International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-03 E-ISSN: 2347-2693

Joint Feature Learning and Clustering Techniques for Clustering High Dimensional Data: A Review

Ghatage Trupti B.[1*], Patil Deepali E.[2], Takmare Sachin B.[3], Patil Sushama A.[4]
[1*,2,3]*Department of Computer Science and Engineering,*
Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India
[4]*DC Branch, Dept of Digital Communication, SSSIST Sehore.*

www.ijcseonline.org

Received: Mar/02/2016 Revised: Mar/08/2016 Accepted: Mar/22/2016 Published: Mar/31/ 2016

***Abstract*—** In many real world applications, we often face high dimensional data. Developing efficient clustering methods for high dimensional datasets may be a challenging problem because of the curse of dimensionality. Common method to deal with this is to use first dimensionality reduction approach and then cluster the data in the lower dimensions. Even though we can initially reduce the dimensionality by any approach and then use clustering approaches to group high dimensional data, performance can also be improved since these two techniques are conducted in sequence. Naturally, if we consider the requirement of clustering during the process of dimensionality reduction and vice versus then the performance of clustering can be improved. This paper presents a review of different techniques for clustering high dimensional data by joint feature learning and clustering.

***Keywords*—**Clustering, high dimensional data, feature learning, dimensionality reduction

I. INTRODUCTION

Clustering is the process of grouping elements together in such a way that elements assigned to the same cluster are more similar to each other than to the remaining data points [1]. Applications in many domains like text mining or web mining and bioinformatics usually result in very high dimensional data. Developing efficient clustering methods for high dimensional datasets may be a challenging problem because of the curse of dimensionality.

Clustering is one of the most generally used statistical tools for data analysis. K-means is a very popular clustering technique among existing techniques because of its easy programming and performance in large high dimensional data sets. But k-means is prone to local minima problem. It does not remove undesirable noise as well as it does not scale well with high dimensional data.

Another approach is called subspace clustering [10] where the attention is on selecting a small number of original dimensions or features in some unsupervised way so that clusters become more obvious in this subspace. Giving focus on the original features has the advantage of easy implementation on a database. Still, the rigidity of original dimension does not have enough flexibility to handle clusters which extends along a mixture of directions.
There is another approach to address problem of clustering high dimensional data is dimension reduction techniques. This includes principal component analysis (PCA) [9] [11] and random projections [12] Here, dimension reduction is carried out as a pre-processing step. Dimension reduction is decoupled from the clustering process. After selecting the subspace, dimension stay fixed during the clustering process. This approach is extended in the adaptive dimension reduction approach [2]. In this method the subspace is adaptively adjusted and incorporated with the clustering process.

There are some techniques which combine dimensionality reduction and clustering in a joint framework. For example, [3] [4] integrate supervised dimensionality reduction like Linear Discriminant Analysis (LDA) and k-means into the clustering framework. It performs clustering and dimensionality reduction LDA simultaneously. Considering the requirement of clustering during the process of dimensionality reduction and vice versus can improve the performance of clustering. This paper presents a review of different techniques for clustering high dimensional data by joint feature learning and clustering.

II. LITERATURE REVIEW

Many researchers have been dedicated for transforming original features and joining clustering with feature learning. Ding et al. [2] and Ding and Li [3] have combined LDA with K-means. This is done by learning a subspace and clustering alternately. They use the K-means algorithm for generating class labels and use LDA to learn the subspace alternately.

De la Torre and Kanade [4] show the benefits of clustering in a low dimensional discriminative space instead of

principal components. They proposed a new clustering algorithm named discriminative cluster analysis.
Ye et al. [13] have also studied the problem of alternating distance metric learning and clustering iteratively. Here data is projected into a low-dimensional manifold such as the separability of the data is maximized.

These methods are based on simultaneous subspace selection by LDA and clustering. Theoretical studies have also been provided in [5]. Niu et al. [14] have presented a method that reduces the dimensionality of the original data for SC. It automatically learns the related dimensions and SC simultaneously.

Gu and Zhou [15] have presented a subspace MMC method that combine dimensionality reduction with MMC in a joint framework. Domeniconi et al. [16] have introduced an algorithm which discovers clusters in subspaces covered by different combinations of dimensions using local weightings of features. They also used trace ratio for learning a subspace for face image clustering [18] and feature selection [17]. Other dimensionality reduction based clustering methods [19] [20], also exists for the simultaneous combination of dimensionality reduction and clustering.

In this paper we study different techniques for clustering high dimensional data using joint dimensionality reduction and clustering.

III. CLUSTERING TECHNIQUES USING JOINT FEATURE LEARNING AND CLUSTERING

A. *Adaptive Dimension Reduction*

Local minima is a problem in clustering where, iterations nearly always get trapped somewhere near the initial beginning configuration. Ding et al. [2] presented an algorithm Adaptive Dimension Reduction (ADM) that utilizes the concept of dimension reduction. This concept is usually utilized in clustering, classification, and lots of different machine learning and data mining applications. In this method most necessary dimensions or attributes are retained and the noisy dimensions or irrelevant attributes are removed. This results in reducing computational cost.
In most of the applications dimension reduction is used as a preprocessing step. The choice of the dimensions using principal component analysis (PCA) [9] [11] using singular value decomposition (SVD) [21] could be a common approach for numerical attributes. In all those applications once the dimensions are chosen, they stay fixed throughout the whole clustering process. The process of dimension reduction is separated from the clustering process.
In Adaptive Dimension Reduction approach, dimension reduction is carried out as a dynamic process that is adaptively adjusted and integrated with the clustering process. Here, projection method is used where the new projected dimensions are linear combination of old dimensions. Cluster membership is effectively used as the bridge for connecting the clusters defined in the reduced dimensional space or subspace and those defined in the full dimensional space. Using this relationship, clusters are discovered in the low dimensional subspace for avoiding the curse of dimensionality and then adaptively re-adjusted in the full dimension space for achieving global optimality. This process is repeated until the convergence.
The key to the effectiveness of this method lies in that working in the subspace containing true cluster centers is sufficient to find the cluster centers.
Subspace which contains cluster centers is of dimension far smaller than the original dimension in many applications. ADM is an effective way to converge to this subspace. Subspace is much easier to find than directly finding cluster centers, due to the flexibility in defining subspace.
This approach could be extended to other clustering methods. The standard practice of reporting the results obtained directly in the reduced dimension subspace is not accurate enough. Therefore, the EM in the d-dim space should be run once using the parameters obtained in the r-dim subspace to get more precise final parameters.

B. *Adaptive LDA-guided K-means Clustering*

Ding and Li [3] combined linear discriminant analysis (LDA) and K-means clustering into a coherent framework that adaptively selects the most discriminative subspace. They use K-means clustering for generating class labels and LDA for subspace selection. The clustering process is thus combined with the subspace selection process. The data are then concurrently clustered while the feature subspaces are selected.

Ding and Li [3] further extended the adaptive dimension reduction approach by explicitly combining LDA and K-means clustering in a consistent way. They show that LDA and K-means clustering are optimizing the same objective function. They both minimize the within-class scatter matrix Sw and maximize the between-class scatter matrix Sb. Based on the theoretical analysis, they show that the objective function for LDA offers a natural generalization which combine LDA and K-means clustering together.
Since LDA and K-means clustering both minimizes Sw and maximize Sb, there should be ways to combine them into a single framework. They propose to combine them into a single framework. For that they used K-means clustering to generate class labels and LDA for subspace selection. The final result of this learning process is such that data are clustered and the feature subspaces are selected simultaneously.

Adaptive dimension reduction optimization problem is then solved by well established LDA and K-means clustering algorithms.

Ding and Li [3] show that this new approach reduces to earlier approaches under various restrictions. This approach provides a generalization that has a concrete theoretical foundation as well as extremely clear and simple implementation. This adaptive dimension reduction approach can be intuitively viewed as an unsupervised LDA.

C. Discriminative Cluster Analysis

A common approach to deal with high dimensional data is to cluster in the space covered by the principal components (PC). De la Torre and Kanade [32] showed the advantages of clustering in a low dimensional discriminative space instead in the PC space. They propose a new clustering algorithm, Discriminative Cluster Analysis (DCA). It jointly performs dimensionality reduction and clustering. Since DCE uses discriminative features for clustering instead of her than generative ones, it outperforms PCA+k-means. Clustering in this space is less prone to local minima problem and removes irrelevant dimensions for clustering.

The aim of DCA is to combine dimensionality reduction and clustering in an unsupervised manner. In the first step DCA find outs a low dimensional projection of the data suitable for clustering. It encourages the preservation of distances between neighboring data points. After projecting data into a low dimensional space, DCA finds a "soft" clustering of the data. Then, this information is fed back to the dimensionality reduction step until convergence.

Clustering in this low dimensional discriminative space is more computationally efficient than clustering in the original space and also less prone to local minima. Computation is faster especially in case of high dimensional data. It removes noisy dimensions that are not useful for clustering. Moreover, it is often difficult to model correlations in high dimensional spaces. But these correlations can be modeled by projecting them into a low dimensional space.

However, it is still unclear how to select the optimal number of clusters. DCA assumes that all the clusters have the same orientation. A number of extensions could be made in the case of non-Gaussian shape clusters such as using kernel methods. However, efficiency of the solution will be lost for huge amounts of high dimensional data.

D. Discriminative K-means

Jieping Ye, Zheng Zhao and Mingrui Wu [5] shown that iterative clustering and subspace selection is equivalent to kernel K-means with a specific kernel Gram matrix. Depending on this equivalence relationship, they propose the Discriminative K-means (DisKmeans) algorithm for simultaneous clustering and LDA subspace selection. They also showed automatic parameter estimation procedure and presented the nonlinear extension of DisKmeans using kernels. They show that the learning of the kernel matrix over a convex set of pre-specified kernel matrices can be incorporated into the clustering formulation [5].

The algorithm Discriminative Clustering (DisCluster) works in an iterative fashion. It alternates between LDA subspace selection and clustering. Here, clustering generates the class labels for LDA, and LDA provides the subspace for clustering. Empirical results have shown the benefits of clustering in low dimensional discriminative spaces instead of principal component spaces (generative). However, the integration between clustering and subspace selection in DisCluster is not well understood, due to the iterative and intertwined nature of the algorithm.

Jieping Ye, Zheng Zhao and Mingrui Wu [5] made following contributions:

- They show that the LDA projection can be factored out from the integrated clustering and LDA subspace selection formulation. This result in a simple trace maximization problem related with a regularized Gram matrix of the data and it is controlled by a regularization parameter λ.
- The solution to this trace maximization problem leads to the algorithm called Discriminative K-means (DisKmeans) for simultaneous clustering and LDA subspace selection. DisKmeans is shown to be equivalent to kernel K-means, in which discriminative subspace selection essentially constructs a kernel Gram matrix for clustering. This offers new insights in the nature of this subspace selection procedure.
- The DisKmeans algorithm is reliant on the value of the regularization parameter λ. They propose an automatic parameter tuning process for the estimation of λ.
- They propose the nonlinear extension of DisKmeans using the kernels and show that the learning of the kernel matrix over a convex set of pre-specified kernel matrices can be integrated into the clustering formulation, resulting in a semi definite programming (SDP) [22].

E. Adaptive Subspace Iteration (ASI)

Document clustering has long been an important problem in information retrieval. Tao Li, Sheng Ma and Mitsunori Ogihara [6] presented a new clustering algorithm ASI, which uses explicitly modeling of the subspace structure related with each cluster. ASI simultaneously performs data reduction and subspace identification using an iterative alternating optimization procedure. Tao Li, Sheng Ma and Mitsunori Ogihara [6] provided a method to determine the number of clusters, motivated from the optimization

procedure. They also discuss the relations of ASI with various existential clustering approaches. An extensive experimental result on real data sets shows the effectiveness of ASI algorithm.

ASI algorithm simultaneously performs two tasks as data reduction i.e. assigning data points into clusters and subspace identification i.e. identifying subspace structure associated with each cluster. The tasks are carried out by an iterative optimization procedure that alternates between updating of the clusters based on the identified new subspace structures and identification of the subspace structure from present cluster partitions.
ASI clustering clearly models the subspace structure associated with each cluster via linear combinations of the original features. A nice property of ASI is that the resulting classes can be easily described in terms of features, since the algorithm explicitly exploit the subspace structure in the feature space.
ASI produces interpretable descriptions of the resulting clusters as an added bonus by explicitly modeling the subspace structure. In addition, ASI performs implicit adaptive feature selection at every iteration and flexibly calculates the distances between data points. It works well for high-dimensional data.
Experimental results suggested that ASI is a practical and competitive clustering algorithm.

F. Discriminative Embedded Clustering

Chenping Ho, Feiping Nie and Dongyun Yi [7] have proposed a framework, Discriminative Embedded Clustering (DEC). It attempts to solve the problem of clustering high dimensional data by using joint clustering and dimensionality reduction.
PCA and K-means [8] are two most commonly used methods in dimensionality reduction and clustering. However it is difficult to combine them. LDA [9] and K-means can be combined because LDA can use label information derived from K-means. But being an unsupervised dimensionality reduction approach, PCA is unable to use label information directly from K-means. Hence, the authors proposed to share the transformation matrix instead of label information between two procedures as dimensionality reduction and clustering.
There are two main objective functions of DEC, the first concerns dimensionality reduction and the second concerns clustering. Different from traditional approaches which conducts dimensionality reduction and clustering in a sequence, DEC alternates them iteratively. It combines clustering and subspace learning in a unified framework. In DEC framework [7] several previous methods can be viewed as special cases by setting different values for a balance parameter of DEC.
Since the devised problem is not joint convex with respect to two group parameters, they are updated alternatively. When one group parameter is fixed, the subproblem is joint convex to the variables, thus alternating minimization can be adopted to obtain the global optimum.
New methods can also be designed by choosing suitable balance parameters in this framework. Experimental results show that DEC out-performs related state-of-the-art clustering approaches and existing joint dimensionality reduction and clustering methods.

IV. CONCLUSION

In this paper we study different techniques for clustering high dimensional data using joint feature learning and clustering. Out of these methods Discriminative Embedded Clustering method outperforms related state-of-the-art clustering approaches and existing joint dimensionality reduction and clustering methods.

ACKNOWLEDGMENT

There have been many contributors for this to take shape and authors are thankful to each of them. They specially would like to thank Prof. Chougule A.B. (Head of Department Computer Science and Engineering (BVCOEK)) and Prof. Takmare S.B

REFERENCES

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Second Edition. Morgan Kaufmann, 2006.
[2] C. Ding, X. He, H. Zha, and H. D. Simon, “Adaptive dimension reduction for clustering high dimensional data,” in Proc. ICDM, Page No (147–154), 2002.
[3] C. Ding and T. Li, “Adaptive dimension reduction using discriminant analysis and K-means clustering”, in Proc., Page No (521–528), ICML, 2007.
[4] F. De La Torre and T. Kanade, “Discriminative cluster analysis”, in Proc. ICML, Page No (241–248), 2006.
[5] J. Ye, Z. Zhao, and M. Wu, “Discriminative K-means for clustering”, in Advances in Neural Information Proc. Systems. Cambridge, MA, USA: MIT Press, 2007.
[6] T. Li, S. Ma, and M. Ogihara, “Document clustering via adaptive subspace iteration”, 27th Annual. Int. ACM SIGIR Conf. Rese. Develop. Inform. Retri., Page No (218–225), Jul. 2004.
[7] Chenping Ho, Feiping Nie, Dongyun Yi, and Dacheng Tao, “Discriminative Embedded Clustering: A Framework for Grouping High-Dimensional Data”, IEEE Trans. Neural Network. Learn. Syst., Volume-26, no. 6, Page No (1287-1299), June 2015.
[8] J. Shi and J. Malik, “Normalized cuts and image segmentation”, IEEE Transaction. Pattern Anal. Mach.

Intell., Volume-22, no. 8, Page No (888–905), Aug. 2000.
[9] Duda R. O., Hart P. E., and Stork D. G., "Pattern Classification", 2nd edition, New York, NY, USA: Wiley, 2000.
[10] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review", ACM SIGKDD Explorations Newslett., Volume-6, no. 1, Page No (90–105), 2004.
[11] Jolliffe, I. "Principal component analysis", Springer. Second Edition, 2002.
[12] Dasgupta S. Experiments with random projection. Proc. 16th Conf. Uncertainty in Artificial Intelligence (UAI 2000).
[13] J. Ye, Z. Zhao, and H. Liu, "Adaptive distance metric learning for clustering", in Proc. IEEE CVPR, Page No (1–7), Jun. 2007.
[14] D. Niu, J. G. Dy, and M. I. Jordan, "Dimensionality reduction for spectral clustering", in Proc. Int. Conf. Artif. Intell. Statist, Volume-15., Page No (552–560), 2011.
[15] Q. Gu and J. Zhou, "Subspace maximum margin clustering," in Proc. CIKM, Page No (1337–1346), Nov. 2009.
[16] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma, "Subspace clustering of high dimensional data", in Proc. SIAM Int. Conf. Data Mining (SDM), Page No (517–521), Apr. 2004.
[17] D. Wang, F. Nie, and H. Huang, "Unsupervised feature selection via unified trace ratio formulation and K-means clustering (track)," in Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discovery Databases (ECML PKDD), Nancy, France, 2014.
[18] C. Hou, C. Zhang, F. Nie and Y. Wu, "Learning a subspace for face image clustering via trace ratio criterion", Opt. Eng., Volume-48, no. 6, p. 060501, 2009.
[19] T. Li, S. Ma, and M. Ogihara, "Document clustering via adaptive subspace iteration," in Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retr., Page No (218–225), Jul. 2004.
[20] R. W. Sembiring, S. Sembiring, and J. M. Zain, "An efficient dimensional reduction method for data clustering", Bull. Math., Volume-4, no. 1, Page No (43–58), 2012.
[21] G. Golub and C. Van Loan, "Matrix Computations", Third Edition, Johns Hopkins, Baltimore, 1996.
[22] L. Vandenberghe and S. Boyd. "Semidefinite programming", SIAM Review, 38:49–95, 1996.

AUTHORS PROFILE

Ms. Ghatage Trupti Babasaheb is a M.E. student in Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India. She has worked as Lecturer in Dr. D.Y. Patil Polytechnic, Kasaba Bawada, Kolhapur, Maharashtra, India. Her research interest lies in Data Mining, Database. She has published one International paper and presented two papers in National Level Conference.

Ms. Patil Deepali Eknath is a M.E. student in Bharati Vidyapeeth's College of Engineering, Kolhapur. Maharashtra, India. Her research interest lies in Networking and Network security. She has published two paper in National Level Conference and one international paper.

Mr. Takmare Sachin Balawant is working as Assistant Professor in Computer Science and Engineering Department of Bharati Vidyapeeth's College of Engineering, Kolhapur with Teaching experience of about 10 years. He has published about three International Papers and five National Papers.

Ms. Patil Sushama Arjun is a Student of M. Tech, DC Branch, Dept of Digital Communication, and SSSIST Sehore.