Volume 3, Issue 2, February-2016, pp. 62-66 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

High Dimensional Data Clustering with Hub Based DEC

[1]Ms. Ghatage Trupti B., [2]Prof. Takmare Sachin B.

[1]Pursuing M.E, CSE Branch, Dept of CSE

[2]Assistant Professor, Department of Computer Science and Engineering,

Bharati Vidyapeeth's College of Engineering, Kolhapur,

Maharashtra, India.

Abstract— Clustering is an important topic in various fields like machine learning and data mining. In many real applications, we often face very high dimensional data. Many dimensions are not always helpful or may even worsen the performance of the subsequent clustering algorithms. To deal with this problem one way is to employ first dimensionality reduction and then apply clustering. But if we consider the requirement of clustering in the process of dimensionality reduction and vice versus then the performance of clustering will be improved. Discriminative Embedded Clustering (DEC) is an algorithm that combines clustering and subspace learning. Hubness is the tendency of high dimensional data to have hubs. Hubs are situated near cluster centeres; therefore major hubs can be successfully used as cluster prototypes or guide during centroid based configurations. Use of hubness for clustering leads to improvement over centroid-based approaches. In this paper we propose a system for clustering high dimensional data using Discriminative Embedding Method with Hub based clustering.

Keywords— Clustering, high dimensional data, subspace learning, hubs, discriminative embedded clustering (DEC)

1. INTRODUCTION

Clustering is the process to group elements together in such a way that elements which are assigned to the same cluster are more similar to each other than to the remaining data points [1]. Clustering algorithms are roughly classified into four groups as partitional, hierarchical, density based, and subspace algorithms. In these algorithms subspace algorithms search for clusters in some lower dimensional projection of the original data. Subspace algorithms are usually preferred when dealing with high dimensional data [2] [3]. In many real applications, we often face very high dimensional data. As dimensions increases clustering become difficult. This is because, as dimensions increases sparsity of data increases and it becomes difficult to distinguish between data points. Many dimensions are not helpful or may even worsen the performance of the successive clustering algorithms. To deal with this problem one way is to employ first dimensionality reduction and then apply clustering. Even we can first do dimensionality reduction and then apply clustering on than lower dimensional space, the performance can also be improved because these processes are conducted in sequence. There are briefly two ways of performing dimensionality reduction and clustering simultaneously, first by joint feature selection with clustering and second by joint feature learning with clustering. The first method commonly called as subspace clustering. It maintains the original features during the process of feature selection. In second method several features are combined to create new representations for clustering. Chenping Hou, Feiping Nie, Dongyun Yi, and Dacheng Tao [4] proposed a framework referred to as Discriminative Embedded Clustering to address problem of clustering high dimensional data using joint dimensionality reduction and clustering. Considering the requirement of clustering in the process of dimensionality reduction and vice versus can improve the performance of

clustering. Discriminative Embedded Clustering (DEC) [4] is an algorithm that clusters high dimensional data by applying joint dimensionality reduction and clustering methods together. But different from traditional methods, DEC performs dimensionality reduction and clustering iteratively instead of sequentially.

M. Radovanovic, A. Nanopoulos, and M. Ivanovic [5] discovered a new aspect of high dimensional data referred to as hubness. In high dimensional data some data points are included in many more k-nearest-neighbor lists than other points. These points are called hubs. Hubness is the tendency of high dimensional data to contain hubs. Hubs also exist in clustered data near cluster centeres. Hubness is a good measure of point centrality inside a high dimensional data cluster [5]. It is proved that hubness can be successfully used in clustering [6]. Major hubs can be used successfully as cluster prototypes or as guides during the search for centroid-based cluster configurations.

Use of hub based algorithm instead of k-means [7] for finding initial clusters can reduce the number of iterations required in DEC. It will also improve the overall performance. In this paper we propose a system for clustering high dimensional data using Discriminative Embedding method with Hub Based clustering.

2. LITERATURE REVIEW

K. Kailing, H. P. Kriegel, P. Kroegerp, and S. Wanka (2003) [2] presented a pre-processing step for traditional clustering algorithms, that detects all attractive subspaces of high dimensional data containing clusters. For this purpose, they defined a quality criterion for the interestingness of a subspace and proposed an efficient algorithm called Ranking Interesting Subspaces (RIS) to examine all such subspaces.

L. Parsons, E. Haque, and H. Liu (2004) [8] presented a survey of the various subspace clustering algorithms together with a hierarchy organizing the algorithms by their defining characteristics. They discussed some potential applications in which subspace clustering could be mostly useful.

F. De La Torre and T. Kanade (2006) [9] proposed a novel clustering algorithm called Discriminative Cluster Analysis (DCA). DCA algorithm jointly performs dimensionality reduction and clustering. DCA is better than PCA + k-means, since it uses discriminative features for clustering rather than generative ones. They had shown the benefits of clustering in a low dimensional discriminative space rather than in the PC space (generative). Their study shows that clustering in these spaces is less prone to local minima and it also removes irrelevant dimensions for clustering. Additionally, clustering in these low dimensional discriminative spaces is more computationally efficient than clustering in the original space.

M. Radovanovic, A. Nanopoulos, and M. Ivanovic (2010) [5] discovered a new aspect of the high dimensionality i.e. hubness, that affects the number of times a point appears among the k-nearest-neighbors of other points in a data set. They showed that under commonly used assumptions as dimensions increase this distribution becomes considerably skewed. This cause the emergence of points with very high k-occurrence called hubs.

Tomasev, Radovanovic, Mladenic, and Ivanovi (2014) [6] shown that the hubness can be successfully exploited in clustering. It is the tendency of high-dimensional data to contain points that frequently occur in k-nearest-neighbor lists of other points. These points are called hubs. They demonstrate that hubness is a good measure of point centrality inside a high dimensional data cluster. They also show that major hubs can be used successfully as cluster prototypes or as guides during the search for centroid based cluster configurations. They proposed several hubness-based clustering algorithms.

Hou, Nie, Yi, and Tao (2015) [4] has proposed a framework referred to as Discriminative Embedded Clustering (DEC) which try to solve the problem of curse of dimensionality by joint dimensionality reduction and clustering. This is different from traditional approaches in which dimensionality reduction and clustering is conducted in sequence. DEC alternates them iteratively. They had done comprehensive analyses, including convergence behaviour, parameter determination, and computational complexity, with the relationship to other related approaches. A large number of experimental results have been proposed to show the efficiency of DEC.

3. OVERVIEW OF METHOD

3.1 DEC based on K-means

Chenping Ho, Feiping Nie and Dongyun Yi [4] have proposed a framework, Discriminative Embedded Clustering (DEC). It tries to solve the problem of clustering high dimensional data by using joint dimensionality reduction and clustering. PCA and K-

means [7] are two most generally used methods in dimensionality reduction and clustering. But it is difficult to combine them. LDA [10] and k-means can be combined because LDA can use label information derived from k-means. But an unsupervised dimensionality reduction approach, such as PCA unable to use label information directly from k-means. Thus, the authors proposed to share the transformation matrix rather than label information between two procedures, i.e., dimensionality reduction and clustering.

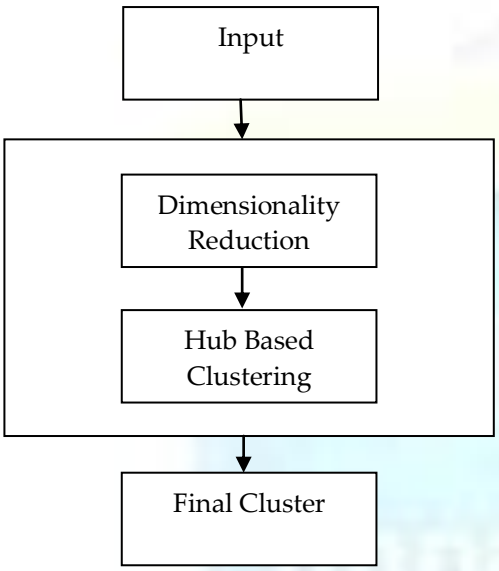

Fig.1 Architecture of Hub based DEC framework

There are two main objective functions of DEC, the first one concerns dimensionality reduction and the second one concern clustering. Different from traditional approaches that conduct dimensionality reduction and clustering in a sequence, Discriminative Embedded Clustering (DEC) alternates them iteratively. It combines subspace learning and clustering in a unified framework. In DEC framework [4] several previous methods can be viewed as special cases by setting different values for a parameter of DEC. PCAKM can be viewed as a special case of DEC when $\lambda \to 0$.

3.2 DEC based on hubs

3.2.1 Hubs

Let D be the set of data points. K-occurrences of point x is the number of times x occurs in k-nearest-neighbor lists of other points from D. With increase in the dimensionality of data, the distribution of k-occurrences becomes considerably skewed [5]. As a result, some data points are included in many more k-nearest-neighbor lists than other points. These points are called hubs. Hubness is the tendency of high dimensional data to contain hubs. Hubness score is the number of k-occurrences of points. Major hub is point with highest hubness score.

Hubs also exist in clustered data, tending to be situated near cluster centers [5]. The points closer to the mean tend to be closer to all other points, for any observed dimensionality. And in high-dimensional data, this tendency is amplified [5]. Such points will have a higher chance of being included in k-nearest-neighbor lists of other points in the data set. Hubness is a good measure of point centrality inside a high dimensional data cluster. It is proved that hubness can be successfully used in clustering [6].

3.2.2 Use of hubs in DEC

Major hubs can be used efficiently as cluster prototypes or as guides in the search for centroid-based cluster configurations. Hubness of points is straightforward to exactly compute. But the computation of cluster centroids and medoids must involve some iterative inexact method basically tied to the process of cluster construction. Centroids and medoids in K-means iterations are likely to be converging to locations close to high-hubness points, which imply that using hubs instead of centroid or medoids could actually speed up the convergence of the algorithms and leads straight to the promising regions in the data space.
In fig 2 the red dashed circle marks the centroid (C), yellow dotted circle the medoids (M), and green circles indicate two elements of highest hubness (H1, H2), for neighbourhood size 3. It shows in two dimensions what normally happens in multidimensional data. This implies that taking hubs as centers in following iterations provide quicker convergence and hence proves helpful in finding the best end configuration.

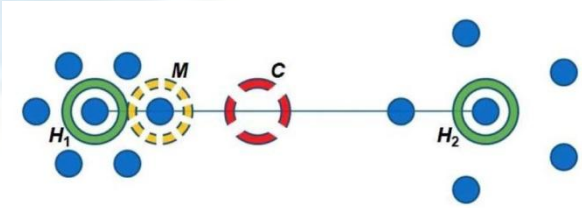

Fig.2 Example for illustrating hubs
Table.1 Notations

D	Dimensionality of High dimensional space;
d	Dimensionality of embedded surface;
n	Number of points;

c	Number of clusters;
λ	Balance parameter;
e = [1, 1, . . . , 1] $\in R^{1\times n}$	The row vector of all ones;
$x_i \in R^D$	The i^{th} high dimensional data;
$y_i \in R^d$	Low dimensional embedding of X_i;
$f_i \in R^c$	Cluster indicator of x_i
$X = [x_1, \ldots, x_n]$	Data matrix in high dimensional space;
$Y = [y_1, \ldots, y_n]$	Data matrix in low dimensional subspace;
$F = [f_1^T, \ldots f_n^T]^T$	Cluster indicator matrix;
$Q \in R^{D\times d}$	The transformation matrix;
$G \in R^{d\times c}$	The cluster centroid matrix;

Nenad Tomasev, Dunja Mladenic, Milos Radovanovic, and Mirjana Ivanovi [6] have presented three algorithms specifically for clustering high dimensional data. Out of these Hubness-proportional K-means (HPKM) algorithm use point hubness scores to guide the search, but select a centroid-based cluster configuration in the end. In this way we are hoping to avoid premature convergence to a local optimum.

The procedure for Hub based DEC is as follows,

Algorithm 1 Hub based DEC

Input:

Data set: {xi | i = 1, 2, . . . , n}, balance parameter λ.

Output:

Transformation matrix Q and cluster indicator F

Procedure:

1. Initialize Q by PCA, i.e., by solving the problem in Eq. (1)

   Initialize F by conducting HPKM on $Q^T X$.

2. Alternatively update Q, G and F until convergence.

   a. Update F by comparison rule in Eq. (2)

   b. Update Q by picking up eigenvectors corresponding to the d largest eigenvalues of

   $S_t - \lambda XX^T + \lambda XF(F^TF)^{-1}F^TX^T$. Update G by Eq. (3).

In hub based DEC the transformation matrix Q is initialized by PCA. The variance of embedding point is measured by following eq.

$$\max T_r(Q^T XX^T Q)$$

$$=\max T_r(Q^T \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T Q)$$

$$=\max T_r(Q^T XX^T Q)$$

.....Eq. (1)

We constrain $Q^TQ = I$, therefore the optimal solution of PCA can be obtained by eigen-decomposition of S_t or XX^T. Here, S_t is total variance matrix in LDA.

The comparison rule for updating Q is as follows,

$$F_{i+1}^* = \begin{cases} F_{i+1}^j, & \left\|(Q_i^*)^T X - G_i^*(F_{i+1}^j)^T\right\|_F^2 \\ & < \left\|(Q_i^*)^T X - G_i^*(F_i^*)^T\right\|_F^2 \\ F^*, & \text{Otherwise} \end{cases}$$

.....Eq. (2)

The rule for updating cluster centroid matrix G is as follows,

$$G = Q^T XF(F^TF)^{-1}$$

.....Eq. (3)

Use of hubness for clustering leads to improvement over centroid-based approaches. In DEC procedure the cluster indicator matrix is initialized by conducting k-means. If we use hub based algorithm instead of k-means for finding initial clusters then the number of iterations required in DEC will be reduced. It will also improve the overall performance.

4. CONCLUSION

In this paper we proposed a system for clustering high dimensional data. The hub based DEC algorithm is proposed which combines the concept of hub with Discriminative Embedded Clustering (DEC). Hubness is good measure of point of centrality within high dimensional data clusters. Therefore this algorithm will be helpful for minimizing the number of iterations required to calculate the initial cluster in DEC and in turn will improve the performance of DEC.

ACKNOWLEDGMENT

There have been many contributors for this to take shape and we are thankful to each of them. We specifically 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 ed. Morgan Kaufmann, 2006.

[2] K. Kailing, H. P. Kriegel, P. Kroegerp, and S. Wanka, "Ranking Interesting Subspaces for Clustering High Dimensional Data," Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003.

[3] C.C. Aggarwal and P.S. Yu, "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. 26th ACM SIGMOD Int'l Conf. Management of Data, pp. 70-81, 2000.

[4] Chenping Ho, Feiping Nie, Dongyun Yi, and Dacheng Tao, "Discriminative Embedded Clustering: A Framework for Grouping High-Dimensional Data", IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 6, pp.1287-1299, June 2015.

[5] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data", J. Machine Learning Research, vol. 11, pp. 2487-2531, 2010.

[6] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovi, "The Role of Hubness in Clustering High-Dimensional Data", IEEE Trans. Knowledge and Data Eng., vol. 26, no. 3, pp.739-751, March 2014.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888–905, Aug. 2000.

[8] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review", ACM SIGKDD Explorations Newslett., vol. 6, no. 1, pp. 90-105, 2004.

[9] F. De La Torre and T. Kanade, "Discriminative cluster analysis", in Proc. ICML, 2006, pp. 241-248.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", 2nd ed.New York, NY, USA: Wiley, 2000.

AUTHOR PROFILE

Ms. Ghatage Trupti B. 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 a paper in National Level Conference.

Mr. Sachin Balawant Takmare 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 3 International Papers and 5 National Papers.

