Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms

Sunita Jahirabadkar
Department of Computer Engineering & I.T.,
College of Engineering, Pune, India

Parag Kulkarni
Department of Computer Engineering & I.T.,
College of Engineering, Pune, India

ABSTRACT
Finding clusters in high dimensional data is a challenging task as the high dimensional data comprises hundreds of attributes. Subspace clustering is an evolving methodology which, instead of finding clusters in the entire feature space, it aims at finding clusters in various overlapping or non-overlapping subspaces of the high dimensional dataset. Density based subspace clustering algorithms treat clusters as the dense regions compared to noise or border regions. Many momentous density based subspace clustering algorithms exist in the literature. Each of them is characterized by different characteristics caused by different assumptions, input parameters or by the use of different techniques etc. Hence it is quite unfeasible for the future developers to compare all these algorithms using one common scale. In this paper, we presented a review of various density based subspace clustering algorithms together with a comparative chart focusing on their distinguishing characteristics such as overlapping / non-overlapping, axis parallel / arbitrarily oriented and so on.

General Terms
Data Mining, Machine Learning, Data and Information Systems

Keywords
Density based clustering, High dimensional data, Subspace clustering

1. INTRODUCTION
Data Mining deals with the problem of extracting interesting patterns from the data by paying careful attention to computing, communication and human-computer interaction issues. Clustering is one of the primary data mining tasks. All clustering algorithms aim at segmenting a collection of objects into subsets or clusters, such that objects within one cluster are more closely related to one another than to objects assigned to different clusters [1].

Many applications of clustering are characterized by high dimensional data where each object is described by hundreds or thousands of attributes. Typical examples of high dimensional data can be found in the areas of computer vision applications, pattern recognition, molecular biology [2], CAD (Computer Aided Design) databases and so on. However, high dimensional data clustering initiates different challenges for conventional clustering algorithms. In high dimensional data, clusters may be visible in certain, arbitrarily oriented subspaces of the complete dimension space [3]. Another, major challenge is the so called curse of dimensionality faced by high dimensional data clustering algorithms, essentially means that distance measures become increasingly meaningless as the number of dimensions increases in the data set [4]. Apart from the curse of dimensionality, high dimensional data contains many of the dimensions often irrelevant to clustering or data processing [3]. These irrelevant dimensions confuse clustering algorithms by hiding clusters in noisy data. Then, common approach is to reduce the dimensionality of the data, of course, without losing important information.

Fundamental techniques to eliminate such irrelevant dimensions can be considered as Feature selection or Feature Transformation techniques [5]. Feature transformation methods such as aggregation, dimensionality reduction etc. project the higher dimensional data onto a smaller space while preserving the distance between the original data objects. The commonly used methods are Principal Component Analysis [6, 7], Singular Value Decomposition [8] etc. The major limitation of feature transformation approaches is, they do not actually remove any of the attributes and hence the information from the not-so-useful dimensions is preserved, making the clusters less meaningful.

Feature selection methods remove irrelevant dimensions from the high dimensional data. Few trendy feature selection techniques are discussed in ref. [9, 10, 11, 12]. The problem with these techniques is that they convert many dimensions to a single set of dimensions which later makes it difficult to interpret the results. Also, these approaches are inappropriate if the clusters lie in different subspaces of the dimension space.

1.1 Subspace Clustering
In high dimensional data, clusters are embedded in various subsets of dimensions. To tackle this problem, recent research is focusing on a new clustering paradigm, commonly called as "Subspace Clustering". Subspace clustering algorithms attempt to discover clusters embedded in different subspaces of high dimensional data sets. Formally, a subspace cluster is defined as (Subspace of the feature space, Subset of data points). Or

$$C = (O, S) \text{ where } O \subseteq DB, S \subseteq D$$

Here, C is a subspace cluster, O is a set of objects in given database DB and S is a subspace projection of the dimension space D.

Fig. 1 show the subspace clusters Cluster1 to Cluster5. Cluster4 represents a traditional full dimensional cluster ranged over dimensions d1 to d16. Cluster3 and Cluster5 are non-overlapping subspace clusters appearing in dimensions {d5, d6, d7} and {d13, d14, d15} respectively. Cluster1 and Cluster2 represent overlapping subspace clusters as they share a common object p7 and a common dimension d6.

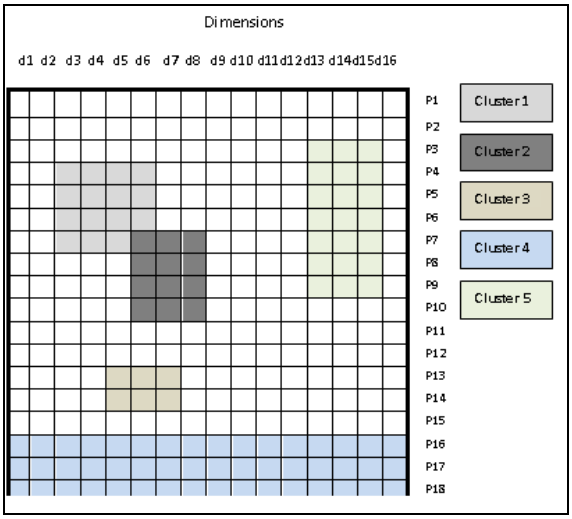

Fig 1 : Overlapping/ non-overlapping subspace clusters

Subspace clustering algorithms face a major challenge. As the search space for relevant subspaces for defining meaningful clusters is in general, infinite; it is necessary to apply some heuristic approach to make this processing of identifying suitable subspaces containing clusters, feasible [13, 8]. This heuristic approach decides the characteristics of the subspace clustering algorithm. Once the subspaces with higher probability of comprising good quality clusters are identified, any clustering algorithm can be applied to find the clusters hidden in those subspaces.

1.2 Density Based Subspace Clustering

Density based clustering algorithms are very popular in the applications of data mining. These approaches use a local cluster criterion and define clusters as the regions in the data space of higher density compared to the regions of noise points or border points. The data points may be distributed arbitrarily in these regions of high density and may contain clusters of arbitrary size and shape. A common way to find the areas of high density is to identify grid cells of higher densities by partitioning each dimension space into non-overlapping partitions or grids. The algorithms following this notion are called as grid based subspace clustering approaches.

The first density based clustering approach is DBSCAN [14]. It is based on the concept of density reachability. A point is called as "core object", if within a given radius (ε), the neighbourhood of this point contains a minimum threshold number (*MinPts*) of objects. A core object is a starting point of a cluster and thus can build a cluster around it. Density based clustering algorithms using the notion of DBSCAN, can find clusters of arbitrary size and shape. Fig. 1 shows clusters built with density notion with threshold ≥ 10. Thus, Fig. 2 (a) builds a cluster with no. of objects > 10 and (b) does not build the cluster.

WaveCluster [15] is another density based clustering approach which uses wavelet transform to the dimension space. The algorithms uses grid to detect clusters of arbitrary shape and size. However, it is applicable to only low dimensional dataset. DenClue [16] is another efficient grid based clustering algorithm, as it keeps information only about dense cells.

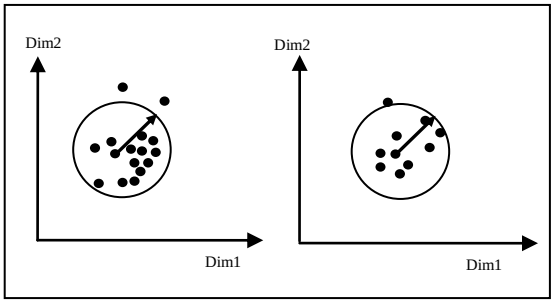

Fig 2 : Clusters by density notion

Density based approaches are commonly and popularly used to discover clusters in high dimensional data. These approaches search for the probable subspaces of high densities and then the clusters hidden in those subspaces. A subspace can be defined as dense subspace, if it contains many objects according to some threshold criteria in a given radius.

CLIQUE [3] is the first grid based subspace clustering approach designed for high dimensional data. It detects subspaces of the highest dimensionalities. SUBCLU [17] is the first subspace clustering extension to DBSCAN to cluster high dimensional data, using the notion of DBSCAN.

While all such clustering approaches organize data objects into groups, each of them uses different methodologies to define clusters. They make different assumptions for input parameters. They define clusters in dissimilar ways as overlapping, non-overlapping, fixed or variable size and shape, and so on. The choice of a search technique, such as top down / bottom up, can also determine the characteristics of the clustering approach. As such, it is quite confusing for developers to select algorithms with which they can compare the results of their proposed algorithm. Hence, we are presenting a short descriptive comparison of few existing, significant density based subspace clustering approaches. It will make easy for the future developers to compare their algorithm with an appropriate set of similar algorithms.

The remainder of this paper is organized as follows. In the next section, we start by reviewing various surveys available in literature on subspace clustering approaches. In Section III, we briefly present key characteristics, functioning, special features, advantages, disadvantages of all significant density based clustering algorithms, followed by a comparison table of all these algorithms focusing on their efficiency, quality etc. in Section IV. Section V discusses the conclusion along with the future research direction.

2. RELATED WORK

There exist few excellent surveys on high dimensional data clustering approaches in literature [2, 18, 19, 20, 21]. In [2], authors have presented a variety of algorithms and challenges for clustering gene expression data. They also discussed different methods of cluster validation and cluster quality assessment. The authors in [19] have presented an extremely comprehensive survey beginning with the illustration of different terminologies used in subspace clustering methodologies. It discusses various assumptions, heuristics or intuitions forming the basis of different high dimensional data clustering approaches. In [20], authors have explored the behaviour of some of the grid based subspace clustering algorithms with their experimental results in 'A survey of Grid Based Clustering Algorithms'. Authors of [18] and [21] presented an excellent survey of subspace clustering approaches along with a classification based on their defining characteristics.

However, there does not exist any explicit comparison among all existing density based subspace clustering algorithms. We present in this paper, such a comparison among clustering algorithms which adopts density based subspace clustering approach for clustering high dimensional data.

3. DENSITY BASED SUBSPACE CLUSTERING ALGORITHMS

In this section, various density based subspace clustering algorithms have been reviewed in brief and they are compared on the basis of their characteristics, such as –

1. Axis parallel / Arbitrarily oriented clusters : The algorithms that aim at finding clusters in axis-parallel subspaces of the data space are called as axis parallel subspace clustering algorithms and the algorithms for finding clusters in arbitrarily oriented subspaces of the feature space are called as arbitrarily oriented clustering, generalized subspace clustering or correlation clustering algorithms [19].

2. Overlapping / non-overlapping clusters : Overlapping clusters allow data points to belong to several clusters in varying subspace projections. Non-overlapping clusters assign each object either to a unique cluster or to a noise.

3. Grid based / Non-grid based : Grid based subspace clustering approaches are based on the grid approximation of data space. These algorithms first partition the range of values in every dimension into equal sized / variable sized cells and then combine the high density adjacent cells to form a cluster (c. f. Fig. 1). The non-grid based approaches or density connected approaches are based on the concept used in DBSCAN [14]. They compute the density around a data object by searching its ε–neighbourhood and if it contains number of objects more than the *MinPts* threshold, it will be called as core object and a cluster will be formed around it. Following algorithms are discussed on the basis of these and many other performance characteristics.

3.1 CLIQUE (CLustering In QUEst)

CLIQUE [3] is the first grid based, non-overlapping, axis parallel subspace clustering algorithm. It uses an apriori-like method which recursively navigates through the set of possible subspaces in a bottom-up way. The data space is first partitioned by an axis-parallel grid into equi-sized blocks. Only units, whose densities exceed a threshold, are retained. The bottom-up approach of finding such dense units starts with 1- dimensional dense units. The recursive step from (k-1) dimensional dense units to k-dimensional dense units takes (k-1) dimensional dense units as candidates and generates the k-dimensional units by self-joining all candidates having the first (k-2) dimensions in common. All generated candidates which are not dense are eliminated. After generating all interesting dense units, clusters are found as a maximal set of connected dense units. The size of the grid and a global density threshold are needed by CLIQUE as input parameters. It detects clusters of highest dimensionality.

3.2 MAFIA (Merging of Adaptive Finite IntervAls)

MAFIA [22] is a more significant modification of CLIQUE. It divides each dimension into variable sized, adaptive grids. A technique based on histograms is used to merge grid cells reducing the number of bins. A so called cluster dominance factor is used as an input parameter to select bins which are more densely populated (relative to their volume) than the average. The algorithm starts to produce such one dimensional dense units as candidates and proceeds recursively to higher dimensions.

MAFIA uses any two k-dimensional dense units to construct a new (k+1)-dimensional candidate as soon as they share an arbitrary (k-1)-face (not only first dimensions). Compared to CLIQUE, the number of candidates generated is much larger. Neighbouring dense units are merged to form clusters and the clusters that are true subsets of higher dimensional clusters, are removed.

3.3 DOC (Density based Optimal projective Clustering)

DOC [23] is a non-grid density based, non-overlapping subspace clustering algorithm. It measures density using hypercubes of fixed width and thus has all those problems as that of grid based approaches. It proposes a mathematical formulation to compute approximations of optimal projective clusters using a Monte Carlo algorithm, concerning the density of points in subspaces. It uses three input parameters as - size of the grid (ω), global density threshold (α) and a balance between data points & dimensions (β). However, DOC only finds approximations of clustering. Also, if a subspace cluster is larger than the defined fixed width hypercube, then few objects may be failed to detect or may contain noise points when the size of the projected cluster is significantly smaller than width of the hypercube.

3.4 PROCLUS (PROjected CLUStering)

PROCLUS [24] is a first, top-down partition based projected clustering algorithm based on the concepts of *k*-medoid clustering. It computes medoids for each cluster iteratively on a sample of data using a greedy hill climbing technique and then improves the results iteratively. The input parameters to be supplied are number of clusters (k) and average number of dimensions (l). Cluster quality in PROCLUS is a function of average distance between data points and the nearest medoid. Also, the subspace dimensionality is an input parameter, which generates clusters of similar sizes. However, PROCLUS is faster than CLIQUE due to the sampling of large datasets, though the use of small number of representative points can cause PROCLUS to miss some clusters entirely.

3.5 SUBCLU (density connected SUBspace CLUstering)

SUBCLU [17] is based on the density notion of DBSCAN which adopts the concept of density connected clusters. It overcomes the limitations of grid-based approaches. It searches for all subspaces of high dimension space using a greedy approach to discover the density connected clusters. It starts with generating all 1-d clusters by applying DBSCAN on every dimension using the input density parameters ε-radius and µ-density threshold. The core object definition used in defining clusters in this approach also holds the monotonicity property. Using this, an Apriori based technique is used to find clusters in all higher dimensional candidate subspaces. Compared to the grid-based approaches like CLIQUE [3], MAFIA [22], DOC [23] etc., SUBCLU achieves a better clustering quality, but requires a higher runtime. Also, it uses global density parameters for subspaces of varying dimensionalities, which degrades the performance by a great extent.

3.6 PreDeCon (subspace PREference weighted DEnsity CONnected clustering)

PreDeCon [25] is another subspace clustering algorithm based on the notion of DBSCAN. It constructs a subspace preference vector for each data object using the variance of the points in the ε-neighbourhood of that object along each dimension. For each data object, PreDeCon checks if it is a preference weighted core object. If it is a core object, then PreDeCon starts building a

cluster around it and adds all points that are preference weighted reachable from p to the current cluster. It gives better performance compared to DOC and PROCLUS, however, it faces problems with subspace clusters of considerably different dimensionalities.

3.7 FIRES (FIlter REfinement Subspace clustering)

FIRES [26] uses an approximate solution for efficient subspace clustering. Rather than going bottom up, it makes use of 1-d histogram information (called base clusters) and jumps directly to interesting subspace regions. Moreover, generation of these base clusters can be done using any clustering approach and may not restrict to DBSCAN [14]. FIRES then generates cluster approximations by combining base clusters to find maximum dimensional subspace clusters. These clusters are not merged in an apriori style. It uses an algorithm that scales at most quadratic with respect to the number of dimensions. It refines these cluster approximations as a post processing step to better structure the subspace clusters. We tested FIRES using OpenSubspace [27] in Weka. Compared to SUBCLU [17], FIRES [26] takes too small execution time and gives more accurate results than SUBCLU.

3.8 DiSH (Detecting Subspace cluster Hierarchies)

DiSH [28] is based on the density based clustering notion of OPTICS [29]. It can find clusters of different size, shape, densities and dimensionality in various subspaces of high dimensional space. The key idea of DiSH is to 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 or are not in a subspace cluster at all. Subspace clusters with small subspace distances are embedded within clusters with higher subspace distances. Using this concept and the variance analysis of each data point, DiSH first computes the subspace dimensionality representing the dimensionality of that subspace cluster, in which object *o* fits best. And then using any frequent item-set mining algorithm (e.g. Apriori algorithm [30]), it determines the best subspace of an object *o*. Compared to OPTICS [29], PreDeCon [25] and PROCLUS [24], DiSH gives better performance and is more effective.

3.9 DENCLUE (DENsity CLUstEring)

DENCLUE [16] is a generalization of DBSCAN and K-means. It works in two stages as pre-processing stage and clustering stage. In pre-processing step, it creates a grid for the data by dividing the minimal bounding hyper-rectangle into d-dimensional hyper-rectangles with edge length 2σ. In the clustering stage, DENCLUE associates an "influence function" with each data point and the overall density of the dataset is modelled as the sum of influence functions associated with each point. The resulting general density function will have local peaks, i.e., local density maxima, and these local peaks can be used to define clusters. If two local peaks can be connected to each other through a set of data points, and the density of these connecting points is also greater than a minimum density threshold ξ, then the clusters associated with these peaks are merged forming the clusters of arbitrary shape and size. The performance of DENCLUE is appealing in low dimensional space, however, it does not work well as the dimensionality increase or if noise is present.

3.10 OptiGrid (OPTImal GRID clustering)

The researchers who created DENCLUE also created OptiGrid [31]. First, it generates a histogram of data values for each dimension. Then it determines the noise level to find leftmost and rightmost maxima in between them. It uses this information to locate lowest density cuts and using these various cut points; OptiGrid creates an adaptive grid, partitioning the data in each dimension. The highly populated grid cells then decide the clusters. OptiGrid looks a lot like MAFIA [22] as it creates an adaptive grid using a data dependent partitioning. However, it does not face the problem of combinatorial search, like MAFIA and CLIQUE, to find the best subspace. It simply searches for the best cutting planes and creates a grid that is not likely to cut any clusters. From an efficiency point, OptiGrid is much better than MAFIA and DENCLUE [16].

3.11 DUSC (Dimensionality Unbiased Subspace Clustering)

To overcome the effect of varying dimensionality of subspaces, DUSC [32] gives a formal definition of dimensionality bias, based on statistical foundations. DUSC assigns weights to each object contained in ε – neighbourhood of each object. Thus an object *o* in subapce S is called dense if the weighted distances to objects in its area of influence sum up to be more than a given density threshold τ. It further uses Epanechnikov kernel estimator [32] to estimate density value at any position in the data space. It assigns decreasing weights to objects with increasing distance. Density of any object is then measured with respect to the expected density α (S). Thus, an object *o* is dense in subspace S according to the expected density α (S), if and only if, $\frac{1}{\alpha(s)}\varphi^{s}(o) \geq F$ where F denotes the density threshold. F is independent of the dimensionality and data set size, and is much easier to specify than traditional density thresholds. DUSC also combines the major paradigms for improving the runtime. The experiments on large high dimensional synthetic and real world data sets show that DUSC outperforms other density connected subspace clustering algorithms such as SUBCLU, PreDeCon etc. in terms of accuracy and runtime.

3.12 INSCY (INdexing Subspace Clusters with in-process-removal of redundancY)

INSCY [33] is an efficient subspace clustering algorithm based on the subspace clustering notion of DUSC [32]. It uses a depth first approach to mine recursively in a region of all clusters in all subspace projections and then continue with the next region. Because of this, it evaluates the maximal high dimensional projection first, quickly pruning all its redundant low dimensional projections. This approach leads to major efficiency gains as it overcomes the drawbacks of breadth first subspace clustering reducing runtimes substantially. Also, this allows indexing of promising subspace cluster regions. INSCY proposes a novel index structure SCY-tree, which provides a compact representation of the data allowing arbitrary access to subspaces. SCY-tree combines in-process redundancy pruning, for very efficient subspace clustering. This makes INSCY fast and concise. Thorough experiments on real and synthetic data show that INSCY yields substantial efficiency and quality improvements over the most recent non-approxive density based subspace clustering algorithms SUBCLU.

3.13 DENCOS (DENsity COnscious Subspace clustering)

A critical problem faced by high dimensional data clustering is "density divergence problem" i.e. different subspace cardinalities have different region densities. DENCOS [34] addresses this problem by formulating a novel subspace clustering model which discovers the clusters using different density thresholds in different subspace cardinalities. It uses a novel data structure DFP-tree (Density Frequent Pattern-tree) to save the complete information of all dense units. To accelerate the search of dense units, using DFP-tree, it calculates the lower and upper bound of the unit counts and uses divide and conquer method to mine dense units. DENCOS uses δ-eq. length intervals, α-unit strength factor and k_{max}-max.sub cardinality as three input parameters; using which it adaptively calculates different density thresholds to locate the clusters in different subspace dimensionalities. DENCOS performs better than the comparative density based subspace clustering algorithm SUBCLU.

3.14 Scalable Density Based Subspace Clustering

Scalable density based subspace clustering [35] is a method that steers mining to few selected subspace clusters only. It reduces subspace processing by identifying and clustering promising subspaces and their combinations directly, narrowing down the search space while maintaining accuracy. It uses the principle that any high dimensional subspace cluster appears in many low dimensional projections. By mining only some of them, the algorithm gathers enough information to jump directly to the more interesting high dimensional subspace clusters without processing the in between subspaces. Database scans are completely avoided with this approach for many intermediate, redundant subspace projections, steering the process of subspace clustering.

Table 1. Characteristics of various density based subspace clustering algorithms

Clustering Approach	Axis parallel	Grid based	Adaptive grid	Overlapping	Search Strategy		Use of Monotonicity property	Input Parameters	global density parameters	Shape of the Cluster	Robust to noise	Independent of order of data	Independent of order of dimensions	Arbitrary Subspace Dimensionality
					Bottom -up	Top-down								
CLIQUE	Y	Y	-	Y	Y	-	Y	ξ-grid interval τ-threshold	Y	F	N	Y	Y	Y
MAFIA	Y	Y	Y	Y	Y	-	Y	α-clus dominance factor	Y	F	Y	Y	Y	Y
DOC	Y	N	-	N	-	Y	N	ω-size of grid α-threshold β-bal bet'n pts & dims	Y	F	Y	Y	Y	Y
PROCLUS	Y	N	-	N	-	Y	-	k-no. of cluster l-avg no. of dims	N	F	Y	Y	Y	N
SUBCLU	Y	N	-	Y	Y	-	Y	ε-radius μ-threshold	Y	A	Y	Y	Y	Y
PreDeCon	Y	N	-	N		Y	Y	ε-radius μ-threshold λ, δ-preference para.s	Y	A	Y	Y	Y	N
FIRES	Y	N	-	Y	-	-	N	ξ-threshold σ-radius	N	F	Y	Y	Y	Y
DiSH	Y	N	-	Y	Y	-	Y	ξ-threshold σ-radius	Y	A	Y	Y	Y	Y
DENCLU	Y	Y	N	Y	Y	-	-	ξ-threshold σ-radius	-	A	N	Y	Y	N
OptiGrid	Y	Y	Y	Y	Y	-	-	ξ-threshold σ-radius	-	A	Y	Y	Y	Y
DUSC	Y	N	-	Y	Y	-	N	μ-threshold	N	A	Y	Y	Y	Y
INSCY	Y	N	-	Y	Y	-	N	τ –threshold ε-radius R-redundancy factor	N	A	Y	Y	Y	Y
DENCOS	Y	N	-	Y	Y	-	N	δ-eq. length intervals α-unit strength factor max subsp cardinality	N	A	Y	Y	Y	Y
Scalable Density Based Sub. Clustering	Y	N	-	Y	Y	-	N	P-jump indicator	N	A	Y	Y	Y	Y

It uses priority queue to initialize the information of density estimates. It gives a basis for selecting the best candidate from the priority queue. The priority queue is split into three levels for multiple density granularities. It skips intermediate subspaces in a best first manner and jumps directly to high dimensional subspaces. The experiments prove that the best first selection of subspace clusters enables a scalable subspace clustering algorithm with enhanced run time and it also produces high quality subspace clustering.

4. COMPARISON AMONG DENSITY BASED SUBSPACE CLUSTERING ALGORITHMS

In this paper, we discussed high dimensional data clustering challenges and also discussed existing, significant density based subspace clustering approaches to solve this problem. We used an open source framework OpenSubspace [27] for evaluation and exploration of few of these algorithms. We tested CLIQUE, DOC, FIRES, INSCY, PROCLUS and SUBCLU using OpenSubspace. We used the Diabetes synthetic data available and supported by OpenSubspace, which consists of 8 dimensions and 779 instances.

A comparative chart (Table 1) is prepared and presented, using different parameters such as run time, shape and size of the clusters, input parameters, use of grid, use of global density parameters, search strategy used, can handle noise and so on to compare all these approaches.

The table uses abbreviations for shape of the cluster as : 'F' for fixed shape and 'A' for arbitrary shape.

5. CONCLUSION

Subspace clustering algorithms help solve the problems of clustering in high dimensional data by using different techniques to locate clusters in different subsets of the complete dimension set. Density based clustering algorithms perform better, compared to other subspace clustering approaches, by generating clusters of adaptive size, shape, densities and dimensionalities.

Lot of such approaches exist for subspace clustering and numerous algorithms are being proposed nearly every day. Proper selection of a clustering approach to suit a particular application and data should be based on the understanding of the exact requirement of clustering application and the principles of working of available approaches.

Hence, in this paper an attempt is made to present various density based subspace clustering algorithms to better understand their comparative characteristics. A comparative chart is prepared on the basis of various performance parameters and presented for a ready reference. We hope, this will surely help future developers to select a set of relevant / appropriate approaches from the given list, against which developers can test / compare the results of their proposed subspace clustering algorithm. Finally, we limited the scope of this paper only to few, significant representative contributions and that too clustering based on continuous valued data. There exist many clustering algorithms which are specially designed for stream data, graph data, spatial data, text data, heterogeneous data etc. We hope to stimulate further research in these areas.

6. ACKNOWLEDGEMENT

We would like to thank my student, Shweta Daptari, for providing help in implementing and testing our algorithm.

7. REFERENCES

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

[2] J. Daxin, C. Tang and A. Zhang (2004) Cluster analysis for Gene expression data: A survey, IEEE Transaction on Knowledge 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. Gao, P. W. Kwan and Y. Guo (2009) Robust multivariate L1 principal component analysis and dimensionality reduction, Neurocomputing, Vol. 72: 1242-1249.

[6] A. Jain and R. Dubes (1988) Algorithms for clustering data, Prentice Hall, Englewood Cliffs, NJ.

[7] K. Fukunaga, (1990) Introduction to statistical pattern recognition, Academic Press, New York.

[8] G. Strang (1986) Linear algebra and its applications. Harcourt Brace Jovanovich, third edition.

[9] A. Blum and P. Langley (1997) Selection of relevant features and examples in machine learning, Artificial Intelligence, Vol. 97:245–271.

[10] H. Liu and H. Motoda (1998), Feature selection for knowledge discovery & data mining, Boston: Kluwer Academic Publishers.

[11] J. M. Pena, J. A. Lozano, P. Larranaga and Inza, I. (2001) Dimensionality reduction in unsupervised learning of conditional gaussian networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23(6):590 - 603.

[12] L. Yu and H. Liu, (2003), Feature selection for high dimensional data: A fast correlation based filter solution, In Proceedings of the Twentieth International Conference on Machine Learning, pp. 856-863.

[13] 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

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

[15] G. Sheikholeslami, S. Chatterjee and A. Zhang "Wavecluster: A multi-resolution clustering approach for very large spatial databases," In Proceedings of the 24th VLDB Conference (1998).

[16] 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. 58-65 (1998).

[17] 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.

[18] A. Patrikainen, and M. Meila (2006) Comparing subspace clusterings, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, Issue 7, pp. 902-916.

[19] H. P. Kriegel, P. Kroger and A. Zimek, (2009) Clustering high-dimensional data : A survey on subspace clustering, Pattern-Based Clustering, and Correlation Clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 3, Issue 1, Article 1.

[20] M. R. Ilango and V. Mohan, (2010) A survey of grid based clustering algorithms, International Journal of Engineering Science and Technology, Vol. 2(8), 3441-3446.

[21] P. Lance, E. Haque, and H. Liu (2004) Subspace clustering for high dimensional data: A review, ACM SIGKDD Explorations Newsletter, Vol. 6 Issue 1, pp 90–105.

[22] 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., Northwestern University.

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

[24] 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.

[25] 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.

[26] H. P. Kriegel, P. Kroger, M. Renz, and S. Wurst (2005) A generic 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.

[27] 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.

[28] 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).

[29] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander 1999. OPTICS: Ordering points to identify the clustering structure. In Proceedings of the ACM International Conference on Management of Data (SIGMOD).

[30] R. Agrawal and R. Srikant, (1994) Fast algorithms for mining association rules. In: Proc. SIGMOD

[31] A. Hinneburg and D. A. Keim, "Optimal grid clustering: Towards breaking the curse of dimensionality in high dimensional clustering," In Proceedings of 25th International Conference on Very Large Data Bases (VLDB-1999), pp. 506-517, Edinburgh, Scotland, September, 1999, Morgan Kaufmann (1999).

[32] 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.

[33] I. Assent, R. Krieger, E. Müller, and T. Seidl (2008) INSCY: Indexing subspace clusters with in process removal of redundancy", Eighth IEEE International Conference on Data Mining In ICDM, pp. 414–425

[34] 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 Data Eng. 22: 16-30.

[35] Muller, E., Assesnt, I., Gunnemann, S. and Seidl, T. (2011) Scalable Density based Subspace Clustering. Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM'11), pp: 1076-1086.