A Comprehensive Study of High Dimensional Data Clustering and Data Reduction Approaches

S.S.Nikam[1], R.M.Raundal[2]

[1]*Assistant Professor, Computer Science and Information Technology Dept, K.K.Wagh College of Agricultural Engineering and Technology, Nashik (MS), India*

[2] *Assistant Professor, Agril.Statistics Dept, K.K.Wagh College of Agriculture, Nashik (MS), India*

***Abstract*—**High dimensional data clustering is the analysis of data with few numbers to large numbers of dimensions. Large dimensions are not easy to handle and in certain cases difficult to visualize. To improve the efficiency and accuracy of clustering on high dimensions, data reduction is required as pre-processing. Some of the applications of dimension reduction are filtering, compression, regression, classification, feature analysis, and visualization. This paper provides effectiveness of various high dimensional data analysis methods, analysis of the popular existing clustering techniques and impact of dimensionality reduction on different algorithm in the prediction process of Data mining.

***Keywords*—** *Data Mining, Clustering, High Dimensional data, Clustering Algorithm, Dimensionality Reduction*

I. INTRODUCTION

Cluster analysis is a task of grouping a set of objects such a way that objects in the one group are more similar to each other than to those in other groups. It is the task of splitting objects of a data set into different groups such that two objects from one cluster are similar to each other, whereas two objects from distinct clusters are dissimilar[1].Clustering is unsupervised learning technique which involves the grouping of data points. Clustering algorithm classifies each data point into a different group. In clustering, data points that are in the same group should have same or similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a technique for statistical data analysis used in many fields. Many applications use clustering to find useful patterns from the data which helps them in prediction or decision making This is helpful to draw certain valid conclusions and proceed further in that direction for development of application.

High Dimensional data containing multiple dimensions are difficult to visualize and for analysis due to the exponential growth of the number of possible values with each dimensions and becomes intractable with increasing values of each dimension. This problem is well known as the curse of dimensionality. A cluster is intended to group similar or related objects based on observations of their attribute's values. However for a large number of attributes some of the attributes will usually not be meaningful for a given cluster. This is called feature selection. Feature selection techniques are mostly used in domains where there are many features and relatively few number of samples. This paper presents an overview of clustering techniques, their association, advantages and disadvantages of them. Section II describes literature survey of clustering approaches; section III of paper discusses dimension reduction approaches with their features and results for high dimensional data.

Clustering can be considered as unsupervised learning problem which aims at finding a hidden pattern in a collection of unlabelled data. A cluster is a group of elements which are "similar" between them and are "dissimilar" to the objects belonging to other group of clusters [2]. In some pattern recognition problems, the training data consists of a set of unlabeled input x without any corresponding target values. The purpose in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering, or to determine how the data is scattered in the plane, known as density estimation. So for a n-sampled space x1 to xn, true class labels are not provided for each sample.

Distance Measure: In Most of the clustering techniques relay on distance measure as an important step for selecting data objects, which will determine the similarity between two elements in the context of a particular problem [2]. The cluster shape or density will be influenced by the similarity between the data objects [3], as some elements may be close to one another according to one distance and farther away according to another. In general there are two types of distance measures. 1) Symmetric measure and 2) Asymmetric measure. The common distance measures used in the clustering process [3][4] are i)The Euclidean distance or Squared Euclidean distance, ii)The Manhattan distance,iii)The Maximum Likelihood Distance,iv)The Mahalanobis distance,(v)The Hamming distances ,(vi)The angle amongst two vectors used as a distance measure when clustering high dimensional data.

II. LITERATURE SURVEY

A. CLUSTERING APPROACHES

Clustering problems arise in many fields particularly in computer vision, pattern recognition, data mining and machine learning. The clustering problem (Jain et al., 1999) is the problem of dividing a given set {x1, . . . , xN} of N data points

into several homogenous similar groups. Each such group or cluster contains similar data items and data items from different groups should not be similar. We refer to a clustering in k groups as a k-clustering. Clustering techniques can be useful in explorative data analysis, e.g. a sales-company might identify different types of customers based on a clustering of data about the purchases that customers made. Clustering can also be used as a preprocessing step for other tasks. For example, in data visualization the data of widely separated clusters may be visualized in a separate displays (Bishop and Tipping, 1998).Many different approaches to the clustering problem have been developed. Some operate on data represented by their coordinates in a feature space and some other operates on a matrix of pairwise similarities between data points. To overview different types of methods, we can categorize them in three groups.

1) Hierarchical clustering methods

These produce a hierarchy of clusters for the data. The first level of the hierarchy contains all data and at each subsequent level of the hierarchy, one of the clusters of the previous level is split in two. The last level contains all data in individual clusters. The hierarchy is based on pair wise similarities between data points and can be constructed either top-down or bottom-up.

A hierarchy of clusters can be represented as a tree; the root node contains all data and the two children of each node contain disjoint subsets of the data contained in the parent. The leaves of the tree contain the individual data points. A hierarchy of clusters, rather than a 'flat' clustering in k clusters, is desired in some applications. For example, consider hierarchical clustering of newspaper articles: in the top-levels general topics are found, such as politics, financial news and sports. At lower levels the sports cluster might be further subdivided into articles on individual sports. Using Such a hierarchy of clusters enables a user to quickly find articles of interest. In this method at each level of the tree the user can discard large clusters of uninteresting articles and explore further only the more promising clusters. See e.g. (Zhao and Karypis, 2002; Blei et al.,2004) for work on hierarchical clustering of documents.

Hierarchical clustering generally consist of two strategies,

Agglomerative: This is a "bottom up" approach and each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach and all observations in hierarchy start in one cluster, and splits are performed recursively as one move down the hierarchy.

2) Partitional clustering methods

These produce a single clustering with a fixed or specified number of clusters. Most partitional clustering algorithms do not operate on the basis of pairwise similarities but with data represented in some feature space. Typically, these methods start with an initial k-clustering and apply an iterative algorithm to improve upon the initial clustering according to some criterion. Most partitional clustering methods make, sometimes implicitly, assumptions on the distribution of data within each cluster.

Partitional methods cluster the data in a specified number of groups. Their main attraction over the hierarchical methods is that partitional algorithms are generally much more efficient. The main drawback is that assumptions on the shape of the clusters have to be made in advance. Also a desired number of clusters have to be specified, which may be known in some applications but not in others. Some work has been done on estimating the number of clusters from the data, see e.g. (Pelleg and Moore, 2000; Rasmussen, 2000; Fred and Jain, 2002). In general however this issue remains unresolved.

The k-means algorithm (Gersho and Gray, 1992) is one of the most frequently applied partitional clustering algorithms. It is also known by different names such as Generalized Lloyd algorithm, Lloyd Max algorithm, Forgy algorithm, or Linde-Buzo-Gray algorithm.

3) Spectral clustering methods

These operate on a matrix with pairwise similarities between the data points. The optimal clustering is defined as the clustering that minimizes the 'normalized cut' criterion that depends on the clusters size and the total sum of the similarities between points that are assigned to various clusters. Finding the clustering that minimizes the normalized cut is an NP-complete problem. However, a relaxation of this optimization problem can be efficiently solved, and the solution is given by an eigenvector of the normalized similarity matrix. The solution of the relaxed problem is then further processed to find an approximate solution for the original problem. The term 'spectral clustering' refers to the eigenspectrum of the normalized similarity matrix which can be used to assess the number of clusters in the data. Spectral methods are used both to find hierarchical clusterings and k-clusterings for a given k. We treat them separately since their working is quite different from the other approaches.

A relatively recent approach based on pairwise similarities is spectral clustering (Scott and Longuet-Higgins, 1990; Weiss, 1999), which draws on results of spectral graph theory (Chung, 1997).Spectral methods are attractive because they (i) make less severe assumptions on the shape of the clusters than partitional algorithms and (ii) can be very fast, depending on the sparsity of the similarity matrix. Furthermore, implementation of most spectral clustering algorithms is quite easy since the main component is a procedure to find a few eigenvectors of a matrix: this is a well studied problem for which highly optimized implementations are available.

B. Comparison of Clustering Methods

All three above mentioned approaches have their role of applications where a particular approach is preferred. Below we compare these approaches in terms of their scalability and the assumptions underlying them.

Scalability: The hierarchical methods are in general computationally quite demanding, for N data points the agglomerative approach takes a time that is either $O(N^2 \log N)$ or $O(N^3)$, depending on whether or not after each merge all

distances between all clusters have to be computed. In the standard algorithm for hierarchical agglomerative clustering (HAC) has a time complexity of $O(n^3)$ and requires $O(n^2)$ memory. Divisive clustering has a time complexity of $O(2^n)$.Most of the partitional clustering algorithms take computation time of O(Nk) for k clusters. This is the case for both the k-means algorithm and the EM algorithm for probabilistic mixture models: In the assignment step each combination of data point and cluster has to be considered to find the optimal assignments.

Spectral clustering approaches need the similarity for each of the N^2 pairs of points and thus take in principle at least N^2 time to compute the similarity matrix. Depending on the similarity measure, speed-ups might be possible. For example, one could use a matrix in which an entry (i, j) is non-zero only if x_i is among the q-nearest neighbors of x_j or vice-versa. Efficient techniques exist to find nearest neighbors among N points in time O(N log N) for fixed q (Bentley, 1980; Karger and Ruhl, 2002). Often the similarity matrix used in spectral clustering is sparse, i.e. each point has non-zero similarity to only a few others. For such sparse matrices the eigenvectors can be efficiently computed using the power-method (Horn and Johnson, 1985). The iterations of the power method take an amount of time proportional to the number of non-zero entries in the similarity matrix rather than N^2 for a dense matrix.

Assumptions on clusters: For all clustering methods the distance or similarity measure that is used plays a crucial role. The measure impacts the clusters that will be found and also determines the amount of computation needed for each distance calculation. The number of distance calculations that have to be performed depends on the clustering algorithm that is used. Probabilistic mixture models have the advantage that assumptions on the cluster shape are made explicit by assuming the distribution of data within a cluster to be in a parametric class of distributions. The assumptions in other clustering methods are often less explicit. Although the assumptions in mixture models are clear, they are often incorrect. The clusters are readily recovered with spectral or single-link agglomerative clustering. Of course, for the latter methods one needs to determine a suitable similarity measure; some interesting work has been done (Bach and Jordan, 2004) on learning the similarity measure on the basis of several example clusterings. Interestingly, the set of densities that can be implemented using standard component classes can be increased by mapping the data to a space of much larger dimensionality where the new dimensions are (non-linear) functions of the original variables (Wang et al., 2003).

III. DIMENSION REDUCTION APPROACHES

Dimension Reduction in general is the process of converting a set of data having high dimensions into data with lesser/lower dimensions ensuring that it conveys similar information concisely without affecting original information. These techniques are in general used while solving machine learning problems to obtain better features for a classification or regression related task. It helps in data compressing and reducing the storage space required. It removes redundant features that improves model performance.

There are many methods to perform Dimension reduction in literature. Some of the techniques with their features and other performance measure are,

Author & Year	Technique	Feature	Result
Gahar, Rania Mkhinini & 2017	Missing Values	Algorithm is proposed for High-Dimensionality Reduction for Heterogeneous Data to deal with curse of dimensionality	Algorithm is based on both PCA and MCA in which PCA method enables the processing of quantitative variables while MCA method enables the processing of categorical variables
El Moudden I, Ouzir M, Benyacoub B, ElBernoussi S & 2016	Decision Tree	Automatically detect and analyze human activities from the information acquired from different sensors	Classification accuracy was maximized with minimum number of feature and technique well suited for pattern recognition problems.
Hur, Jae-Hee, Sun-Young Ihm, and Young-Ho Park & 2017	Random Forest	Clarify which variable affects classification accuracy to study variable impact	Prediction based on priority of the variables was obtained to study the classification result
Hui, Kar Hoou & 2017	Forward Feature Selection and Backward Feature Elimination	Most representative Subset of variables from set of variables was selected from multidimensional data	Increased accuracy rate and reduced run-time when searching data consisting of a large number of multidimension al data
Ali, M. Usman & 2017	PCA and Factor Analysis	Bioinformatics data having hundreds of attributes	Number of Attributes was reduced in proposed study.PCA was applied on the data set and 9 components were selected out of the 500 components and then Factor Analysis was used to extract the important features

IV. CONCLUSION

The purpose of this paper is to present a broad classification of different clustering techniques for high dimensional data. Clustering high dimensional data sets is a complex task for some problems. As there is tremendous growth in the fields of communication and technology, there is tremendous growth in high dimensional data spaces [10]. This study focuses on issues and major limitations of existing algorithms. As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality which results in de-grading the quality of the results. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless and complex. This problem has been studied broadly and there are various solutions and each solution is appropriate for different types of high dimensional data problems and data mining procedures [11]. There are many potential applications like compression, regression, visual analysis, bioinformatics, text mining with high dimensional data where techniques such as subspace clustering, projected clustering approaches could help to reveal patterns missed by current clustering approaches or to discover hidden patterns in data. As with any clustering techniques, finding meaningful and useful results depends on the selection of the appropriate clustering technique [12]. In order to do this, one must understand the dataset in a domain specific context for any specific problem in order to be able to best evaluate the results from various approaches. From the above discussion it is observed that the current techniques will suffers with many problems [1].To improve the performance and time complexity for some techniques of the data clustering in high dimensional data, it is necessary to perform research in the areas like dimensionality reduction, redundancy reduction in clusters and data labelling [13].Also Dimensionality reduction can be used to eliminate irrelevant and redundant features from the datasets. It can be categorized into two sub-categories i.e. feature extraction and feature selection [20]. The feature extraction approach uses multiple features to compose a new feature with lower dimensional feature space. Feature extraction methods are Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA) [20].Feature selection approach can be use in some problems to selects a subset of features from the dataset and aims to minimize feature redundancy.

REFERENCES

[1] P. Berkhin, "A Survey of Clustering Data Mining Techniques" Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press, 25-72, 2011.

[2] Guha S., Rastogi R., Shim K,"CURE: An efficient clustering algorithm for large databases", Proc. Of ACM SIGMOD Conference, 2012.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2010.

[4] A.K.Jain and R.C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 2009.

[5] Rui Xu and W. Donald, "Survey of Clustering Algorithms," IEEE Transaction on Neural Network, vol. 16, 2009.

[6] Gan Guojan, Ma Chaoqun, and W. Jianhong," Data Clustering: Theory, Algorithm and Applications",Philadelphia, 2012.

[7] A.Jain and R. Dubes, "Algorithms for Clustering Data", New Jersey, 2011.

[8] A. K. Jain, M. N. Murtyand, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys vol.31, pp. 264-324, 2012.

[9] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya A, Foufou S, Bouras A. A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans Emerg Topics Comp. 2014;2(3):267–79.

[10] K. Bache and M. Lichman. (2013). UCI MachineLearning Repository. Available: http://archive.ics.uci.edu/ml/machinelearningdatabases/.

[11] M. Steinbach, L. Ertoz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, Ed New Vistas: Springer, 2010.

[12] A. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Volume 31(3), pp. 264-323, 2011.

[13] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2010.

[14] https://en.wikipedia.org/wiki/Clustering_high-dimensional_data

[15] Gahar, Rania Mkhinini, et al. "Dimensionality reduction with missing values imputation." arXiv preprint arXiv:1707.00351