGr(k,n)=\{\omega|\omega=p_\perp g,\quad g\in R(n-k,k)\} \tag{2}$$

The common method for defining the metric structure on manifold M is to assign inner product $\langle\cdot,\cdot\rangle$ for the tangent space T_pM of each point $p\in M$, that is Riemann metric. For any point $p\in Gr(k,n)$,Where $Gr(k,n)$ is Grassmann manifold, the tangent space is :

$$T_pGr(k,n)=\{\omega|\omega=p_\perp g,\quad g\in R(n-k,k)\} \tag{3}$$

Where $p_\perp$ is the orthogonal complement for point p. The metric on the $Gr(k,n)$ is defined as:

$$\|\omega\|=Tr(\omega^T\omega) \tag{4}$$

Let $\gamma : t \Rightarrow \gamma(t)$ be the geodesics for the initial point $\gamma(0)$, the initial vector $\frac{d\gamma}{dt}(0) = \omega$, where the exponential map $\mathrm{Exp}_p(\omega) = \gamma(1)$ defines the end of the geodesics:

$$\mathrm{Exp}_p(\omega) = pV\cos(\theta) + U\sin(\theta) \tag{5}$$

Where $U\theta V^T = SVD(\omega)$. The corresponding inverse mapping is $\mathrm{Log}_p(q) = U\theta V^T$, where $\theta = \arctan(S)$ and $USV^T = p_\perp p_\perp^T q\left(p^T q\right)^{-1}$. Therefore, the geodesic distance between the points (p,q) on Grassmann manifold is defined as:

$$d_G(p,q) = \left(\sum_{i=1}^{k}\theta_i^2\right)^{1/2} = \|\theta\|_2 \tag{6}$$

An example for $Gr(2,3)$ is shown in Figure 1. The line between Y_1 and Y_2 represents the geodestic on Grassmann manifold.

Its sectional curvature is:

$$K_{Gr(n,k)}(X,Y) = \|[X,Y]\|^2 \tag{7}$$

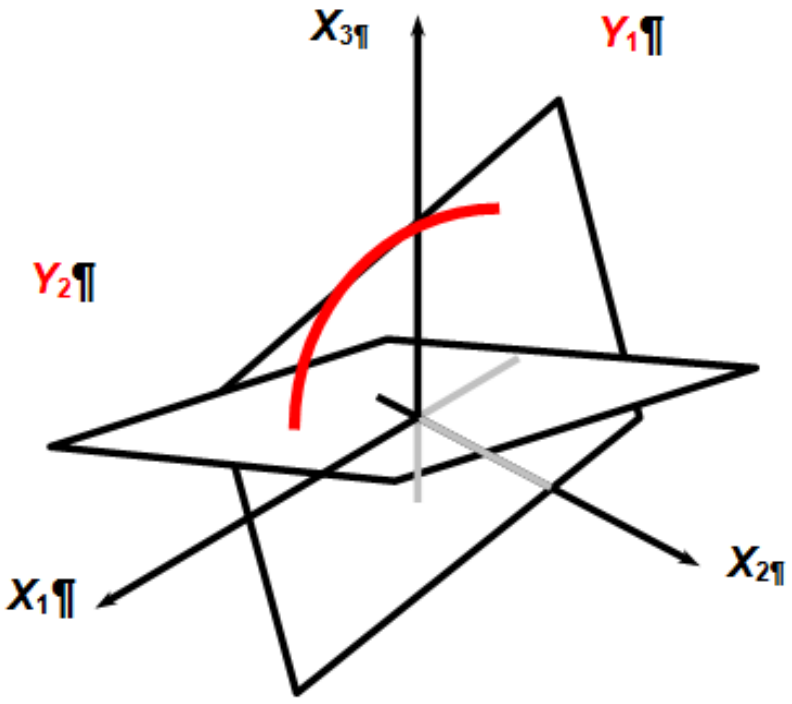


Figure 1. Visualization of Grassmann Manifold $Gr(2,3)$

From formula (2), each point on Grassmann manifold has non negative sectional curvature. Especially, when p equals 1 or n-1, the Grassmann manifold is degenerate into a spherical manifold, whose sectional curvature is constant 1.

3. Spectral Clustering Analysis Algorithm Based on Grassmann Manifold

3.1. Improved Spectral Clustering Algorithm

Spectral clustering algorithm is based on spectrum division theory, and it is a kind of high performance computing method. It regards the data clustering as a undirected graph multiway division problem. Ng-Jordan-Weiss (NJW) algorithm is a popular spectral clustering algorithm. The algorithm is to select the eigenvectors corresponding to Laplasse matrix of the largest K eigenvalue, which correspond with the original data representation in Rk space, and then clustered in the space [22].

In this paper, taking into account that the local manifold topological structure used by NJW algorithm is constructed on Euclidean distance, chaotic topology of local manifold may exist. In order to make the spectral clustering algorithm has better clustering accuracy for different data, we propose a spectral clustering algorithm based on distance metric of Grassmann manifold, thereby improving the accuracy of clustering.

The method is carried out according to the following steps:

step 1:

input n data points $\{x_i\}_{i=1}^n$ and number k to be clustered.

step 2:

Compute the distance between data points based on the distance formula on the Grassmann manifold $d_G(p,q)=\left(\sum_{i=1}^{k}\theta_i^2\right)^{1/2}=\|\theta\|_2$, then build the similarity matrix $S\in R^{n\times n}$.

where p and q are the two points on Grasmann manifold, the main angle between p and q is $\theta_1,\ldots,\theta_k$.

step 3:

Construct Laplasse matrix $L=D^{-1/2}SD^{-1/2}$, where D is diagonal matrix $D_{ii}=\sum_{j=1}^{n}S_{ij}$.

step 4:

compute the eigenvectors $v_1,v_2,...,v_k$ corresponding to Laplasse matrix of the largest K eigenvalue, and build matrix $V=[v_1,v_2,...v_k]\in R^{n\times k}$, where v_k is column vector.

step 5:

Normalize the row vector of V, and get the matrix Y, where $Y_{ij}=V_{ij}/(\sum_j V_{ij}^2)^{1/2}$.

step 6:

Regarding each row of Y as a point in space R^k,cluster them using k-means algorithm.

step 7:

If row i belong to class j, the original data X_I is to be classified to cluster j. Then output the division $c_1,c_2,...,c_k$.

3.2. Kernel Parameter Determination

This paper uses the Gauss kernel function as the distance similarity measure, the same with standard spectral clustering algorithm. But the difference is, the computation of the distance between the elements is carried on the more accurate Grassmann manifold, and the result is used to assess the similarity. The kernel function used in this paper are as follows:

$$K_{ij} = \exp(\frac{-D_{ij}}{2\sigma_1\sigma_2}) \tag{8}$$

Nuclear width:

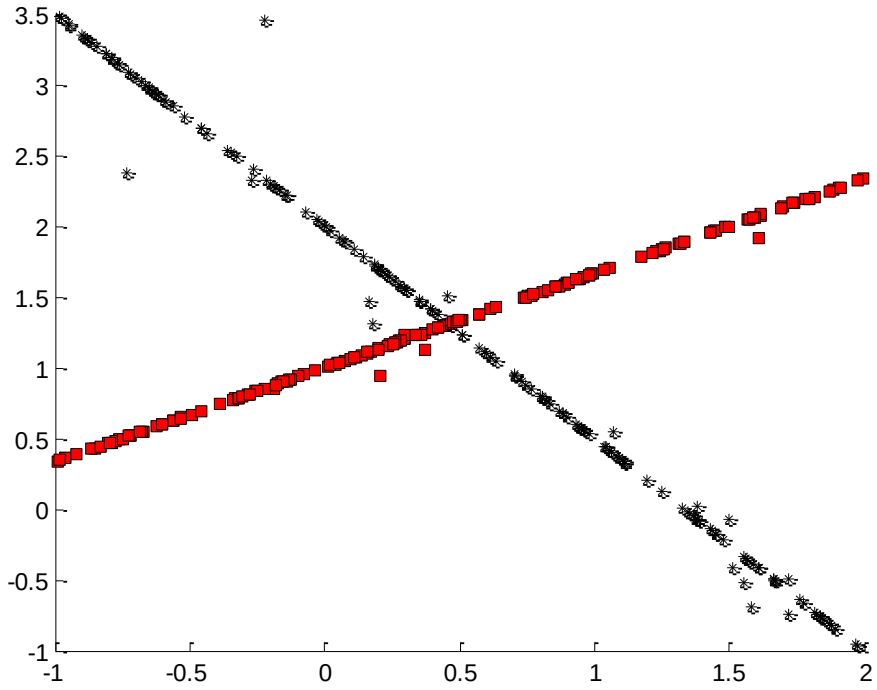
$$\sigma_i = d(x_i, x_{il}) \tag{9}$$

Where $d(x_i, x_{il})$ is computed by formula (5). X_{il} is the lth neighborhood point of X_i. σ_i changes adaptively with the nearest neighbor distribution to ensure that the similarity between the same class in the sample is the largest, while the similarity between the different class in the sample is much lower.

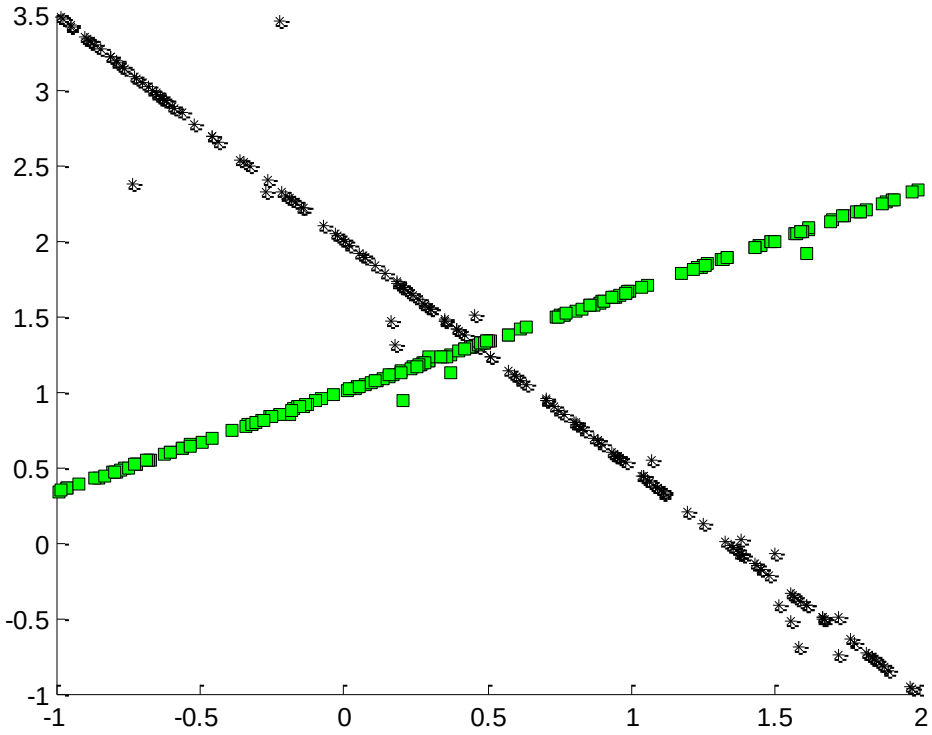
4. Experimental Results and Analysis

In order to verify the effectiveness of the algorithm, the proposed algorithm is compared with the standard spectral clustering algorithm.

Enter 200 data points with 100 dimension $\{x_i\}_{i=1}^{200}$, the number to be clustered is 2. Each data point is a column vector with 100 dimension, and 200 data points consist of a 100*200 matrix.



(a) Clustering Result by Standard Spectrum Clustering Algorithm



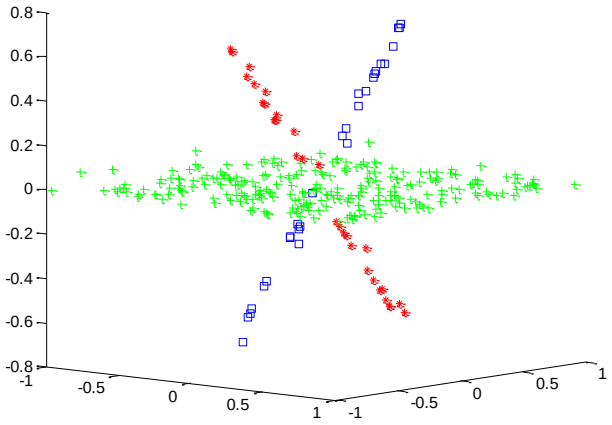
(b) Clustering Result by the Proposed Algorithm

Figure 2. Clustering Result for Data Set in Independent Subspace

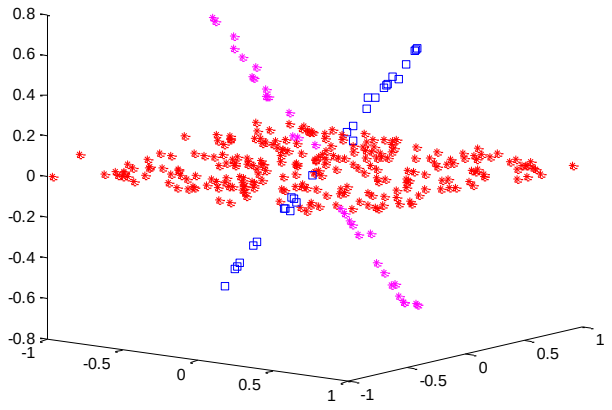
If row i belong to class j, the original data x_l is to be classified to cluster j. Then output the division C_1, C_2. The results of the two algorithms are the same.

In the first group of experiments, we classify the points in the subspace independently. This case they are divided into two categories. The results of the two algorithms are the same, as shown in Figure 2.

In the second group of experiments, the selected sample set belongs to dependant sub space, and the number of cluster is set manually, which is 3 in this experiment. The results of the two algorithms are the same, as shown in Figure3. They can also classify the sample set.



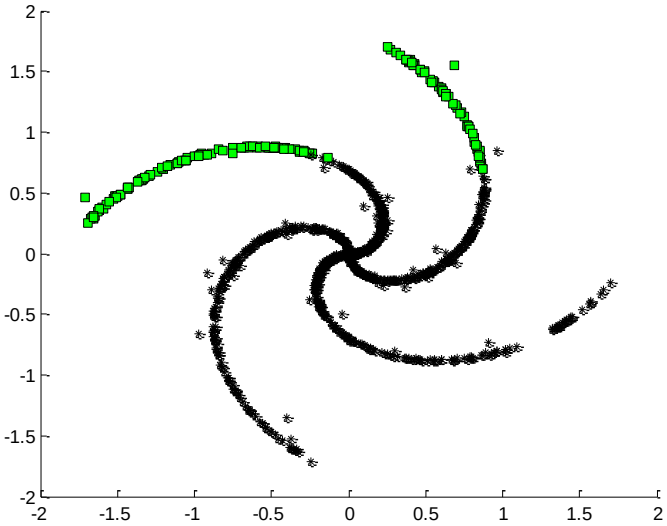
(a) Clustering Result by Standard Spectrum Clustering Algorithm



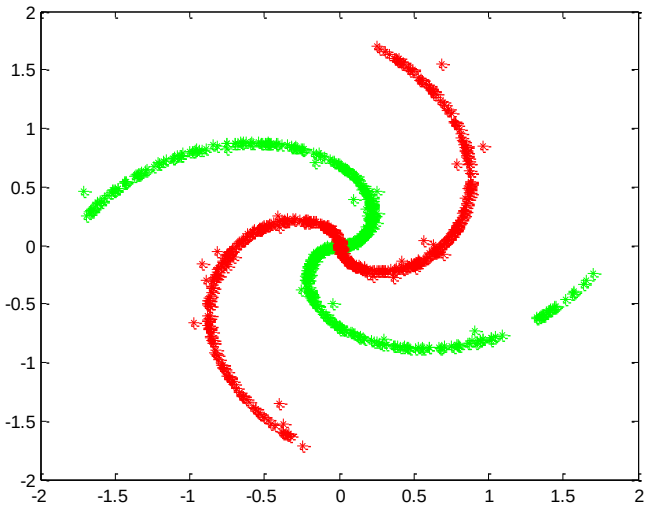
(b) Clustering Result by Standard Spectrum Clustering Algorithm

Figure 3. Clustering Result for Data Set in Dependent Subspace

In the second group of experiments, we validate the effectiveness of the proposed algorithm for manifold spatial data clustering. And the number of cluster is set manually, which is 2 in this experiment. The results of the two algorithms are shown in Figure4. From it , we can conclude that the standard spectral clustering algorithm failed to cluster the data set, while the proposed algorithm can classify the data set accurately.



(a) Clustering Result by Standard Spectrum Clustering Algorithm

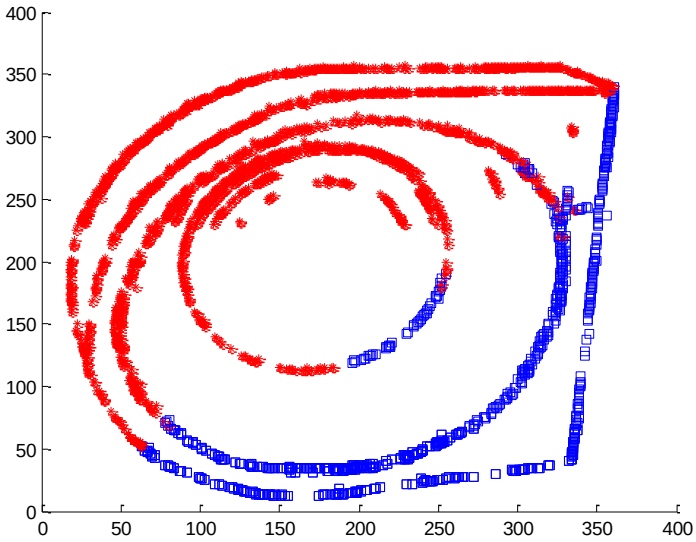


(b) Clustering Result by the Proposed Algorithm

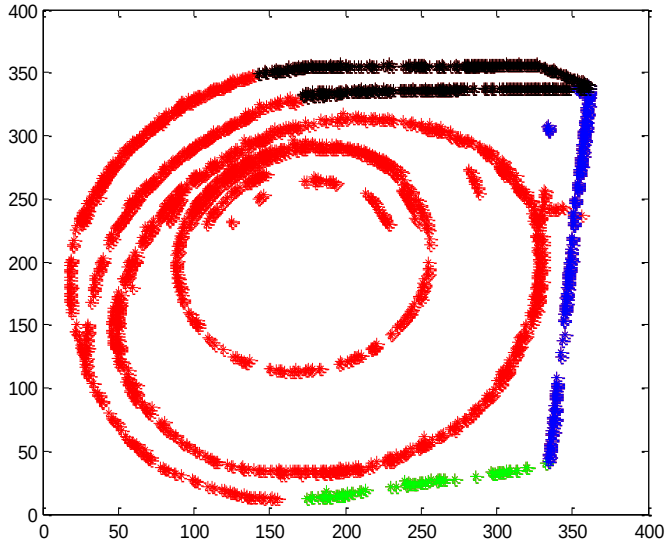
Figure 4. Clustering Result for Data Set in Manifold Space

In the second group of experiments, we validate the effectiveness of the proposed algorithm for complex manifold spatial data clustering. And the number of cluster is set by the algorithm. The results of the two algorithms are shown in Figure 5. And the experimental data set is divided into 2 clusters by the standard spectral clustering algorithm. but the number is 4 by the proposed algorithm. From the results shown in the Figure, we can conclude the proposed algorithm can classify the data set more accurately.

In summary, the proposed algorithm can not only effectively to cluster the data set in different sub space, but also can analyze the data set with complex geometry and effectively cluster them in manifold space.



(a) Clustering Result by Standard Spectrum Clustering Algorithm



(B) Clustering Result by the Proposed Algorithm

Figure 4. Clustering Result for Data Set in Manifold Space

5. Conclusion

In the standard spectral clustering analysis algorithm, the metric based on Euclidean space can not fully reflect the complex spatial distribution characteristics of data clustering ,which leads to the clustering results being not accurate enough. While manifold space can describe the geometrical structure relationship between data more accurate. Considering the Grassmann manifold is a manifold entropy in Lie group manifold, which not only has a smooth surface expression of space, but also has the characteristics being more suitable to measure the distance between data points, this paper proposed an analysis method for data clustering based on Grassmann manifold, which compares the similarity of data points in Grassmann manifold space that can cluster the

data points either in dependent or independent subspace effectively. At the same time it can effectively cluster the manifold space data set.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant [No. 61503274, No.61603415, No.61472274], Science and technology research projects of Education Department, Liaoning Province [No. L2014481], and Doctor Startup Fund Program Funded by Liaoning Province Education Administration [No.201501090]. Fundamental Research Funds for the Central Universities [No. 140403005, No. 140404014].

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] ZHOU Lin,PING Xi-Jian,XU Sen,ZHANG Tao. Cluster Ensemble Based on Spectral Clustering[J], ACTA AUTOMATICA SINICA, 2012,38(8):1335-1342.
[2] Deng Zhao-hong, Choi Kup-sze, Chung Fu-lai, Enhanced soft subspace clustering integrating within cluster and between-cluster information[J]. Pattern Recognition, 2010. 43(3): 767-781.
[3] Hathaway R J, Hu Y. Density-weighted fuzzy k-means clustering[J]. IEEE Trans on Fuzzy Systems, 2009, 17(1):243-252.
[4] Wang J, Wang S T. A dsouble-indexed FCM algorithm based on Hybrid distance metric learning[J]. J of Software, 2010, 21(8): 1878-1888.
[5] WANG Jun, WANG Shi-tong, DENG Zhao-hong. Survey on challenges in clustering analysis research[J], Control and Decision, 2012, 27(3):321-327.
[6] JAIN A K. Data clustering: 50 years beyond k-means[J], Pattern Recognition Letters, 2010, 31(8): 651-666.
[7] XIE Y. Segmantation method combined with mean shift and K-mean clustering algorithm for clothing image[J], Electronic Measurement Technology,2013,36(8):53-60.
[8] Qian P J, Wang S T, Deng Z H. Fast spectral clustering for large data sets using minimal enclosing ball[J]. Acta Electronica Sinica, 2010, 38(9): 2035-2041.
[9] CHI Y, SONG X D. On evolutionary spectral clustering[J]. ACM Transactions on Knowledge Discovery from Data, 2009,3(4):17-47.
[10] Wang L, Bo L W, Jiao L C. Density-sensitive semi-supervised spectral clustering[J]. Journal of Software, 2007, 18(10): 2412- 2422.
[11] ZHOU W G, CHEN L T, DONG S. Network traffic classification algorithm based on spectral clustering[J]. Journal of Electronic Measurement and Instrument.2013,7(12):1114-1119.
[12] CHANG H, YEUNG D Y. Robust path-based spectral clustering[J].Pattern Recognition, 2008, 41(1): 191-203.
[13] VIDAL R. Subspace clustering, IEEE Signal Processing Magazine, 28(2):52–68, 2011.
[14] ELHAMIFAR E,VIDLA R. Sparse subspace clustering, Algorithm, theory, and applications.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013,35(11):2765–2781.
[15] WANG Y, JIANG Y,WU Y, ZHOU Z. Spectral clustering on multiple manifolds, IEEE Transactions on Neural Networks, 22(7):1149–1161, 2011.
[16] LIU Y P, LI G W, SH Z L. Projective registration algorithm based on Riemannian manifold[J]. ACTA Automatica Sinica. 2009,35(11): 1378-1386.
[17] Ehsan Elhamifar, Rene Vidal, Sparse manifold clustering and embedding[C], Proc of the 25th Annual Conf on Neural Information Processing Systems, Sierra Nevada,2011: 55-63.
[18] XIE Y H, WU C D. Object Tracking with dual modeling based on projection group and covariance manifold[J]. Chinese Journal of Scientific Instrument, 2014,35(2):374-379
[19] GUO Q, WU C D, FENG Y, LU X H. Conjugate Gradient Algorithm for Efficient Covariance Tracking with Jensen-Bregman LogDet Metric[J]. IET Computer Vision.2015,9(6):814-820.
[20] SI X D, LIU G. Method of water and land segemantation in optical remote sensing images[J]. Forgein Electronic Measurement Technology 2014,33(11):29-32.
[21] LU J, TAN Y .P, WANG G. Discriminative multi-manifold analysis for face recognition from a single training sample perperson[J], IEEE International Conference on Computer Vision, 2011:1943-1950.
[22] Yang Yifang,Wang Yuping. Spectral clustering algorithm based on kernel fuzzy similarity measure[J]. Chinese Journal of Scientific Instrument, 2015,36(7):1562-1569.

Authors

Xie Yinghong, received her B.Sc. degree in 1999 from Shenyang Jianzhu university, received her M.Sc. degree in 2005 from Northeastern University, received her Ph.D. degree in 2014 from Northeastern University. Now she is an associate Professor in Shenyang University, and a post doctor in Tianjin University. Her main research interests include video image processing, pattern recognition.

Yuqing He, received his Ph.D. degree in Electronic Engineering from the Tianjin University, China, in 2008. He is currently a Lecturer with School of Electronic Information Engineering, Tianjin University, China. His research interests include image processing, image restoration, computer vision, and pattern recognition.

Xiaosheng Yu, received his Bachelor degree majored in artificial intelligence and robotics from University of Bedfordshire in 2006. He received his Master degree in mechanical systems engineering from University of Liverpool in 2007. He received his Ph.D. degree in pattern recognition and intelligent system from Northeastern University in 2014. Now he is currently a lecturer in College of Information Science and Engineering, Northeastern University, and engages in post doctoral research work in Northeastern University. His main research interests include image processing, machine learning, pattern recognition and wireless sensor networks.

Xindong You, is a post-doctoral of Beijing Institute of Graphic Communication union with Tsinghua University Before as a post-doctoral, she is an associate professor in Hangzhou Dianzi University. And before joining Hangzhou Dianzi University, she was a PhD candidate in Northeastern University from 2002 to 2007. She received her PhD degree in 2007. Her current research areas include distributed computing, Cloud Storage, Energy Management, Data Replica Management, *etc*.

