I.J. Modern Education and Computer Science, **2015, 1, 38-46**
Published Online January 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.01.06

Clustering Techniques in Bioinformatics

Muhammad Ali Masood, M. N. A. Khan
Shaheed Zulfikar Ali Bhutto Institute of Science and Technologies, Islamabad, Pakistan.
Email: muhammad.ali.masood@gmail.com, mnak2010@gmail.com

Abstract—Dealing with data means to group information into a set of categories either in order to learn new artifacts or understand new domains. For this purpose researchers have always looked for the hidden patterns in data that can be defined and compared with other known notions based on the similarity or dissimilarity of their attributes according to well-defined rules. Data mining, having the tools of data classification and data clustering, is one of the most powerful techniques to deal with data in such a manner that it can help researchers identify the required information. As a step forward to address this challenge, experts have utilized clustering techniques as a mean of exploring hidden structure and patterns in underlying data. Improved stability, robustness and accuracy of unsupervised data classification in many fields including pattern recognition, machine learning, information retrieval, image analysis and bioinformatics, clustering has proven itself as a reliable tool. To identify the clusters in datasets algorithm are utilized to partition data set into several groups based on the similarity within a group. There is no specific clustering algorithm, but various algorithms are utilized based on domain of data that constitutes a cluster and the level of efficiency required. Clustering techniques are categorized based upon different approaches. This paper is a survey of few clustering techniques out of many in data mining. For the purpose five of the most common clustering techniques out of many have been discussed. The clustering techniques which have been surveyed are: K-medoids, K-means, Fuzzy C-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Self-Organizing Map (SOM) clustering.

Index Terms—Clustering Techniques, Data Mining, DBSCAN, Hierarchical Clustering, Performance Analysis.

I. Introduction

The field of data mining is used to extract useful information, identify the concealed patterns and identical attributes within big body of dataset. Data mining provides a powerful support for decision-making through the application of supervised and unsupervised data analysis techniques.

Data mining tasks utilize different techniques such as clustering, prediction, association, classification, sequential patterns and decision tree. These data mining techniques are briefly explained as under:

A. Association

Association is also known as relation technique as it is based on a relationship between items in the same operation a pattern is discovered. Most common example of this technique is market basket analysis to recognize the purchasing trends of consumers associated with different products.

B. Classification

Classification is a typical data mining technique which is used to classify predefined set of classes based on each item in a dataset. Mathematical techniques like neural network, statistics, decision trees and linear programming are used to perform classification.

C. Prediction

Prediction is a data mining techniques that discovers relationship and dependencies of different attributes. In this technique independent variables relationships and dependent variables relationship are discovered. Based on the historical data, fitted regression curve can be drawn for future prediction.

D. Sequential Patterns

Sequential patterns analysis technique seeks to explore or identify similar patterns, consistent events or trends in transaction data over ascertain timeframe.

E. Decision trees

Decision tree is one of the mostly used data mining techniques because of its ease to understand and use. The root of the decision tree is a condition that has different answers and each answer leads to a set of conditions to help process the data so that final decision can be made.

F. Clustering

Clustering is a data mining technique that automatically creates suitable cluster of objects which have similar characteristics. Clustering technique is unsupervised as compared to classification technique in which objects are assigned into predefined classes. Clustering defines the classes and places objects in each class based on similar properties.

Data mining systems are either supervised or unsupervised, depending on whether the domain is already known or not. If domain is known then separate supervised classes are defined for making it supervised classification, or if domain is unknown then unsupervised

clustering is performed where exploratory data analysis is done to identify the hidden data patterns.

Clustering technique is an unsupervised data mining technique which is used to place individual artifacts into relevant groups without prior knowledge of distinct group properties to explore structure in the data. Clusters automatically link hidden patterns by learning the data pattern which is then utilized for learning. The aim of clustering is to make unlabeled dataset into isolated set of data structures by means of learning hidden data concept. For example, the spending behaviors of different population segments can be compared to find out which segments to target for a new product release.

Clustering is an initial and fundamental step in data analysis. Historically, clustering has its foundations laid down by mathematics, statistics and numerical analysis making it a classification of patterns, in unsupervised manner, into groups of similar objects. So patterns in a cluster are more alike to each other than to a pattern related to other cluster. It identifies groups of related records that can be used as an opening point for exploring further associations. Clustering can be classified into the five major types based on criteria like: Hierarchy, Density, Partition, Grid and Model

One of the biggest challenges in clustering is to decide which algorithm is to be used for a specific problem. Algorithms differ in their execution characteristics, creating discrete cluster analysis models. Understanding these analytical models is very important in identifying the variances between the outputs of various algorithms. These clustering models include Connectivity models, Centroid models, Density models, Subspace models, Graph-based models, Group models and Distribution models.

Cluster formation is one of the most difficult techniques used for knowledge extraction process. The goal is to identify clusters without any prior knowledge to differentiate the attributes of different clusters. Clustering techniques are used for correlating identified artifacts into groups based on the following criteria:

- Each cluster is homogeneous in nature.
- Each cluster should be diverse in nature from other clusters.

The usefulness of clustering lies in various arenas e.g. Geo-informatics, web mining, Bio-informatics, market research, market segmentation, Image processing, Document categorization, learning and pattern recognition.

The unsupervised characteristics of the task require that its structural properties are unknown making dimensional distribution of the data in terms of the number, volume, density, shape, and orientation unknown. When applied to data mining applications; clustering encounters three additional complications, including huge data repositories, objects having different characteristics and numerous attributes types.

By default clustering poses different problems for which each solution might be violating at least one rule regarding scale invariance, richness, and cluster consistency. All these properties and rules are defined to enhance the credibility of clustering techniques as if we do not have equal variance then it will be impossible to avoid clusters that are dominated by variables having most variation. Same is the case with cluster consistency and richness; if there is lack of consistency between data partitions then it will be again a serious threat to the credibility of clusters formed.

Based on different assumptions, clustering techniques uses certain data model, and there are chances that due to misguided assumptions, we might have chosen wrong model to apply on sample data causing erroneous or unrelated results. So, it is important that domain knowledge of data is available for successful clustering and there are chances that even domain experts might not be able to provide such crucial information. To establish strong grounds for the sample data's distribution or processing tin to the proper number of clusters we need to identify relevant subspaces or visualization of domain knowledge. Hence efficient and effective methods are required to strengthen the individual clustering algorithms due to exploratory nature of clustering tasks.

II. Literature Review

Jelili et al. [1] implemented k-mean clustering analysis technique to examinestudents'academic performance data. The k-means clustering technique is used in combination with Euclidean distance, a deterministic statistical analysis technique, to analyze the students' performance. Main aim of the paper is to present predictive power of clustering algorithms and statistical techniques. A futuristic approach for data analysis as used for 79 student's results for nine courses offered to each student. The trends were presented as distance. A qualitative data analysis approach was used tomeasure the similarity distances and produce the numerical explanation of the results for the performance assessment. Usually time complexity is dependent on speed and type of the system.

The technique proposed is not only a model for academic forecasts but is an improved version of the existing models by removing their limitations. The existing methods described in this paper are fuzzy models which uses the dataset of only two course results to predict students' academic behaviors. Another approach described is rough Set theory to analyze student data using Rosetta toolkit. The purpose of using this toolkit is to assess data in relation to identifying association between the affecting factors and student grade.

Reference [2] utilizes input data points based on arbitrary distribution to analyze the performance and quality of two clustering algorithms i.e. k-Means and k-Medoids. Clustering algorithm is dependent on the type of data chosen for processing. The data points are clustered according to the distribution of arbitrary shapes of the data points. Partition based algorithms are known to perform well to analyze small or medium datasets to identify cluster of spherical shape. The results of both the

algorithms are analyzed on the exact figure of data points and the computational time required for each algorithm. Time complexity analysis is a part of computational complexity theory which is used to describe an algorithm's utilization of computational resources; in terms of the best case and worst case execution time expressed.

The behavior of algorithm is analyzed using performance as a benchmark by calculating the computational time required by each algorithm to process the datasets, Quality assessment is based on analysis and measurement of distance between two data points. The author observed that average execution time of the k-Means is comparatively less than k-Medoids algorithm.

Processing high dimensional dataset has always been a challenging task because of its multiple dimensions. K-means is determined as the best technique to partition a dataset into groups according to their patterns among all the partition based clustering techniques. It is also known that in k-means algorithm results are dependent on the starting points known as centroids. Dimensionality reduction is an important task in the determination of centroid. Numbers of techniques have been proposed for the purpose of improving k-means efficiency but as compared to these techniques results shown by the method proposed by Tajunisha and Saravanan [3] are far better showing near perfect accuracy. K-means algorithm application results depend on the initial value of centroid. To find starting centroid for k-means the author has proposed the usage of Principal Component Analysis (PCA) for dimensional reduction of the datasets and heuristics approach to reduce the number of iterations in distance measurement in assignment of data point to clusters. Reason for application of PCA on microarray data prior to application of clustering technique is to improve the accuracy of the obtained results based on the assumption that improvement can be the linked to the application of centroid values obtained by proposed method that are very much close to the optimum solution.

The authors compared results obtained by the application of k-means algorithm on microarray data with randomly initialized centroid and PCA generated centroid. Execution time of proposed technique was less than the average k-means execution time with random centroid initialization. The comparison of the results on IRIS dataset of UCI machine learning repository show that proposed technique is more effective, accurate and efficient than the existing methods.

Accuracy and efficiency in clustering of a really large high dimensional datasets having huge number of samples is quiet a challenging task. To resolve this issue it is usually advised to apply data reduction techniques in order to achieve the ultimate goal of acquiring efficiency and accuracy. The application of vertical data reduction techniques is required prior to implementing clustering technique. However, dimensionality reduction methods has disadvantage of damaging the result's quality and causes data loss. In this study Khalilian, Mustapha, Suliman and Mamat [4] have proposed a method for improving K-Means algorithm performance by using divide and conquer strategy, i.e. application of Hill Climbing algorithm, in relation to compatibility and equivalence concept.

This study presents experimentation on data from PEIVAND web site, purpose of this website is to find suitable partners who have similar personality's based on 8 pages of psychiatric questions for assessing personality. Different aspects of the hidden data patterns is extracted based on the recursive simulation of proposed method i.e. use of both HC and K-Means algorithm's advantageous attributes. Property of K-Means is that, it is more qualified in dealing with big sized data set and dimension yielding low quality clusters, whereas HC algorithm has the capability of constructing structured clusters with high quality attributes but possesses complexity issues making it unsuitable to be used for huge datasets having high dimension data. Using the advantageous aspects of these two techniques allows us to define similarity within referred domain and it also allows us to subdivide our space according to certain criterion. The results of proposed method application demonstrate remarkably improved efficiency and accuracy in cluster formation allowing creation of structured / nested clusters as end result.

Reference [5] proposed a new algorithm by combining "A Fast DBSCAN" and "Memory effect in DBSCAN" to speed up the performance and to improve the output quality of the DBSCAN clustering technique. As the original DBSCAN algorithm uses the distance measures to compute the distance between objects, therefore it requires too much processing time.

To know the real performance difference achieved in the new algorithm, authors did not use any additional data structures (like spatial tree) to improve the performance. The new algorithm has small number of object loss than the Fast DBSCAN algorithm. The proposed algorithm analyzes all the border objects during the clustering process.

Reference [6] made an attempt to enhance the clustering effect on uneven density distribution of the datasets with high-dimensional attributes. The authors put forward an improved version of density distribution function based clustering algorithm using local-scale and boundary limits. The proposed technique employs partition, hierarchy, density, grid and model driven clustering techniques. Density point is with maximum value is identified as the centroid from which the cataloging is prolonged to density based boundary threshold. K Nearest Neighbor (KNN) method is applied for measuring each point density and then a center point is defined having the maximum density value / point. The proposed technique enhances density-based clustering algorithms sensitivity to parameters and improves the clustering on uneven density distribution of the high-dimensional datasets.

One of the major difficulties of data mining in medical field is the identification of understandable information from spatial data. Aim of clustering is to classify data into different clusters by partitioning datasets into subsets. Spatial data mining can handle huge amounts of spatial

knowledge gathering data collected through numerous applications. Pratap, Devi, Vani and Rao [7] propose a density based k-medoids clustering algorithm to overcome the limitations in DBSCAN and k-medoids clustering algorithms. Weka software has been utilized as a tool for algorithm execution and testing the proteins data base generated by Gaussian distribution function. Proposed algorithm's performance is better than DBSCAN. Clustering technique proposed has an efficient way of identifying information from raw data utilizing K-means and K-medoids as basic methods, it also caters circularly distributed data points clusters and marginally overlapped clusters. Based on quality of classification measured by Rand index the proposed Density based K-medoids technique performed better than combined implementation of DBSCAN and k-medoids.

Most of the classical clustering techniques algorithms are dependent on static statistical data to identify the hidden patterns and relationships. As an upcoming requirement from the implementers of critical systems, dynamic data also needs to be clustered. For the purpose dynamic clustering techniques are required to be implemented. By dynamic clustering we mean to identify & analyze the clusters in live data environments. Clustering and visualizing high-dimensional dynamic data has always been great challenge for researchers.

To coup up with this problem incremental data mining is one of the many solutions with practical application in data warehousing and sensor network, both of these are dependent on dynamic data clustering algorithms. Mary and Kumar [8] have addressed the problems of clustering a dynamic dataset in which the dataset is increasing in size over time by adding more and more data. Density based dynamic data clustering algorithm for incremental clustering dataset has been evaluated, to compare performance of proposed techniques with normal DBSCAN, Chameleon algorithms on the dynamic dataset and successful implementation and evaluation of the proposed density based dynamic clustering algorithm has been claimed.

For the purpose of examining the performance of the proposed algorithms in comparison to DBSCAN and Chameleon algorithm UCI Data repository live datasets have been used. On comparing the Chameleon and DBSCAN algorithms with proposed algorithm it was concluded that proposed algorithm's performance was significantly better for efficiency and accuracy. The speed of the proposed algorithm was better than Chameleon and normal DBSCAN algorithm.

In past corporate and military organization were the only ones having access to the location-acquisition technology. But with the advancement of technology and reduced costs ordinary people are also able to lay their hands on global positioning enabled devices generating huge amount of spatiotemporal data and enabling them to record location based personalized data. On the other hand communication technology advancement allows them to transmit this data to internet to be shared within their social circle. This new type of data may include location-tagged images, GPS coordinates, and mobile networks cell / sensors properties data.

The availability of huge amount of spatiotemporal data and its properties has opened new arenas for research and its application in various fields. As a consequence of this availability, new investigative techniques are required to be developed because of challenges posed by processing diverse nature of generated data. Kisilevich, Mansmann and Keim [9] have proposed, a new clustering algorithm P-DBSCAN which is built on original DBSCAN clustering algorithm to process and analysis geo-tagged pictorial spatiotemporal data of events and places of Washington, D.C.

Following two improvements in original definition of DBSCAN were introduced:

- Adaptive density approach to optimize search for dense zones and rapid correlation of algorithm with high density clusters.
- A well-defined density benchmark based on the statistical figures of people taking picture in the neighborhood.

Based on critical analysis it was concluded that adaptive density approach is the key ingredient of creating packed clusters with high density on one hand whereas on the other hand density based on ownership leads to creation of smaller clusters with low density.

Various domains of knowledge discovery have utilized Self-organizing Map (SOM), a renowned neural networking technique model, as tool for analyzing data. It also provides mapping technique based on topology-preserving to a lower-dimensional output space from a high-dimensional input space. Hidden features and characteristics of underlying data can be explored, identified and analyzed by utilizing powerful tools of sophisticated visualization techniques. Mayer and Rauber [10] is evaluating novel visualization technique enabled to explore and present structure inherent in the datasets by its application on benchmarked data. Proposed technique is not only able to expose similar data object groups, but also facilitates visualization of similar data objects in graph formation of identical datasets by utilizing minimum spanning trees. This technique is well capable to identify the similar attribute groups on the basis of graphs drawn using either input data or the SOM nodes. MST visualization technique is used for presenting the relationship & dependencies of the categories utilizing minimum spanning tree technique. On comparing this technique with density graph method by their application on a dataset of three categories of Iris flowers by their characteristics, i.e. length of Sepal, width of Sepal, length of Petal and width of Petal, consisting of 50 samples in each category; it was concluded that both techniques produce almost similar results. The only difference identified was that SOM is more efficient than density graph method and the reason for this is that SOM nodes have smaller magnitudeas compared to number of data samples. The visualization does not require expert

user making it a more user friendly technique for rookie users.

To minimize the risk in stock trading investors select portfolios. Portfolio selection is the process of minimizing financial investment risk by distributing investor's capital into set of stocks. Silva and Marques [11] have strived to develop a tool for helping investors to select stocks for investment purpose. The idea is to build an intelligent mechanism that will help the investors in the process of selecting portfolios. SOM is proposed because of its relational and visualization characteristics using Component Planes technique, which has the capability of identifying the indirect and hidden associations in input data making it an ideal technique for modeling stock market systems. Implementation of improved *Rv*co efficient is suggested for the purpose of comparing Component Planes and to produce a distance matrix between features leading to hierarchy oriented clustering method used to obtain feature's clusters. The main aim of building portfolio is to diversify the investor's profile by limiting the purchase of dissimilar stock as it is risky to invest in the stock of similar behavior.

For the purpose of normalization of values of different stocks results were interpreted. Analysis of the values obtained by distance matrix of two products determined that they are not closely correlated. It was also observed that feature clustering based on SOM is good option to cluster time-series data. The proposed technique was used to visualize time-series data of insurance companies stocks and financial institutions stocks presenting good empirical results for portfolio determination. On the basis of these finding it can be safely assumed that consecutive clusters are usually correlated but depend heavily on organizational performance.

With the increase in the amount of data, requirement for the better information extraction technique with better presentation of data is growing, for the purpose clustering techniques of data mining are utilized. Clustering is process of separating data into groups. K-Means is one of the famous clustering techniques due to its simplicity and efficiency. When k-means id used to cluster large datasets it has accuracy of problems. Therefore there is a need to improve this technique, making it suitable for processing all kinds of data. To solve this problem Sakthi and Thana Mani [12] have proposed to include some constraints in the algorithm naming it Constrained K-Means Clustering. The constraints introduced are Cannot-link constraint, Must-link constraint, ε-constraint and δ-constraint. To generate Cannot-link and Must-link constraints SOM has been utilized.

Data from UCI machine learning Repository has been used to test proposed technique using two datasets i.e. IRIS and Wine Dataset.50 samples of four features for each species of three Iris flowers (virginica, setosa and versicolor) were taken. Fisher developed a discriminant linear model, on the basis of four feature's combination, for differentiating species from each other. Experiments of both techniques i.e. existing K-Means algorithm and constrained K-Means were performed on the datasets to evaluate the efficiency and accuracy differences between the techniques. Based on experimental results it was concluded that proposed technique produces centroids more closely to true centers more efficiently as compared to traditional method. Hence it was proved that proposed method is capable of classifying all kind of data accurately with less time requirement to process.

Researchers tend to use fuzzy C-Means algorithm (FCM) for Computational geometry, pattern recognition, data compression, vector quantization, and pattern classification. Velmurugan and Santhanam [13]are presenting an efficient application of FCM clustering algorithm on three types of inputs; i.e. Data points are first distributed manually, Statistical distributions of normal data points using the Box-Muller formula and Statistical distributions of uniform data points using the Box-Muller formula; are given to algorithm. Clustering quality is the benchmark for smooth execution of the algorithm and the data points along with number of clusters determine the behavior of the algorithm. Multiple executions of the program elaborate the performance of the algorithm based on the input data points and the execution time required. The execution time required for the execution of the algorithm is dependent on the power of the processor and available computational resources.

In broad spectrum picture results obtained from the execution of Java based program shows that execution of program on Uniform distribution of input data points required much lesser execution time as compared to time required for execution of program on manual and Normal distribution of data points, whereas on comparing the time required for manual distribution with normal distribution it was found that manual distribution require lesser time. On the basis of iterative executions of the program it was concluded that FCM algorithm is efficient for Uniform distribution of input data points.

C-means clustering algorithm is highly dependent on centroid value and is sensitive to noise in data therefore data requirements are set on the higher end, in this paper it is proposed to utilize information entropy method to identify the initial centroids. Introduction of weighting parameter is also suggested for making adjustments in the cluster center location to reduce the noise problem, to reduce its dependence on the centroid and to make arbitrary shape clusters and later on sort them through certain predefined rules. For the purpose of testing the idea experiments were conducted utilizing the IRIS as the dataset to verify the functionality of suggested modifications to the algorithm. The test dataset contains 150 iris species information which was distributed into following three Iris classes i.e. setosa, versicolor and virginica. Each category consist of 50 type's datasets containing five kinds of attributes with each dataset.

Reference [14] claim, based on experiment's findings, that improved c-means model will be able to differentiate arbitrary shape clusters, and reduce the dependence of algorithm on centroid. Experiment results also showed that the error fraction of the entropy weighting c-mean algorithm is much lower than actual algorithm.

Suggested modifications tend to improve the algorithm's efficiency.

Gath–Geva (GG) algorithm is a famous procedure for numerical data clustering in fuzzy c-means technique which based on assumption that GG generated clusters are more flexible than FCM generated spherical clusters. To handle mixed category and numerical attributes data; traditionally fuzzy k-prototypes algorithm is used which is an extended version of FCM but it does not use same dissimilarity function. To cater this issue Chatzis [15] has proposed a GG algorithm extension for effective handling mixed numeric and categorical properties data. The method proposed is dependent on KullbackLeibler (KL) fuzzy objective function and is based on probability oriented supposition regarding shape of the clusters derived, providing important in-built advantage over the novel objective function of FCM algorithm.

Efficacy of the proposed method is compared with other non-fuzzy and fuzzy techniques for clustering on standard data from UCI repository of Vote dataset, Heart disease dataset and Australian credit approval dataset. Quality of the clustering results is assured by using used data ground truth labeling. Number of objective criteria were employed to measure intersection between obtained results and used data ground truth results. The experimental results for performance evaluation of proposed method for pattern recognition show that the proposed technique's performances far efficient and in improved manner than other competing non-fuzzy and fuzzy clustering algorithms for mixed category and numerical attributed data.

Iqbal et al. [16, 17] proposed performance metrics for software design and software project management. Process improvement methodologies are elaborated in [18, 19] and Khan et al. [20] carried out quality assurance assessment. Amir et al. [21] discussed agile software development processes. Khan et al. [22] and Khan et al. [23] analyzed issues pertaining to database query optimization and requirement engineering processes respectively. Umar and Khan [24, 25] analyzed non-functional requirements for software maintainability.

Khan et al. [26, 27] proposed a machine learning approaches for post-event timeline reconstruction. Khan [28] suggests that Bayesian techniques are more promising than other conventional machine learning techniques for timeline reconstruction. Rafique and Khan [29] explored various methods, practices and tools being used for static and live digital forensics. In [30], Bashir and Khan discuss triaging methodologies being used for live digital forensic analysis.

Table 3. Critical evaluation

Reference	Methodology	Proposed Solution	Strength	Weakness	Suggestive Improvements
[1]	M-Means	K-means clustering algorithm has been applied as an efficient and simple tool to monitor performance of students.	Proposed application of k-means clustering algorithm and Euclidean can act as a baseline for progression evaluation of the students' performance in institution of higher study.	The proposed application of both techniques do not present any modification to reduce effort redundancy and resources required for its application.	For the application of this technique, efficient centroid determination mechanism to reduce redundant efforts required for random sampling technique should be incorporated.
[2]	K-Means & K-Medoids	Efficiency of K-Means & K-Medoids has been assessed on the basis of time required for computing small and medium datasets	This research work presents k-means as more efficient and accurate technique for clustering arbitrarily distributed input data points in comparison to k-Medoids	Type of data and application of that data in particular scenario determines the selection of a particular clustering algorithm.	The experimentation was executed on arbitrary dataset to perform performance evaluation it is suggested that same experimentation may be performed on actual data from data repositories.
[3]	K-Means	PCA has been applied on the dataset prior to the application of clustering technique for the purpose of obtaining the initial centroid and clustering data into lower dimensions.	Application of three principal components along with usage of PCA technique scrutinized about 99.48% of processed data consistency causing bare minimum loss of data with aspect of dimension reduction.	The proposed technique should have been applied to multiple types of data sets for the purpose of evaluating the actual potential.	It is suggested that the proposed technique may be tested / experimented / applied to a variety of datasets for the purpose of exploring new avenues and possibilities.
[4]	K-Means	For improving the K-means algorithm's application performance in high dimensional datasets, proposed method uses the Hill climbing technique for compatibility and equivalency in relations concept	Proposed algorithm is based on K-Means complexity that is in lines with amount of samples, clusters, iteration and dimensions therefore it is scalable.	Both techniques i.e. Hill Climb and K-Means utilize similarity measurement of pairs which not very efficient and time consuming and the technique should have been tested on different types of datasets.	Experimentation with different Data type can be an interesting application of this methode.g. Improvement in processing data stream. Amount of sub spaces determination should be studied.

[5]	Density Based Clustering	DBSCAN algorithm has been improved and a new algorithm has been presented as ODBSCAN.	Proposed ODBSCAN algorithm improves the performance of clustering along within advantage of reduced object loss.	No additional data structure has been utilized to improve and analyze the real performance of the proposed technique.	For obtaining the accurate results with similar performance it is important to address the problem of possibly missing core objects during data loss.
[6]	Density Based Clustering	Density distribution function based improved clustering algorithm has been proposed using the ideas of local scale and boundary threshold	The proposed algorithm improves the clustering result of the high-dimensional datasets having uneven density distribution and also enhances the density-based clustering algorithm's sensitivity to parameters	Quality aspects of clustering using improved DENCLUE, have not been discussed, as quality of clustering results dependent on the two parameters: noise threshold and density	Clustering quality aspects based on two influencing parameters of DENCLUE algorithm should be studied.
[7]	Density Based Clustering & K-Medoids	An effective density based k-medoids algorithm has been proposed for overcoming the problems in DBSCAN and k-medoids algorithms	Proposed K-Medoids clustering algorithm can provide improved clustering results having a huge scope of enhancing the clustering of medical image datasets.	K-medoids has severe problems of resource and time consumption, which needs to be addressed before its application.	To address issues of application proposed algorithm in any particular domain combination of multiple techniques can be utilized for effective and efficient clustering.
[8]	Density Based Clustering	For clustering incremental dataset a dynamic clustering algorithm has been presented.	The efficiency of the proposed algorithm better than DBSCAN and Chameleon algorithm.	In dynamically altering dataset operations like modifications of data points, deletions and remodeling the algorithm for data clustering have not been addressed.	The future work may address operations like modifications of data points, deletions and remodeling the algorithm for data clustering in dynamically altering dataset.
[9]	P-DBSCAN	P-DBSCAN algorithm has been presented as an improvement to DBSCAN clustering algorithm for processing collection of geo-tagged photos	Specialized for the problem of analysis of places and events using large collections of geo-tagged photos.	Different aspects of the proposed approaches were not mentioned in this paper as it is an ongoing research.	Efforts should be focused on assessment approaches, runtime performance enhancement and database incorporation.
[10]	Self-organizing Maps	based on Minimum Spanning Trees a visualization technique has been presented for SOM	On the basis of graph built on data input of self-organizing maps nodes data, the proposed method is able to explore similar item groups	The proposed method is a generalized technique suitable for novice users, making this technique uninteresting for experts as clustering requires specific field related knowledgebase to deal with	The visualization can be implemented on other techniques enabling the user to view the numerous types information in one go without the requirement of comparing there figures.
[11]	Self-Organizing Maps	SOM based feature clustering method was presented.	Such kind of applications and techniques can help Investors analysis portfolio for identifying similar performance stocks and enumerating the stock on suitable time.	Mathematical evaluations with other techniques has not been provided for the purpose of assessing the quantifiable performance.	Mathematical assessments with other techniques may be provided
[12]	K-Means & Self-organizing Maps	To suit for all kinds of data Constrained K-Means Clustering has been proposed as an improved version of k-means.	Evaluation results of proposed technique display a significant improvements in efficiency and quality of clustering as compared to other techniques	Proposed technique requires to enhance its mechanism for noise reduction.	More appropriate constraints and mechanism should be defined for noise reduction in the Constrained K-Means.
[13]	Fuzzy C-Means	FCM clustering algorithm implementation in simple and efficient manner is presented.	FCM algorithm has shown improved efficiency for Uniform distribution of input data points.	Structure and behavior of proposed algorithm is similar to K-Means algorithm.	Performance of results for manual distribution and uniform distribution have bare minimal difference therefore efficient mean of uniform distribution needs to be explored.

[14]	Fuzzy C-Means	To initialize the cluster centroids and handling the noisy data implementation of information entropy centers along with c-means clustering algorithm has been proposed.	The results confirmed the improved fuzzy clustering algorithm based on the FCM algorithm can distinguish clusters of arbitrary shape, and greatly reduce the dependence on the initial cluster centers	In order to explore the number of duplicate centroids it is better to engage information entropy but there are still some defects remaining.	Apply the improved algorithm to specific areas (such as a data of banks, telecommunications data clustering) is also the direction of further research
[15]	Fuzzy C-Means	GG algorithm extension has been proposed to effectively handle data with mixed categorical and numerical qualities.	The experimental assessment of the design recognition performance of projected model various applications has revealed that it beats other algorithms for data clustering with varied numeric and categorical characteristics	This radical algorithms should be able to process any kind of dataset either numerical, graphical or textual	Other types of datasets should be dealt by this clustering algorithm making it a universal algorithm that can deal with all kind of data types.

III. Conclusion and Future Work

Clustering is the organization of related objects into similarity based groups of different shapes and density. In other words we can safely say that, the dividing a dataset into groups so that the data in each subset has mutual characteristic. Data clustering is a common method for arithmetical study of data, which is utilized by numerous fields of life, like bioinformatics, image analysis, pattern recognition and machine learning. While conducting this study, few aspects were found that lured us to continue our research in particular direction. By working further in the direction of covering the gaps identified implied clustering techniques can further be improved in the future using amalgamation of clustering technique to achieve more precision in result and reduce the time required for data and/ or information repossession from large data set.

References

[1] O. O. Jelili, O. O. Ojeniyiand I. C. Obagbuwa. Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance. International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, No. 1, 2010.

[2] T. Velmurugan. Efficiency of K-Means & K-Medoids Algorithms for Clustering Arbitrary Data Points. International Journal of Computer Technology & Applications (IJCTA), Vol. 3 (5) Sept-Oct 2012.

[3] Tajunisha and Saravanan. Performance analysis of k-means with different initialization methods for high dimensional data. International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010.

[4] M. Khalilian, N. Mustapha, M. N.Suliman and M. A.Mamat. A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. International Multi Conference of Engineers and Computer Scientists (IMECS). Vol. I. March 17, 2010.

[5] J.H. Peter and A. Antonysamy. An Optimized Density Based Clustering Algorithm. International Journal of Computer Applications, Volume 6– No.9, September 2010.

[6] J. Zhang, W. Li and J. Tan. An Improved Clustering Algorithm Based on Density Distribution Function. Computer and Information Science Vol. 3, No. 3; August 2010.

[7] A. R. Pratap A, J. R. Devi, K. S. Vani and K. N. Rao. An Efficient Density based Improved K-Medoids Clustering algorithm. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 6, 2011.

[8] S. A. L. Maryand K.R. S. Kumar. A Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset. Journal of Computer Science 8 (5) 2012.

[9] S. Kisilevich, F. Mansmann and D. Keim. P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. 1st International Conference and Exhibition on Computing for Geospatial Research & Application Article No. 38 ACM New York. 2010.

[10] R. Mayer and A.Rauber. Visualizing Clusters in Self-Organizing Maps with Minimum Spanning Trees. K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part II, LNCS 6353, pp. 426–431.Springer-Verlag Berlin Heidelberg. 2010.

[11] B. Silva and N. Marques. Feature Clustering With Self-Organizing Maps and an Application to Financial Time-Series for Portfolio Selection. International Conference on Neural Computation (ICNC). 2010.

[12] M.Sakthi and A. S. Thanamani. An Efficient Constrained K-Means Clustering using Self Organizing Map. International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4. April 2011.

[13] T.Velmurugan and T.Santhanam. Clustering Mixed Data Points Using Fuzzy C-Means Clustering Algorithm for Performance Analysis. International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 3100-3105.

[14] X. SU, X. WANG, Z. WANG and Y. XIAO. A New Fuzzy Clustering Algorithm Based on Entropy Weighting. Journal of Computational Information Systems (JOFCIS) 6:10 (2010) 3319-3326. October, 2010.

[15] S. P. Chatzis. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Systems with Applications 38, 8684–8689. (2011).

[16] Iqbal S., Khalid M., Khan, M N A. A Distinctive Suite of

Performance Metrics for Software Design. International Journal of Software Engineering & Its Applications, 7(5), (2013).
[17] Iqbal S., Khan M.N.A., Yet another Set of Requirement Metrics for Software Projects. International Journal of Software Engineering & Its Applications, 6(1), (2012).
[18] Faizan M., Ulhaq S., Khan M N A., Defect Prevention and Process Improvement Methodology for Outsourced Software Projects. Middle-East Journal of Scientific Research, 19(5), 674-682, (2014).
[19] Faizan M., Khan M NA., Ulhaq S., Contemporary Trends in Defect Prevention: A Survey Report. International Journal of Modern Education & Computer Science, 4(3), (2012).
[20] Khan K., Khan A., Aamir M., Khan M N A., Quality Assurance Assessment in Global Software Development. World Applied Sciences Journal, 24(11), (2013).
[21] Amir M., Khan K., Khan A., Khan M N A., An Appraisal of Agile Software Development Process. International Journal of Advanced Science & Technology, 58, (2013).
[22] Khan, M., & Khan, M. N. A. Exploring Query Optimization Techniques in Relational Databases. International Journal of Database Theory & Application, 6(3). (2013).
[23] Khan, MNA., Khalid M., ulHaq S., Review of Requirements Management Issues in Software Development. International Journal of Modern Education & Computer Science, 5(1), (2013).
[24] Umar M., Khan, M N A., A Framework to Separate Non-Functional Requirements for System Maintainability. Kuwait Journal of Science & Engineering, 39(1 B), 211-231, (2012).
[25] Umar M., Khan, M. N. A, Analyzing Non-Functional Requirements (NFRs) for software development. In IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS), 2011 pp. 675-678), (2011).
[26] Khan, M. N. A., Chat win, C. R., & Young, R. C. (2007). A framework for post-event timeline reconstruction using neural networks. Digital investigation, 4(3), 146-157.
[27] Khan, M. N. A., Chat win, C. R., & Young, R. C. (2007). Extracting Evidence from File system Activity using Bayesian Networks. International journal of Forensic computer science, 1, 50-63.
[28] Khan, M. N. A. (2012). Performance analysis of Bayesian networks and neural networks in classification of file system activities. Computers & Security, 31(4), 391-401.
[29] Rafique, M., & Khan, M. N. A. (2013). Exploring Static and Live Digital Forensics: Methods, Practices and Tools. International Journal of Scientific & Engineering Research 4(10): 1048-1056.
[30] Bashir, M. S., & Khan, M. N. A. (2013). Triage in Live Digital Forensic Analysis. International journal of Forensic Computer Science 1, 35-44.

Authors' Profiles

Muhammad Ali Masood obtained Master's degree in Business Administration with specialization in Human Resource Management from National University of Modern Languages, Islamabad in 2005 and Computer Science with specialization in Software Engineering from University of Arid Agriculture, Rawalpindi in 2001. In 2014, he has completed his Masters of Science degree with specialization in Software Engineering from Shaheed Zulfikar Ali Bhutto Institute of Science and Technologies, Islamabad. Currently he is serving in National Information Technology Board (Ministry of Information Technology, Islamabad) as Project Director (HMIS). Mr. Masood is a certified Project Manager and has obtained his PMP certification with good standing in 2010.

M. N. A. Khan obtained D.Phil. degree in Computer System Engineering. His research interests are in the fields of software engineering, data mining, cyber administration, digital forensic analysis and machine learning techniques.

How to cite this paper: Muhammad Ali Masood, M. N. A. Khan,"Clustering Techniques in Bioinformatics", IJMECS, vol.7, no.1, pp.38-46, 2015.DOI: 10.5815/ijmecs.2015.01.06