International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3, February 2019

Comparison between Subspace and Conventional Clustering for High Dimensional Data Analysis

Kahkashan Kouser, Amrita Priyam

Abstract: ***Clustering High dimensional data is a propitious research area in current scenario. Now it becomes a crucial task to cluster multi-dimensional dataset as data-objects are largely dispersed in multi-dimensional space. Most of the conventional algorithms for clustering work on all dimensions of the feature space for calculating clusters. Whereas only few attributes are relevant. Thus their performance is not very Precise. A modified subspace clustering is proposed in this research paper, which does not use all attributes of high-dimensional feature space simultaneously rather, it determine a subspace of attributes which are important for each individual cluster. This subspace of attributes may be same or different for the different cluster. The comparison between conventional K-Means and modified subspace K-means clustering algorithms were done based on various validation matrices. Results of the modified subspace clustering is compared with the conventional clustering algorithm. It was analyzed based on different matrices such as SSE(sum of squared error), WGAD-BGD (Within group average distance minus between group distances) and DBI(Davies-Bouldin index) or Validity index. Artificial data set were used for all the experiments. Results represent the better efficiency and feasibility of modified subspace clustering algorithm over conventional clustering methods.***

Index Terms: ***clustering, high-dimensional data, Subspace clustering, COSA, clustering on subset of attribute***

I. INTRODUCTION

The act of mining or eradicating useful information from massive data saved inside the database is defined as Data mining. This technique is particularly used to extract information from a data collection and convert into a comprehensible shape which could be use in decision making [1]. Clustering is unsupervised data mining task. It constructs physical class of identical data-objects, identified as cluster. So, in cluster data-objects are much similar to each other. At the same time, they are different from data object of another cluster. [2] Conventional strategies do not provide desired result despite continuous research on clustering techniques over the long time mainly in the fields of database, statistics and machine learning. In real-world situations when clustering is applied on large-dimensional space, most of the algorithms are suffering from various key challenges. Data objects in distinct clusters are usually associated with each other by using some sets of features, i,e. a cluster can additionally occur in one of a kind subspace or in a certain subspace of all dimensions. It is found that data-objects are actually some distance apart from one another in some dimensions of high-dimensional space. Recently various subspace clustering techniques were evolved to overcome this problem. The intention of subspace clustering is to discover clusters in exclusive subspace or in certain subspace of the original feature space. Subspace clustering is a suitable technology concerned with the clusters which are identified totally on the basis of their association with subspace of high dimensional feature space [3, 14]. Subspace clustering algorithms are categorized into two main classes as Soft subspace clustering and Hard subspace clustering. Soft subspace clustering attaches weight to every feature on the basis of feature contribution for constructing distinct cluster in the clustering process. In soft subspace clustering method, an initial weight is attached to individual feature by using a random approach and then allocates each object to the cluster. The process is repeated iteratively to refine the and clusters [11]. Example of some soft clustering algorithm are Fuzzy Subspace clustering given by Gan and Wu in 2008. Fuzzy Weighted K-means algorithm given by Jing et al., in 2005 and Attribute Weighted algorithm given by Chan et al. in 2004. Whereas in hard subspace clustering, a particular subset of the feature is chosen for individual cluster and remaining other features are rejected i.e. hard subspace clustering method split the features space into distinct subspace where individual feature can be member of a particular subspace or not [4]. It is further categories into two categories primarily on the basis of their searching strategy. They are top-down and bottom-up hard subspace clustering. Top-down approach locates clusters initially in full-dimensional feature space after that it determines the subspace for individual clusters iteratively. Top-down start via thinking that clusters made of every attribute in the dataset. The attributes are incrementally removed and the status of evolved clusters is evaluated. The process is continued until the overall functioning of the algorithm attains a desired level. During the beginning of the process the clusters are constructed by using all attributes [5]. Examples of some Top- down approach are given as follows. In 1999 Aggarwal et al. has given PR Ojected CL Ustering (PROCLUS).

Manuscript published on 28 February 2019.
***** Correspondence Author (s)
Kahkashan Kouser*, Research Scholar Department of Computer Science, Birla Institute of Technology, Ranchi 834001, India.
[2]Assistant Professor, Department of Computer Science & Engineering, Gaya College of Engineering, Gaya 823003, India
Amrita Priyam, Associate Professor, Department of Computer Science, Birla Institute of Technology, Ranchi 834001, India.

In the year 2000, Aggarwal and Yu developed Arbitrarily Oriented Projected Cluster generation (ORCLUS). Further in the year 2004 Woo et al. has given Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting (FINDIT).
In 2004 Friedman and J. J. Meulman has developed Clustering on Subset of Attribute(COSA). Bottom-up approach firstly locates dense regions in low-dimensional feature space. After that integrates them in the shape of clusters.
Some data structure such as, windows, cells are prerequisite for using Bottom-up approach. Examples of some bottom-up clustering algorithm are as follows. In 1998 Aggarwal et al. has developed CLIQUE. In 1999Goil et al. has given MAFIA. In 1999 Cheng et al. has given ENCLUS and SUBCLU is given by Kailing et al. in 2004.

II. METHODOLOGY

The modified Subspace clustering algorithm have two steps: In first step COSA algorithm was applied, which identifies subspace for each individual cluster. This subspace is consisting of most important attribute for each individual cluster, and may be same or different for each individual cluster. Whereas, in second step k-means clustering algorithm applied on the subspace obtained in the first stage for the formation of the cluster.

A. *Clustering on subset of attribute (COSA)*

J. H. Friedman and J. J. Meulman in 2004[6] introduce an algorithm known as Clustering on subset of attribute (COSA). This algorithm is basically used to determine the set of attributes which construct the subspace for each and every cluster. In the subspace of each cluster, larger weight is assigned to the attribute which have smaller dispersion i.e. the attribute which are important for a particular subspace. the execution of algorithm, gives as output a distance matrix. which can also use as input for any further clustering algorithm. Steps in COSA algorithm are shown below:

1. Assign equal weight to each dimension of full-dimensional feature space, with the constrain sum of weight of all dimensions are equal to 1.

i.e., $\sum_{k=1}^{n} W_k = 1$

2. Calculate distance matrix with the help of weight allocated to individual dimension in step1.this distance matrix δijk, gives distance among each pair of data objects. the value of distance matrix is obtained by using following formula.

$\delta ijk = |xik - xjk|$

3. Calculate closeness between the attributes for all clusters by using weight assigned to attributes and the distance matrix.

$$Sk = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ijk}$$

4. Based on of this closeness measure, modify the distance matrix δijk to create a modified distance matrix Dijk , in which the individual distance is divided by the Sk as follows,

$$Dijk = \frac{\delta ij}{Sk}$$

5. now modify Dijk by multiplying the weight of respective attribute for every cluster, Wkl, which is computed in step1. As follows,

$$Dij[W] = \sum_{k=1}^{n} W_{kl} \cdot d_{ijk}$$

6. Run any clustering algorithm such as k-means, KNN which uses the new modified Dij[W] distance matrix for calculating distance between the objects.
7. After getting clusters of objects we calculate dispersion of data objects in the Kth attribute for Lth cluster by using formula.

$$Skl = \frac{1}{N_l^2} \sum_{c(i)=l} \sum_{c(j)=l} d_{ijk} \quad \ldots.(I)$$

i.e.,

$$Skl = \frac{\text{sum of distance between all objects in } lth \text{ cluster using } kth \text{ attribute}}{\text{total number of objects in } lth \text{ cluster}}$$

8. calculate the weight of attribute for every cluster with the help of following formula.

$$Wkl = \frac{\exp^{(-skl/\Lambda)}}{\sum_{k'=1}^{n} \exp^{(-skl/\Lambda)}} \quad \ldots..(II)$$

Where,

$$Wkl = \frac{\text{disperion of } kth \text{ attribute in } lth \text{ cluster}}{\text{disperison of all attribute in } lth \text{ cluster}}$$

the attribute with smaller dispersion within every group lth will receive larger weights. Parameter Λ is used for controlling the rate of increase.
9. Change value of Λ by
Λ= Λ+ Λ.n
go to step 5 and continue entire procedure until weight matrix get establish.
Note: In step 5 of 1st iteration the weight of each attribute for every cluster represented by Wkl is same for every cluster, but in later 2nd and successive iteration the weight of Wkl is different for all cluster so for calculating the Dij[w] use following formulas.

B. *1st formula*

$$Dij^{(1)}[w] = \sum_{k=1}^{n} \max(W_k c(i/w), W_k c(j/w)) \cdot d_{ijk} \quad \ldots.(III)$$

Whereas, $\sum_{k=1}^{n} \max(W_k c(i/w), W_k c(j/w))$ denotes maximum weight of any attribute k obtained from all clusters. First determine the maximum weight of each attribute, then multiply this maximum value of weight to dijk. to evaluate distance matrix Dij[w].

C. *2nd formula*

$Dij^{(2)}[w] = \max(D_{ij}[Wc(i/w)], D_{ij}[Wc(j/w)])$ ….(IV)
In this method first find value of distance between objects in every dimension by multiplying the weight of that particular dimension for each cluster. Then choose maximum distance to construct distance matrix Dij[w].

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

III. RESULT AND DISCUSSION

Functioning of both algorithm conventional clustering and subspace clustering is tested on the dataset with numeric attributes.

Both clustering algorithms are applied on Synthetic Fisher's IRIS data set which contains eight numeric attributes and total 300 data-objects of three species Setosa, Virginica and Versicolor. For implementation all the programs are written in C language.

It compares the conventional clustering and subspace clustering in term of SSE, WGAD-BGD, and DBI. The corresponding graphs are shown in fig1-3 to plot the SSE, WGAD-BGD, DBI values.

Table 1: Meaning of Notation

Notation	Meaning
N	Total number of data objects present in a given dataset
D=d (xi,xj)	"distance Matrix" between data points
C_1,……..C_k	Group of K clusters
K	Total number of clusters
|Ci|	Total data elements present in cluster *i*
x_1,….x_N	Set of data points

A. *Sum of squared error (SSE)*

For measuring intra-cluster cohesion SSE is a useful tool. SSE helps in determining cohesion of data objects inside cluster. To find value of SSE the squared distance of all data object Xj to its cluster centroid Cm is added. value of SSE a cluster configuration can obtained as below [8].

Sum of square error=$\sum_{i=1}^{i=k}\sum_{j=1}^{j=|c_i|} dist(x_j, C_m)^2$

The cluster configuration that have lower value of total SSE will be better.

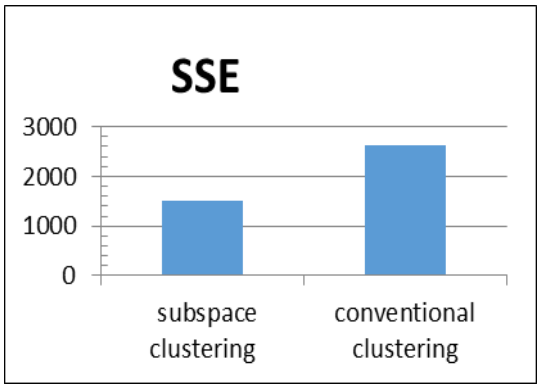

Figure I: Comparison of SSE for conventional clustering and Subspace clustering.

The above graph demonstrates that the value of SSE of subspace clustering is less than conventional clustering .so subspace clustering is superior to conventional clustering algorithm.

B. *BGD (between group distances)*

BGD (Between Group Distance) is method of determining distance of the clusters from each other. Basically it is the distance of center of every cluster to the center of cluster of whole data set. Larger is the separation, outcome is superior cluster configuration. BGD is find out by using formula given below.

TOTAL BGD = $\sum_{i=1}^{i=k} dist(c_i, c)$

C. *WGAD (Within group average distance)*

Excellence of clustering method is measured with the aid of cohesion in the cluster. WGAD (Within group average distance) is used for measuring cohesion inside the cluster. The minimized value of WGAD shows higher cohesiveness of the cluster. WGAD (Within group average distance) can be measured by formula given below.

TOTAL WGAD = $\sum_{i=1}^{i=k} \frac{\sum_{j=1}^{j=|c_i|} dist(x_j, c_i)}{|c_i|}$

to gain a fine cluster configuration, try to minimize the WGAD and maximize BGD So, both cohesion and separation can combine into a single unit called Eval ,that may be determine by the following formula

Eval=WGAD-BGD

So, for obtaining a good cluster configuration one should try to minimize *Eval*.

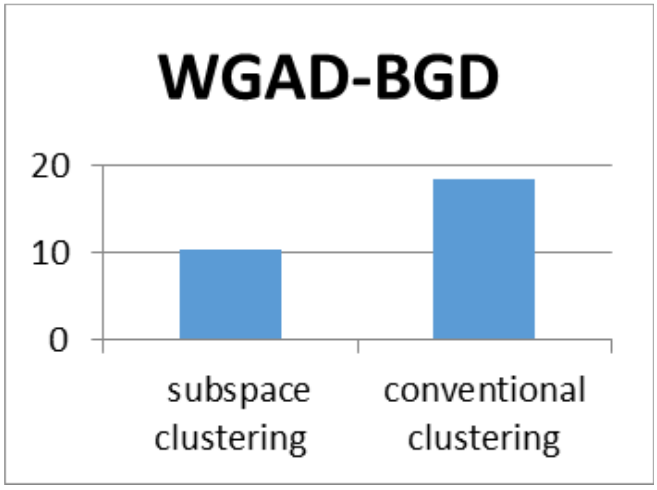

Figure II: Comparison of WGAD-BGD for conventional clustering and Subspace clustering.

The above figure demonstrates that value of WGAD-BGD for subspace clustering is less than conventional clustering. Thus it's determined the subspace clustering is superior to conventional clustering algorithm.

D. *DAVIES-BOULDIN Index*

Davies-Bouldin Index is a mixture technique that measures the couple of level of Intra-cluster cohesion and level of inter-cluster separation. Davies-Bouldin index for specific cluster configuration is obtained by adding up the greatest proportions of the Intra-cluster separation to the inter-cluster separation of every cluster.

Davies-Bouldin Index = $\frac{1}{k}\sum_{i=1}^{i=k} R_i$

Here R_i is greatest proportion among each cluster *i* and another cluster *j*. In which $1 \leq j \leq K$ and $j \neq i$. For obtaining a desirable cluster design try to reduce the estimate of DBI . Unique proportion R_{ij} representing intra-cluster separation to the inter-cluster separation of *ith* cluster respecting *jth* cluster is computed as [8]:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$R_{ij} = \frac{S_i - S_j}{D_{ij}}$$

Here D_{ij}, *is* separation of centroid of *ith* cluster with centroid of *jth* cluster. Si and S_j represents average separation of data-objects from its own cluster. The value of S_i and S_j is determined as follows:

$$S_j = \frac{1}{|c_i|} \sum_{n=1}^{n=|c_i|} d(x_n, c_i)$$

Here x is data-object inside cluster *i.* centroid of cluster *i* is represented by C_i.

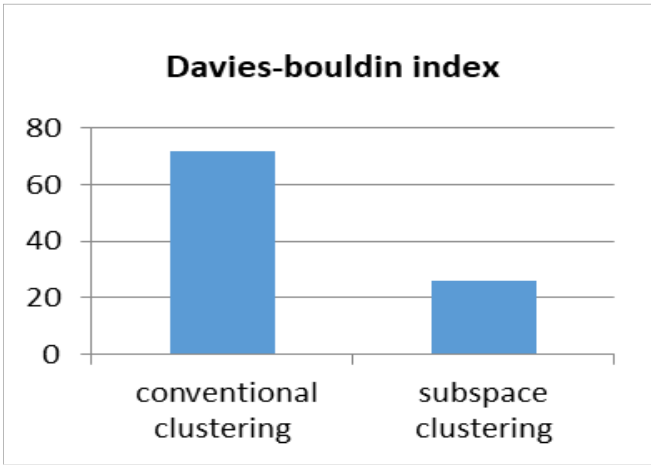

Figure III: Comparison of Davies-Bouldin index for conventional clustering and subspace clustering

The above graph demonstrates that the value of Davies-Bouldin Index of subspace clustering is less than conventional clustering .so subspace clustering is superior to conventional clustering algorithm.

IV. CONCLUSION

In the paper comparison between modified subspace clustering algorithm and conventional clustering algorithm has been done. The overall performance of each algorithms is analyzed primarily based on different validation matrices such as, sum squared blunders (SSE), Davies-Bouldin index (DBI) and WGAD- BGD. Based on the experimental result as depicted in the Fig1, 2, 3, it is determined that the above mentioned validation matrices have better value for modified Subspace clustering than conventional clustering algorithm. So it is not feasible to apply conventional clustering algorithm for extremely massive datasets. hence, one can draw the conclusion that the subspace clustering algorithm offers more appropriate way of clustering multi-dimensional data. The limitation of the above algorithm is that it works only on the numeric dataset. In future it may be extended to handle all kind of data.

REFERENCES

1. H.J.Sun,L.H.Xiong ,Genetic Algorithm based high dimensional data clustering techniques. *IEEE 2009 Sixth International Conference on Fuzzy System anf Knowledge Discovery.*
2. M. Anusha ,J.G.R. Sathiaseelan , Feature Selection using K-Means Genetic Algorithm forMulti-objective Optimization .*Procedia Computer Science 57 (2015) 1074 – 1080.*
3. Z, Deng, K.C.Choi, Y. Jiang, J. Wang, S. Wang, A Survey on Soft Subspace Clustering, *Information Sciences: an International Journal, June 2016, pp. 84-106*
4. A. Surekha, S. Anuradha, B. Jaya Lakshmi, B. Madhuri , A survey on hard subspace clustering algorithms ,*International Journal of Science & Engineering Development Research, August 2016.*
5. S. Jahirabadkar, P. Kulkarni , Clustering for High Dimensional Data: Density Based Subspace Clustering Algorithms, *International Journal of Computer Applications, February 2013.*
6. J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *Journal of the Royal Statistical*
7. *Society: Series B (Statistical Methodology), 66(4):825-849, 2004.*
8. S. S. Singh, N. C. Chauhan, K-means v/s K-medoids: A Comparative Study, *National conference on Recent Trends in Engineering &Technology, Jan2014*
9. S. Chaimontree, K. Atkinson, F. Coenen, Best Clustering Configuration Metrics: Towards Multiagent Based Clusterin. , *Advanced Data Mining and Applications, Volume 6440 of the series Lecture Notes in Computer Science pp 48-59.*
10. Q.Liu, J.Zhang, J.Xiao, H.Zhu, Q.Zhao A Supervised Feature Selection Algorithm through Minimum Spanning Tree Clustering, *2014 IEEE 26th International Conference on Tools with Artificial Intelligence.*
11. Manolis C. Tsakiris and Ren´e Vidal "Filtrated Spectral Algebraic Subspace Clustering" 2015 IEEE International Conference on Computer Vision Workshops.
12. M. C. Tsakiris and R. Vidal. Abstract algebraic subspace clustering. CoRR, ABS/1506.06289, 2015.
13. Hongchang Gao, Feiping Nie, Xuelong Li, Heng Huang,Multi-View Subspace Clustering. 2015 IEEE International Conference on Computer Vision.
14. E.Castro,X.Pu A simple approach to sparse clusterin, *Computational Statistics & Data Analysis*
15. *Volume 105 Issue C, January 2017 Pages 217-228*
16. H.P.Kriegel, E.Ntoutsi .Clustering high dimensional data: Examining differences and commonalities between subspace clustering and text clustering A position pape. *SIGKDD Explorations Volume 15, Issue 2.*
17. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Academic Press, 2nd edition, 2006.
18. https://en.wikipedia.org/wiki/Clustering high-dimensional_data

AUTHORS PROFILE

Ms Kahkashan Kouser is presently working as Assistant professor in the Department of Computer Science and Engineering at Gaya College of Engineering, Gaya. she is pursuing Ph.D. from Birla Institute of Technology, Ranchi. She has done M.Tech**.** in Computer Science from Birla Institute of Technology, Ranchi(Jharkhand). She has published three papers in National journal and attended several Conferences and workshops. Her area of interest includes Data mining, Artificial Neural Network and Cluster Algorithm.

Dr.(Mrs.) Amrita Priyam is Associate Professor in the Department of Computer Science and Engineering at BIT Ranchi Campus. She has done M.Tech. in Computer Science and Ph.D. in Engineering. Her area of interest lies in Object Oriented Modelling and Design, Artificial Intelligence, Software Engineering, Soft Computing Techniques, Decision Models, Optimization Technique.She has credit of publishing papers in various journals. She has been reviewer of many papers. She is Member of International Association of Engineers and Life Member of Indian Society for Technical Education.

.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)