Efficient Density-Based Subspace Algorithms for High-Dimensional Data

Ms. M Pallavi
Assistant Professor
CSE Department
MNR College of Engineering and Technology, Sangareddy, Medak.

***Abstract* - Density-based Clustering algorithms are fundamental technology's for data clustering with many attractive properties and applications. In high dimensional data, clusters are embedded in various subsets finding of dimensions. Density based subspace clustering algorithms treat clusters as the dense regions compared to noise or border regions. The major challenge of high dimensional data is Curse of dimensionality, means that distance measures become increasingly meaningless as the number of dimensions increases in the data set. Another major challenge is, the high dimensional data contains many of the dimensions often irrelevant to clustering. These irrelevant dimensions confuse the clustering algorithms by hiding clusters in noisy data. The task is to reduce the dimensionality of the data, without losing important information.**

***Index Terms* - Clustering, DBSCAN, CLIQUE, SUBCLU, PROCLUS, MAFIA, Optigrid, SUBCLU, PreDeCon, INSCY, DENCLUE, DISH, DENCOS**

I. INTRODUCTION

Data mining is used to extract the useful information from a collection of databases or data warehouse. Data mining helps to predict future data trends, and can be used as a reliable basis in the decision making process. Clustering is the process of grouping similar objects that are different from other objects. It is an unsupervised classification technology, means it does not have any prior knowledge of its data. Clustering is one of the major task in data mining. The main aim of clustering algorithm is segmenting a collection of objects into clusters or subsets, such that objects within one cluster are more closely related to one another than to objects assigned to different clusters [1]. Typical of high dimensional data can be found in the areas of pattern recognition, molecular biology [2], CAD (computer aided design) databases, and computer vision applications and so on. The major challenge of high dimensional data is Curse of dimensionality, means that distance measures become increasingly meaningless as the number of dimensions increases in the data set. Another major challenge is, the high dimensional data contains many of the dimensions often irrelevant to clustering. These irrelevant dimensions confuse the clustering algorithms by hiding clusters in noisy data. The task is to reduce the dimensionality of the data, without losing important information.

1.1 Subspace Clustering

Subspace clustering is the problem of finding a multi-subspace representation that best fits a collection of points taken from a high-dimensional space. In case of high dimensional data, clusters are embedded in various subsets of dimensions. To overcome this problem, recent research is focusing on a new clustering technique called "subspace Clustering". It mainly focus on clusters which are embedded in difference subspaces of high dimensional datasets.

Fig. 1 shows the subspace clusters .The image on the right shows a two-dimensional space where a number of clusters can be identified. In the one-dimensional subspaces, the clusters C_a (in subspace $\{x\}$) and C_b, C_c, C_d(in subspace $\{y\}$) can be found. C_c cannot be considered as a cluster in a two-dimensionsl subspace ,because it is too sparsly distributed in the x axis. In two dimensions, the two clusters C_{ab} and C_{ad} can be identified.

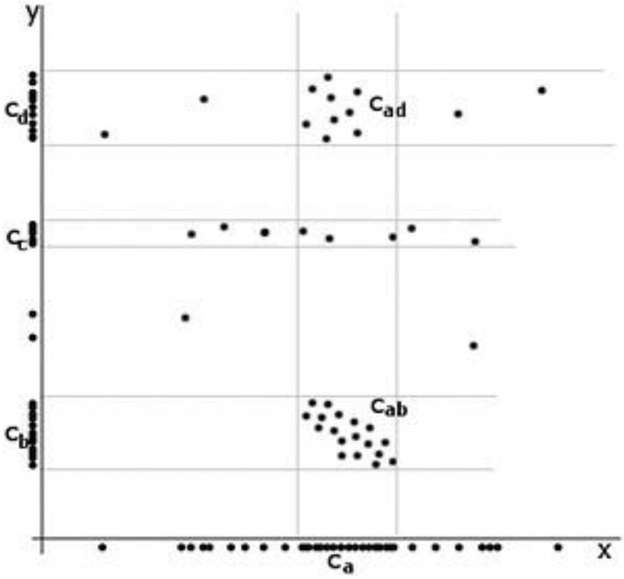

Figure 1: 2D space with subspace clusters

Once the subspaces with higher probability of comprising good quality clusters are identified, various clustering algorithms can be applied to find the hidden clusters.

1.2 Density Based Subspace clustering

In Density-Based method, clusters are made according to the density of the data. Here clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas that are required to separate clusters are considered to be noise points and border points. The data points are distributed arbitrarily in these areas of high density and may contain clusters of arbitrary size and shape.

Consider the Density based clustering of applications with noise (DBSCAN)[6], this is the first density based approach. It is based on the concept of density reachability. Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points. First it arbitrary selects a point p, and then retrieves all points density-reachable from w.r.t Eps and MinPts. If p is a core point, a cluster is formed. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database. It continues the process until all of the points have been reached.

CLIQUE is the first density based subspace clustering technique designed for high dimensional data. Clique can be considered as both density based and grid based. It detects subspaces of highest dimensionalities. Wavecluster[7] is density based clustering approach. It is a novel clustering approach based on wavelet transforms. To detect clusters it uses grid. It is applicable to only low dimensional dataset. DENCLUE [8] (density-based clustering) is density based clustering algorithm. This algorithm is used to find clusters in high dimensional data. A subspace is defined as dense subspace, if it contains many objects according to given threshold. SUBCLUE (density-based Subspace Clustering method relies on the density-based clustering and it is extension to DBSCAN.

II. Literature Survey

According to Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu [6], large spatial databases should have some minimal requirements to determine input parameters and to discover clusters with arbitrary shape. Martin etl proposed an algorithm called DBSCAN, depends on Density-based technique. This method discovers arbitrary shaped clusters also. Martin etl performed an experimental evaluation regarding efficiency and effectiveness of DBSCAN using real data of the SEQUOIA 2000 benchmark. The results of the experiments are DBSCAN discovers arbitrarily shaped clusters very effectively than CLARANS algorithm and more efficient than CLARANS.

According to Priyanka Trikha and Singh Vijendra [21], the existing algorithms do not meet the multiple requirements. DBSCAN algorithm is a density-based algorithm, based on the density of the data points. It can discover clusters with arbitrary shapes. Priyanka trikha and Singh Vijendra proposed a new method called kd-tree based on the DBSCAN. This method increases the efficiency of memory.

According to Sunitha Jahirabadkar and Parag kulkarni [10],” Subspace clustering is an evolving methodology, it aims at finding clusters in various overlapping and non-overlapping subspaces for high dimensional data. There are many clustering algorithms are there in data mining. It is impossible for the future developers to select best clustering algorithm. Sunitha Jahirabadkar and Parag Kulkarni presented a review of various density based subspace clustering algorithms together with a comparative table focusing on different characteristics such as axis parallel/arbitrary oriented, over lapping and non-overlapping. According to Sunitha Jahirabadkar and Parag Kulkarni density-based clustering methods are more efficient than other clustering methods like Partition-based clustering, Grid-based clustering, Model-based clustering methods etc.

According to Rahmat Widia Sembiring, Jasni Mohamad Zain [25], in high dimensional data, clustering faces a problem called ‘ curse of dimensionality ‘. Normally data consists of many dimensions. By using density based approach multi- dimensional data clustering can be done based on paradigm introduced by DBSCAN clustering method. With the changes in the density of each object, cluster changes will occur. Here density of each object neighbors with Minpoints will be calculated. By using Euclidean distance the neighbors of each object is determined. Rahmat Wid Sembiring and Jasni Mohammad Zain proposed SUBCLU, FIRES, INSCY clustering methods for better performance , efficiency and accuracy.

According to Hans Peter Kriegal, Peer Kroger, Matthias Renz, Sebastian Wurst, [19], In high-dimensional data, traditional algorithms often fails to detect meaningful cluster. There are some issues like, a) algorithms typically scale exponentially with the data dimensionality and subspace dimensionality of the clusters, b) Many algorithms use global density threshold for clustering, since the clusters in subspaces of different dimensionality will exhibit varying densities. Hans Peter Kriegal etl proposed a framework relies on efficient filter refinement architecture which deals with most quadratic w.r.t to dimensionality of data and dimensionality of subspace clusters. This method can be applied for any clustering notions that are based on local density threshold. This method achieves a significant gain of runtime and quality when compared to other clustering methods.

III. Problem Statement

In data mining many clustering algorithms exists like, Partition-based clustering method, Hierarchical clustering method, Grid-based clustering method, Model-based clustering method etc. Clustering high-dimensional data is an important task in cluster analysis .All the mentioned clustering methods are not efficient for clustering high-dimensional data. They cannot find clusters with complex shapes and sizes and some clustering methods are very expensive i.e., cost of processing is high. In the proposed method we are using Density based Subspace Clustering Algorithm (DBSCA) which is more dominant than the existing methods. This DBSCA handles noise and works efficiently in high dimensionality.

IV. Issues and Challenges

High Dimensional data clustering has been a major challenge due to the inherent sparsity of the points. We have several clustering methods in data mining. In Partitioning Method, it creates an initial partitioning. To achieve global optimality in partitioning-based clustering it requires the inclusion of every possible object’s information of the possible partitions. It uses algorithms like k-means, where each cluster is represented by a mean value of the objects in the cluster, and k-mediods, where each cluster is represented by one of objects located near the center of the cluster. Partition-based clustering method cannot find clusters in arbitrary shape. In Hierarchical method, it creates a hierarchical decomposition of the given set of the data objects. This method can be classified as Agglomerative (bottom-up) and divisive approach (top-down). This method suffers from the fact that once a step (split or merge) is done, it can never be undone. This technique cannot correct erroneous decisions. In grid-based method, it uses a multi resolution grid data structure. STING is a grid-based method, which explores statistical information stored in the grid cells. The quality of STING clustering depends on the granularity of the lowest level of the grid structure. If the granularity is very fine, the cost of processing will increase substantially. If the bottom level of the grid structure is too coarse, it may reduce the quality of the cluster analysis. Most clustering methods are based on the distance between the objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shape. Density-based clustering methods group neighboring objects into clusters based on local density conditions rather than proximity between objects. These methods regard clusters as dense regions being separated by low density noisy regions. Density-based methods have noise tolerance, and can discover non-convex clusters. Density-based techniques encounter difficulties in high dimensional spaces because of the inherent scarcity of the feature space, which in turn, reduces any clustering tendency. In data mining clustering methods like DBSCAN, CLIQUE, SUBCLU, PROCLUS etc, these are Density-based clustering techniques. These algorithms are developed to discover clusters with arbitrary shape, density in spatial databases with noise. Density-based clustering methods works efficiently in case of high-dimensional data when compared to other clustering algorithms.

V. Subspace Clustering Algorithms: Density-based

In this section, various algorithms for density-based subspace clustering have been reviewed on the basis of their characteristics.

5.1 CLIQUE (Clustering In Quest)

CLIQUE [3] can be considered as both density based and grid based. CLIQUE can automatically finds subspaces with high density clusters. It automatically identifies subspaces of a high dimensional data space that allow better clustering than original space. It partitions each dimension into the same number of equal length interval. It partitions an m-dimensional data space in to non-overlapping rectangular units. A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter. A cluster is a maximal set of connected dense units within a subspace. It partitions the data space and find the number of points that lie inside each cell of the partition. It identifies the subspaces that contain clusters using the Apriori principles. It identifies the clusters by determining dense units in all subspaces of interest and determines connected dense units in all subspaces of interest. After determining all dense units, clusters are found as maximal set of connected dense units. The clusters of highest dimensionality are determined by the size of grid and global density threshold. The accuracy and the efficiency of CLIQUE depends on granularity and positioning of the grid.

5.2 MAFIA (Merging of Adaptive Finite Intervals)

MAFIA [12] is advancement for CLIQUE approach. It achieves higher efficiency by using non-uniform grid cells and find better clusters. This approach partitions each dimension using a variable number of adaptive intervals, which reflects the distribution of the data .MAFIA, starts with a large number of small intervals for each dimension and then combines adjacent intervals of similar density to end up with a smaller number of larger intervals.

5.3 PROCLUS (Projected Clustering)

PROCLUS[14] is focused on a method to find clusters in small projected subspaces for data of high dimensionality. It finds a method for finding regions of greater density in high dimensional data in a way which has good scalability. Proclus is a first top-down partition algorithm and it is based on k-mediods clustering concept. Iteratively it computes mediods for each cluster on a data sample using a technique called greedy hill climbing. We need two parameters for this computation 1) number of clusters 2) average number of dimensions. The average distance between the data points and the nearest medoid determines the cluster's quality. PROCLUS is faster than CLIQUE[3].

5.4 FIRES (Filter Refinement Subspace Clustering)

FIRES[16] is an efficient subspace clustering. It works with all clustering method. It uses 1D clusters that can be constructed with any clustering method and merge these 1D clusters to generate approximations of subspace clusters. A refinement step can compute the true subspace clusters by using any clustering method. FIRES consist of three steps:

A. Preclustering: In the first step all 1D clusters called base clusters are computed. This can be done by any clustering method, because it is similar to existing subspace approaches.

B . Generation of Subspace cluster Approximations: In this step , the base clusters are merged to find maximal-dimensional subspace cluster approximations. These clusters are not merged by using Apriorimethod, it uses an algorithm that scales at most quadratic w.r.t the number of dimensions.

C. Post processing of Subspace Clusters: The third step can be applied to refine the cluster approximations which are retrieved after the second step.

The benefits of this method are less runtime and it produce accurate results.

5.5 Optigrid (Optimal Grid clustering)

Optigrid is a grid based clustering method. It first generates histograms of data values of all dimensions. If the dimensionality is not too high, then it determines the noise level by manually inspecting the histogram, otherwise needs to be automated. It determines the noise level to find leftmost and rightmost maxima in between them. By using this information it can determine lowest density cuts. Optigrid creates an adaptive grid by partitioning the data in the clusters .Optigrid is more efficient than MAFIA [12] and CLIQUE [3] .

5.6 SUBCLU (density-based SUBspace Clustering)

SUBCLU [9]is a subspace clustering algorithm based on the density notion of DBSCAN. SUBCLU uses an Apriori technique to find clusters in all higher dimensional subspaces. It overcomes the limitations of greedy based approach. Using greedy approach it searches for all subspaces of high dimension space. the core objects used in defining clusters in this approach holds monotonicity property. Compared to CLIQUE [3] and MAFIA [12], SUBCLU achieves a better quality.

5.7 PreDeCon (Subspace Preference Weighted Density Connected Clustering)

PreDeCon [15] is a subspace clustering algorithm based on the idea of DBSCAN. It generates a subspace vector for each data object along each dimension. For each point, the local neighborhood is examined to determine a subspace tendency. PreDeCon checks each object to determine core objects. If core objects are found then PreDeCon start forming a cluster around it and add all points that are reachable to the current cluster. It is efficient when compared to PROCLUS [14].

5.8 INSCY (Indexed Subspace Clusters with In-Process- Removal of Redundancy)

INSCY [21] follows depth-first approach, recursive mining in a region of all clusters in all subspaces, before continuing with the region. In INSCY, as the maximal high dimensional projection is evaluated first, immediate pruning of all its redundant low dimensional projections leads to major efficiency gains. Indexing of potential subspace cluster regions is possible. INSCY merges the in-process redundancy pruning with novel index structure, the SCY-tree, for efficient clustering. The SCY-tree is a compact representation and it reduces the database scans. INSCY is fast and it automatically reduces the output size.

5.9 DISH (Detected Subspace Cluster Hierarchies)

DISH [18] based on density on density based clustering. DISH can detect the clusters in subspaces of different dimensionalities. DISH uncovers complex hierarchies of subspace clusters. It can be able to find clusters of different shape, size and density. The major idea of DISH is be define the subspace distance that assigns small values if two points are in a common low-dimensional subspace cluster and high values if two points are in a common high-dimensional subspace cluster. The clusters with small subspace distances are embedded with in clusters with higher subspace distances. It computes the subspace dimensionality representing the dimensionality of that subspace cluster. Any mining algorithm is used to represent the best subspace of an object. DISH is more effective than PreDeCon [15] and PROCLUS [14].

5.10 DENCLUE (Density Clustering)

DENCLUE [8] is a clustering method that employs grid-partition mechanism for data object handling and retrieving. The major features includes, it allows a compact mathematical description of arbitrarily shaped clusters in high dimensional data sets. It is good at datasets with large amounts of noise. It is faster than DBSCAN algorithm. It models the overall density of the set of points as the sum of influence functions associated with each point. The resulting density function will have local peaks, and these local peaks can be used to define clusters. For each data point, a hill climbing procedure determines the nearest peak associated with that point. The set of points associated with a particular peak becomes a cluster. If the local peak is too low, then these points are considered as noise and discarded. DENCLUE doesn't work well if noise is present.

5.11 DENCOS (Density Conscious Subspace Clustering)

Density Divergence Problem is the major issue in high dimensional data clustering. Density divergence means having different subspace cardinalities for different region densities. To overcome this problem DENCOS uses a novel subspace clustering model. It discovers the clusters using different density thresholds in different subspaces. DENCOS uses a data structure DFT-tree, Density Frequent Pattern-tree to save information of dense units. To mine the dense units it calculates the lower and upper bound of the units and uses divide and conquer method. To find the clusters in different subspace dimensionalities, it adaptively calculates the density thresholds. DENCOS is more efficient than SUBCLU [9].

VI. CONCLUSION

Density-based Clustering algorithms are fundamental technology's for data clustering with many attractive properties and applications. In Density-based method, clusters are made according to the density of the data. Identifying clusters in high dimensional data is a challenging task as the high dimensional data consists of hundreds of attributes. In high dimensional data, clusters are embedded in various subsets finding of dimensions. Density based subspace clustering algorithms treat clusters as the dense regions compared to noise or border regions. The task is to reduce the dimensionality of the data, without losing important information. These algorithms are developed to discover clusters with arbitrary shape, density in spatial databases with noise. Density-based clustering methods works efficiently in case of high-dimensional data when compared to other clustering algorithms.

VII. REFERENCES

[1] L. Kaufman, and P.J. Rousseeuw (1990) Finding groupsindata: An introduction to cluster analysis. John Wiley andSons, New York.

[2] J. Daxin, C. Tang and A. Zhang (2004) Cluster analysis forGene expression data: A survey, IEEE Transaction onKnowledge and Data Engineering, Vol. 16 Issue 11, pp.1370-1386.

[3] R. Agrawal, J. Gehrke, D. Gunopulos and Raghavan (1998)Automatic subspace clustering of high dimensional data for data mining applications, In Proceedings of the SIGMOD, Vol. 27 Issue 2, pp. 94-105.

[4] M. Steinbach, L. Ertöz and V. Kumar, "The challenges of clustering high dimensional data", [online] available :http://www.users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf

[5] J. Friedman (1994) An overview of computational learning and function approximation, In: From Statistics to Neural Networks. Theory and Pattern Recognition Applications. (Cherkassky, Friedman, Wechsler, eds.) Springer-Verlag 1

[6] M. Ester, H.-P. Kriegel, J. Sander and X. Xu (1996) ADensity-based algorithm for discovering clusters in largespatial databases with noise, In Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR., pp. 226-231.

[7] Rahmat Widia Sembering, Jasni Mohammad Zian, " Cluster Evaluation of Density based subspace clustering", Journal of computing, Vol 2, Issue 11, Nov 2010, ISSn 2151-9617.

[8] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise, "Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, pp. 58-

[9] K. Kailing, H.P. Kriegel and P. Kroger (2004) Density connected subspace clustering for high dimensional data, In Proceedings of the 4th SIAM International Conference on Data Mining, Orlando, FL, pp. 46-257.

[10] Sunitha Jahirabadkar and Parag kulkarni ," Clustering of High Dimensional data:Density-based subspace clustering Algorithms", International journal of Computer Applications, vol 63-No.20,February 2013.
2(8), 3441-3446.

[11] P. Lance, E. Haque, and H. Liu (2004) Subspace clustering for high dimensional data: A review, ACM SIGKDD Explorations Newsletter, Vol. 6 Issue1, pp 90–105.

[12] Technical Report CPDC-TR-9906-010 (1999) MAFIA: Efficient and scalable subspace clustering for very large data sets, Goil, S., Nagesh, H. and Choudhary, A., NorthwesternUniversity.

[13] C. Procopiuc, M. Jones, P. K. Agarwal and T. M. Murali, (2002) A montecarlo algorithm for fast projective clustering, In Proceedings of the 2002 ACM SIGMOD International conference on Management of data, pp. 418-427.

[14] C. C. Aggarwal, J. L. Wolf, P. Yu, C. Procopiuc, and J. S. Park (1999) Fast algorithms for projected clustering, In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp.61-72.

[15] C. Bohm, K. Kailing, H. P. Kriegel, and P. Kroger, (2004)Density connected clustering with local subspace preferences, In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM-04), Washington DC,USA, pp. 27-34.

[16] H. P. Kriegel, P. Kroger, M. Renz, and S. Wurst (2005) Ageneric framework for efficient subspace clustering of high dimensional data, In Proceedings of the 5th International Conference on Data Mining (ICDM), Houston, TX, pp. 250-257.

[17] E. Müller, S. Günnemann, I. Assent and T. Seidl (2009)Evaluating clustering in subspace projections of high dimensional data, In Proc. of the Very Large Data Bases Endowment, Volume 2 issue 1, pp. 1270-1281.

[18] E. Achtert, C. Bohm, H. P. Kriegel, P. Kroger, I. Muller and A. Zimek 2007. Detection and visualization of subspace cluster hierarchies. In Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA).

[19] Hans Peter Kriegal, Peer Kroger, Matthias Renz, Sebastian Wurst, " A Generic Framework for Efficient Subspace clustering of High-Dimensional data", in proceedings of 5th IEEE international Conference on Data Mining (ICDM), HOUSTON, TX,2005

[20] I. Assent, R. Krieger, E. Muller, and T. Seidl, (2007) DUSC: Dimensionality Unbiased Subspace Clustering. In Proc. IEEE Intl. Conf. on Data Mining (ICDM 2007), Omaha, Nebraska, pp 409-414.

[21] Amandeep kaur Mann and Navneet kaur, "survey paper on clustering techniques", International Journal of Science, Engineering & Technology Research (IJSETR), vol 2, Issue 4, April 2013.

[22] Y. H. Chu, J W. Huang, K. T. Chuang, D. N. Yang and M.S. Chen. (2010) Density conscious subspace clustering for high dimensional data.IEEE Trans. Knowledge DataEng.