CHAPTER 1
INTRODUCTION

1.1 Data Mining

The word data mining is known as the technique which deals with the removal or distillation of unseen predictive knowledge from large database. It includes different sorting of data through large amounts of data sets and discover useful and essential information from it. It is commonly applied in the field of business, and also plays a major role in the field of finance, but its most important in the field of science in this area it is used to refine valuable information from large data sets produced by latest experiments, techniques and observational methods. It is seen as a progressively important tool to transform data into business intelligence and can be achieve useful result. Data mining is usually the task of finding new interesting and uncommon sequence or patterns from stored data. It usually discovers the sheer within the data which is beyond the simple analysis with the help of algorithm. Data mining can be enforced on processes of knowledge discovery as well as prediction. As the knowledge discovery and data mining are mainly put to use conversely. Knowledge discovery is the process of choosing necessary information from the data. The process of mining using algorithms to uncover hidden information and patterns derived by the KDD process. Data mining techniques include the integration of various disciplines of machine learning approaches, high performance computation, recognition of patterns, neural networks, various data visualization, retrieval of information and statistics etc. Various tools of Data mining predicts future sheer and management of allowing businesses to create ardent knowledge-driven decisions. With extremely large amount of data increasing every year, more data is gathered and hence it becomes an important tool to transform data into useful information. Extensive process of research and product advancement expanded data mining. This progress started when business related data was feed in computers and continued with advancement in accessing data and new technologies that is helpful in exploring data in real time. Data mining is the process of change through access to data for reflection and expected navigation and dedicated to delivering information.

The two target of data mining are as follows:

Prediction process: It performs inferences on the current data to make prediction and makes profit by using already stored values in form to figure out hidden or expected results.

Description process: It characterizes and modeled the basic properties of data in the database which focuses on discovering patterns which define the data and the consecutive submission of user interpretation.

1.1.1 Motivation for Data Mining

The conceptual motivation behind Data Mining is getting some new innovative ideas for the improvement of existing business and technologies as well as applying these technologies in making some great decision making.

There are following points that motivate the organization for using data mining technology and to boost up their working.

- **Large Databases and Data warehouse:** Data mining collect its name from the fact that finding analyze valuable information from gigabytes. The increased use of electronic devices assembly as a staging data, remote sensing devices or web logging has contributed to this explosion of data. Data Warehouse gathers data from many sources, it rearranges it according to its uses and stores it in an operative repository that can be used for data mining for decision making.
- **Price drop in data storage and efficient computer processing:** The recent development of graphical initial statistical methods, the new machine learning procedure which depends on artificial intelligence, logical programming field and genetic algorithms unlocked the area for fruitful data mining. Whenever the tools of mining are implemented on high-performance parallel processing system, it can easily examines the huge amount of data in few minutes. Faster processing of data means that users can certainly experiment with more models for understanding complex data. The high processing speed makes it useful for users to determine massive amounts of data.
- **Growth in analytical methodology:** New and advance analytical models and algorithms, as exploration and data visualization, clustering and segmentation, neural networks, decision trees, memory-based reasoning and basket analysis gives greater analytical depth. So, mining data quality is now possible with the availability of new analytical solutions.

- **Data mining benefits:** There are great data mining benefits which are as follows:
    - Uncover and reduces the fraud.
    - Increase customer achievements and reservations.
    - Sales of goods and services based on combinations of market basket analysis.
    - Increase up selling and cross selling.

1.1.2 Process of Data Mining

The process involves following steps these are as follows:

- **Requirement analysis:** Data mining without a good idea can't be taken that what kind of results or outcomes the company is looking for, whatever the technique which are going to be used and the requirement of data are usually be different for different types of goals. If all the objectives have been clearly defined, it will be easy to evaluate and calculate the results of the project. After achieving the desired goal further steps are taken as a continuous process.
- **Data selection and collection:** In this step it consisting of finding the best source databases from which the data should be taken. The company has its own data warehouse from which they will be getting bulk of data. If the required amount of data is not available in the warehouse then the source OLTP systems need to be identified and all the necessary information extracted from it and stored in other temporary systems.
- **Cleaning and preparing data:** This task is usually done for cleaning the data if a data warehouse already contains the required data, the cleaning of data already been done when it was loaded in the warehouse. Normally this task is time taking a lot of effort is wasted during this task.
- **Data mining exploration and validation:** When appropriate and useful data has been taken and cleaned, it is likely to start the data mining exploration. It may be possible to take a sample of data and put it to a number of relevant techniques. For particular technique the result should be evaluated and their significance has to be interpreted. This is probably to be a repetitive process which should lead to the selection of one or more techniques which are suitable for further needs.

- **Implementation, evaluating, and monitoring:** When a model is selected and validated, it is authorized for the decision makers. This may include software development for report generation, results visualization and explanation for managers. It is then necessary to calculate the results and choose the best suitable technique and continuous monitoring is required.
- **Result visualization:** The most essential step is to read the result of data mining to the decision makers or to visualize the result regularly in this process.

1.1.3 Data Mining Architecture

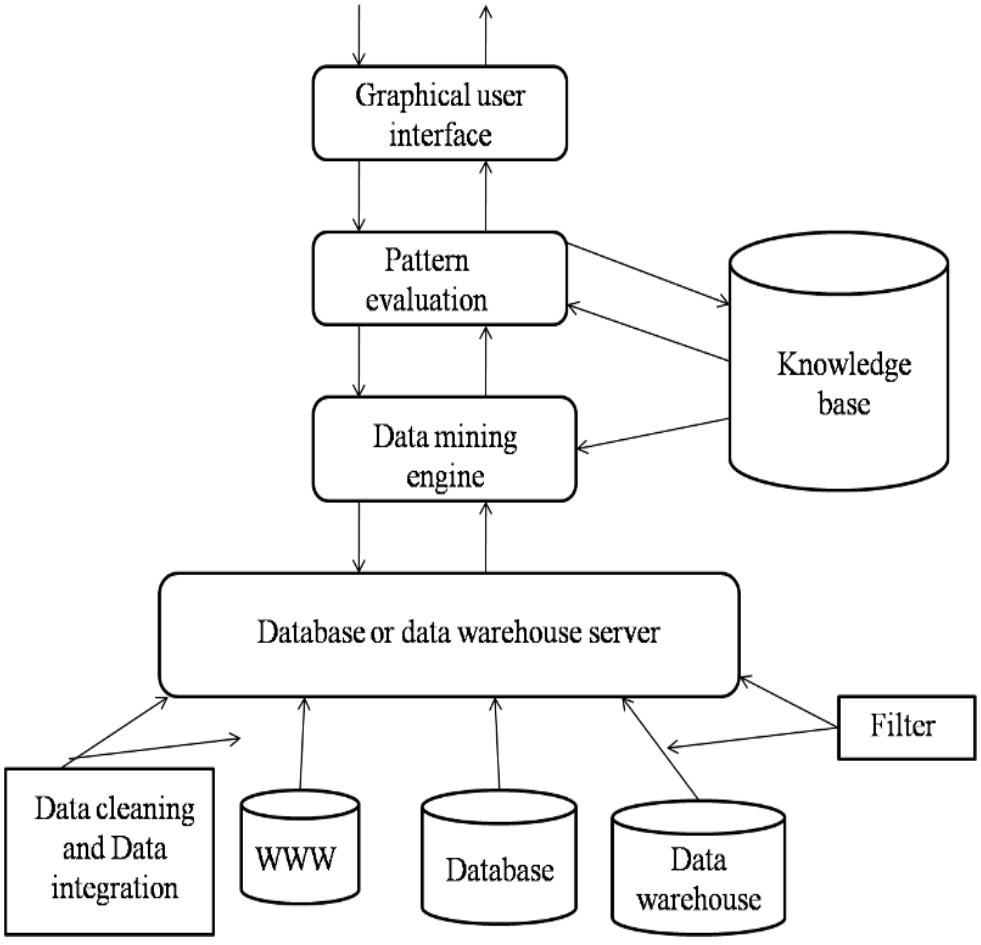


Figure 1.1.3: Data Mining Architecture [1]

The architecture consists of following components these are [1]:

- Database, data warehouse, World Wide Web: This contains databases sets, data warehouses and number of spreadsheets etc. Methods of data cleaning and data integration are carry out on the data. From this a user can get an enormous amount of data for processing it further in the development of science and business field.
- Database and warehouse server: In this component, database servers and data stores are liable for collecting the necessary data based on users questions.
- Knowledge base type: It is the important real idea consider to monitor and evaluate the search for interesting patterns of results. This type of beneficial knowledge can involve view ordering that are used to constructs attributes and their values at next levels of thinking.
- Data mining engine: The parts of the architecture system that consists of modules to perform various things, such as characterization of data, classification of data, prediction functions, analysis of cluster and outliers.
- Pattern evaluation type: It is normally used interesting portions and also communicates with mining module that focuses the search for interesting patterns. Can be used as a threshold level of interest to the well known filtering patterns. The model evaluation component can be unified with the mining module, which depends entirely on the method used for the application of data mining. Data mining for effective and efficient, it is strongly considered to accelerate the calculation of portion of interesting mining process for finding interesting patterns only.
- Graphical u/i: The component convey between users and extraction structure takes place which gives authorization to the user to cooperate directly with the system query or task and also provides information about search, and performs basic data mining which is depending on the results of current process. It also allows the user to explore design databases and data warehouses or data structures, performs the evaluation of the patterns and display the various forms of mined patterns. This component can directly deals with the user end as well as the overall structure of the system that will be easily helpful for the users to communicates with the absolute world application in a short period of time and can also take some effective measures and valuable ideas to implement into the processing of the mining process.

1.1.4 Data Mining Tasks

The task of data mining can be used to describe the essential features of the data when the data or a part of it is taken for the required tasks.

There are several Data mining tasks these are as follows:

- Class description task : The description of the class is used to depict the alone type of classes and different ideas in a short, brief and literal condition.
- Association task: The disclosure association rule which shows the total values of attribute in terms of that often comes jointly in a dataset.
- Classification: It comprise discovery of rules that partition the data into dislocate groups. It observes the data and builds a structure based on the standard label, and the desire of establishing a class label of future unlabeled records. As known class field so this type of classification is called supervised learning. Rules are created by this process, which helps to indentify future data and develops a structure in database. There are various models such as classification discovery decision tree types, genetic algorithms and statistical models.
- Clustering: It is known as the separation of data into similar objects groups. It is the main and valuable part in the field of mining, and is noticed as the initial stride of discovering interesting ideas. Clustering is the arrangement of data into different groups, so that the data in each and every group share common sheer and sequence. Objects are arranged and grouped according to the similarity of intra-class and the similarity of inter class.
- Outlier analysis process: The term outliers are those things which unable to meet the standard common behavior or data. Such kind of data objects displayed are uncommon or incompatible with rest sets of data is referred to as outliers. Many mining algorithms attempt to reduce or overcome the effect of outliers or eliminate them all together or can be discovered using the distance measurement and statistical test.
- Evolution analysis: Usually describes the trends and patterns of objects whose behavior reflects as the time passes. It also comprise of number of time series analysis of data, sequence matching and the type of similarity based process of analysis of data.

1.2 Introduction to Clustering

1.2.1 Clustering

Clustering is that type of process of arranging data in other groups, so every group containing those data share common sheer and patterns. It blends important class types of data mining working algorithms process. Efforts algorithm working which partitions the data field into a set of regions or groups, which are selected as that type of portion of the table, either deterministic or may be any kind of probability-wise. The main objective is to recognize all sets of alike occurrences in those type of data, in some excellent manner. On the basis of similarity clustering it is an idea that arises in different types of other ideas. If an area is available which is same, then there is a large and many types of methods and techniques for creating clusters. And also there is different kind of approach for construction of the adjustment functions that calculate a different property of the groups. Following this approach achieves what is known as the separation optional.

The few goals of clustering are:

• To disclose common grouping types.

• To begin assumption of the data.

• To achieve constant and right type of data.

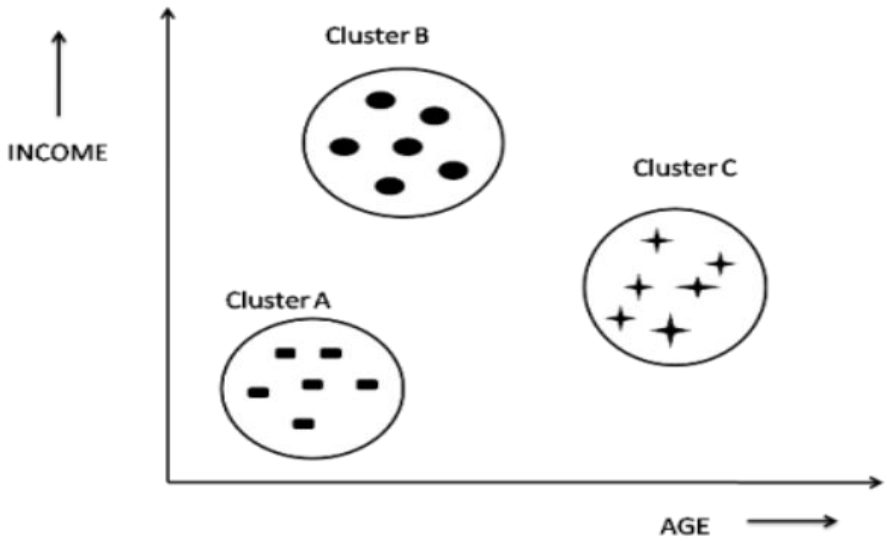


Figure 1.2.1: Clustering example

The above figure shows a dataset of customer which consist two attributes: income and age. The clustering method after applying on the data set groups the data sets into three segments for the two attributes. Cluster A shows the data of young population with low income. Cluster B shows middle aged people with high incomes. Cluster C represents a group of some senior peoples which has low income.

Clustering refers to an unsupervised task of data mining. No particular attribute is used to start the guidance process. Total input attributes are evaluated fairly. The majority of clustering algorithm to build the model through an iteration number and stop when the model is gathered when the boundaries of these segments are stabilized.

1.2.2 Requirements of Clustering

Clustering is an area that demands research into its possible appliance pretence their rare demands. There are several clustering requirements in data mining which are as follows [1]:

- Scalability: Many algorithms of clustering perform good on small sets which contains fewer data objects; since, large database may consist of huge amount of objects. Such that clustering in a field of very large sets of data can lead to partial type of results. So algorithms process of high working capability is required.
- Ability to deal with dissimilar types of characteristics: Different types of algorithms process are performed to cluster that is based on numerical data intervals. However many types of applications may depend on the clusters of other data, such as data and binary mixtures of these categorical and ordinal data types.
- Exploration of clusters with random shapes: Many types of clustering algorithms process is taken out on the basis of Euclidean and Manhattan distance measurements. The cluster can posses with different kinds of shapes and sizes. It is very essential to develop those type of algorithms that can detect groups randomly.
- Vital requirements for basic knowledge to compute input guidelines: Several algorithms of clustering needs positive guidelines in the cluster analysis by user. The clustering result can be quite complicated for the input values. The values are usually difficult to be achieved, for high dimensional data. This can lead to the decrease in the clusters quality when not being properly implemented.

- Ability to handle data with noise: Many databases contain real world data outliers or missing value. Most of the algorithms didn't work properly and are delicate to this type of poor quality data and hence degrade clusters.
- Additive clustering and lethargy to the order of aided records: Few types of algorithms can not include new data in existing group structures. For a given a set of objects type, the algorithm can recover completely from different group depend on delivery of input objects. It is necessary to decide increased algorithms that are unkind to the order of entry.
- High dimensionality type: A database consist of a lot of dimensions or many different attributes. A number of clustering process are fine at managing low type of dimensional data, includes only few dimensions. The search of clustered objects in high types of dimensional space is a challenge in taking that these data may be inadequate and highly distorted.
- Clustering on constraint types: Applications of real absolute world performs clustering process in various and different constraints. So this is a demanding work in clustering.
- Usability and interpretability: Users predicts the clustering results to be usable, understandable and interpretable. It is possible that the clusters to be attached to particular linguistic applications. So it is compulsory to test how a targeted appliance can be able to emerge as the valuable types of ultimate clustering process.

1.2.3 Types of Data in Clustering

As clustering works supports many types of data in performing its valuable process to develop some useful outcomes. So there will be many types of supportive data available in the absolute mining field.

Data used in clustering are as follows [1]:

- Data matrix type (object by variable types of structure): In this r object like a persons with q variable which is also called as attributes or measurements such as height, age, gender, weight etc. Structured relational table is presented or r by q matrix (r objects x q variables):

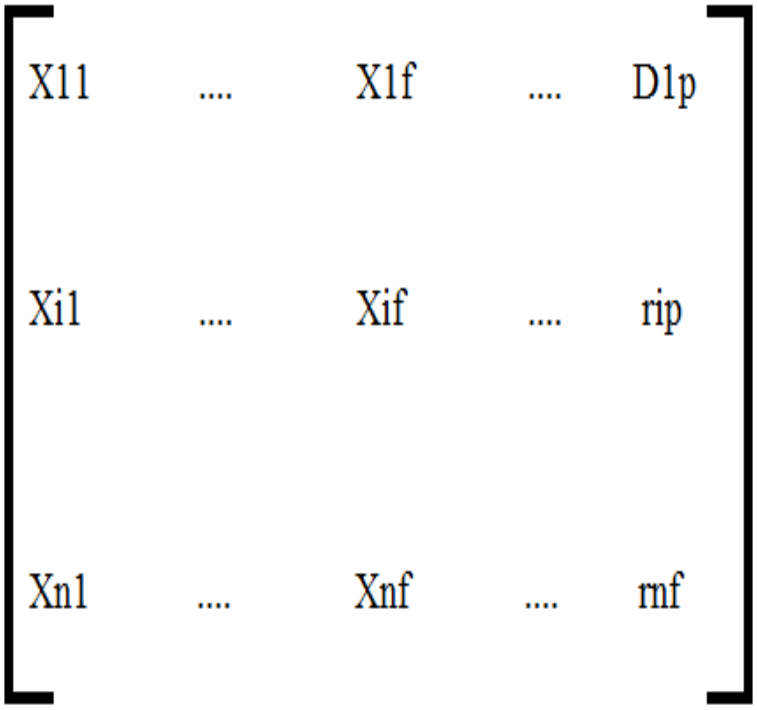


Figure 1.2.3(a): Data matrix [1]

- Dissimilarity matrix type (object by object types of structure): This has a number of closeness which are vacant for all pairs of r objects. It is usually represented by an r by r table:

  Where d(x, y) is the measured differences or dissimilarity between objects x and y. d(x, y) is a nonnegative number which is close to 0 when objects x and y are strongly same as the later one and it also shows the closeness among them so that it shows the number of differences measurement. Since d(x, y) = d(y, x) and d(x, x) = 0.

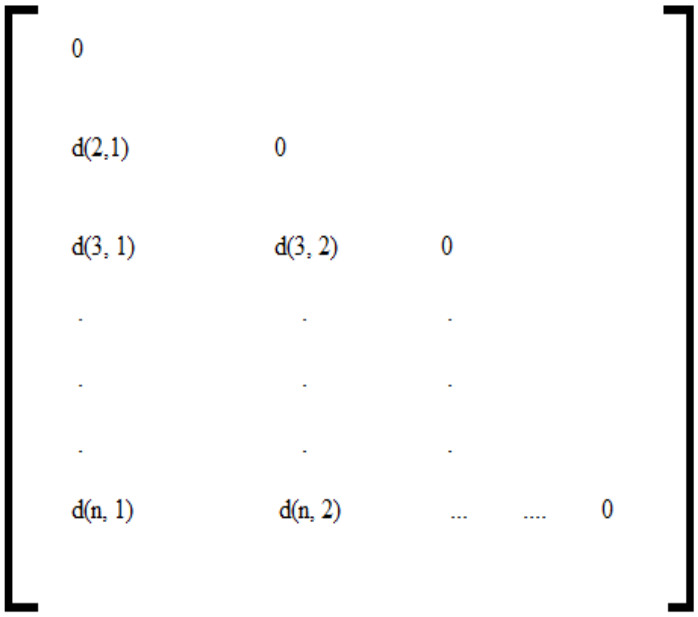


Figure 1.2.3(b): Dissimilarity matrix [1]

Classification of data types: Many types of data classification are as follows

- Interval scaled variables type: In this continues measurements of a linear scale is done. Examples are height and weight, age, coordinates, and temperature.
- Binary variable type: It contains two states 1 or 0, where 1 means absent variable, and 0 means that it is present. In this there are two modes these are as follows:
  (a) Symmetric binary variables: It is symmetric if both of its states have equal type of weights; it means there is no preferences on results seen as 0 or 1.
  (b) Asymmetric binary variable: If the result of the states are not very necessary, such as the negative and positive result of a disease test. By such assembly we will consider mainly the important result that is normally the limited one by 0 (e.g. HIV-) and the other by 1 (e.g. HIV+).
- Categorical variables and ratio scaled variables types:
  (a) Categorical variable: This type of variable are represented by symbolic references, such as red, orange etc.

(b) Ratio scaled variable: These variables make a positive measurement on a nonlinear scale. Simple examples include the decay of a radioactive element or the growth of a bacteria population.

- Variable of mixed data types: Several types of objects are characterize by a combination of different variables in several databases. All the six variable types data bases are listed above.
- Vector objects: These are symbolic references and are represented in vector form like X (a, b). These are mainly used to measure distance between two end points.

1.2.4 Types of Clustering

Clustering refers to an unsupervised task of data mining. No particular attribute is used to start the guidance process. Total input attributes are evaluated fairly. The majority of clustering algorithm to build the model through an iteration number and stop when the model is gathered when the boundaries of these segments are stabilized. The existing clustering algorithms can be categorized in four ways such as [1, 2]:

- **Partitioning**: Given a database tuples of n objects or data partitioning method creates k data partitions, where each partition is shown a cluster $k \leq n$. That is, it classifies the data in groups of k, which together meet the requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group.
- **Hierarchical**: In the hierarchical method develops hierarchical breakdown of data types objects. This process can be as any agglomeration or division, which is on the kind of hierarchical process is created. Hierarchical decomposition is shown or usually represented by a tree structure called dendogram.
- **Density based**: In this grouped objects based on covered range of objects. Such type of process searches for spherical clusters and groups find difficulty in finding arbitrary shapes. New clustering methods have grown on the basis of the notion of density. Their idea for increase of the given cluster as longer the density resides (number of objects or data points) in the "zone" will increase the saturation point for each type of data points, the neighborhood of a limit given must contain at least a minimum number of points.

- **Grid based**: This methods quantify the object type of space into a compact cells that form a grid type of structures. Grouping of processes on the grid type of structure (ie quantized space). The overall benefit of the method is of rapid type of processing amount of required time, This number is typically the number of independent and depends only on the data objects in the cell space quantization of each dimension..

1.2.5 Clustering in Subspace

As we know that in high dimensional space clustering is generally uncertain this means that we cannot easily recognize clusters in high dimensional space as the theoretical study which is proposed in [29] questioned about the meaning of the closest or nearest match in the high dimensional space. Since several researches work [30, 37] has paid attention towards the discovering of clusters which is entrenched in high dimensional data set subspaces. So this type of problem can be known as clustering in subspace and is based on the similarity measure.

1.2.6 Evaluation of Clustering

There are lots of approaches [27, 28] planned for calculating the outcome of the type of clustering based algorithm process. Each and every algorithms of clustering has some kind of advantages and disadvantages. The dataset groups having different size, density and with different shapes, so that various algorithms are important for detection of other types of groups. There is no other way to join the influence of these different clustering based algorithms and by neglecting their circumstances.

1.3 Introduction to Cluster Outlier Detection

In modern era there is lot of algorithms for data mining which basically focuses on methods of clustering. There are also different types of approaches and techniques designed for outlier detection process. There are some outliers are considered as those data objects which generally do not satisfied with basic behavior of the data model. Many algorithms reduce the effect of outliers or eliminate them. This can cause in the loss of very important information hidden because if there will be a noise from one side could be the signal for another side. In different

saying, the outliers is of general concerns for fraud detection cases, where outliers may identify fraudulent activity. So outlier detection and analysis is a task of data mining that is called as outlier mining or outlier detection.

There are several and unique types of techniques and approaches created and designed for outlier detection from the data. In several different conditions the meanings of cluster and different outlier are relevant and connected together, particularly for the data sets which contain lots of noise. So it is very important to deal with those types of clusters and those outliers as a perception of important factor.

There are lots of studies and research on cluster outlier detection.

1.4 Contribution to the Research

In this work we mainly target on the cluster outlier detection approach. In modern era there is lot of algorithms for data mining which focuses on methods of clustering. There are several and unique types of techniques and approaches created and designed for outlier detection from the data. Outliers are considered as those data objects which generally do not satisfied with the general and basic behavior of the data model. In several different conditions clusters and outliers meanings relevant and connected together, especially for the data sets which contain a lot of noise. So it is very important to deal with clusters as well as outliers as a perception of important factor.

In the chapter 3 we have to check the difficulties of grouping and finding of serious errors and make improvements.

- We will discuss a new procedure for determining the two clusters distances, two extreme values or outliers and the cluster and the outliers distances.
- We implemented few new formulas for defining the clusters and the outliers qualities and the data set partition.
- We will introduce an algorithm for the detecting the clusters and the outliers. So in this algorithm the clusters are discovered and managed on the basis of intra-relationship of the clusters and on the basis of inter-relationship between the clusters and the outliers. The whole management, fixing and correction of the clusters and of the outliers are done repeatedly before a reached termination.

1.5 Thesis Organization

- Chapter 2 in this it discusses about some of the most significant approaches to cluster and outlier detection in high dimensional data which has been noticed in the field of data mining.
- Chapter 3 it describes the overall methodology of the proposed approach and about its problem as well as the algorithm.
- Chapter 4 in this chapter we describe about the experiments which were conducted on the real world data set and also discusses about the tool as well as the result.
- Finally, Chapter 5 it concludes the overall conclusion and possible future extensions.

CHAPTER 2
LITERATURE REVIEW

This dissertation is intently linked to data mining and its exploration and especially for cluster outlier detection in high dimensional data. We had already mentioned in chapter 1, that there are many and different types of clustering algorithm as well as several types of outlier detection algorithm exist today. Now in part 2.1 we will investigate a short overview for different types of clustering algorithm and also some of their core concepts. In part 2.2 we will talk about different types of outlier detection algorithms.

2.1 Clustering of High Dimensional Data

Clustering is that type of process of arranging data in other groups, so every group containing those data share common sheer and patterns. It blends an important class types of data mining working algorithms process. Efficient algorithms to automatically partition the data field in a group or groups of regions, wherein a table is selected in the example, either deterministic or probabilistic wise. The main purpose of this method is to identify the data set of all instances of the same, in some of the best mode. Now according to the similarity grouping is an idea that arises in divers regulation. If an area is available in common, then forming of clusters will be in number of methods. There is another approach for construction of the adjustment functions to calculate a different property of the groups. Following this approach achieves what is known as the separation optional. An extremely flushed literature on clustering has been evolved by the former few decades. This is a problem ion the database, the theoretical literature, artificial intelligence and also incorporates our valuable importance in the field of science and business. The fundamental steps involved in developing a process is compiled as the selection function, applying a clustering type algorithm, the checking of the results and calculation. Apart from these measures, the algorithm of clustering and result checking is crucial, there are a lot of methods been proposed. The clustering algorithms can be summarized into four types such as[2]:

- **Partitioning method:** Tuples of n objects creates k data partitions, where each partition is shown a cluster $k \leq n$. That is, it classifies the data in groups of k, which together meet the following requirements: (1) each group must contain at least one object, and (2) each object

must belong to exactly one group. Detect that the second requirement can relax on some of the techniques of fuzzy partition. References to these techniques are given in the bibliographic notes. Given k, the number of partitions for a partition method for early partitioning. Then use an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The common prototype of a good deal is that objects in the same cluster are "close" or related to each other, while objects of the other groups are "distant" or very different. There are several other types of prototypes to judge the quality of the partitions. To achieve global optimality in partitioning based clustering require exhaustive enumeration of all possible partitions. Some of the popular portioning algorithm includes: K means algorithm [38], Graph based, and model based etc.

- **Hierarchical method:** This type of method forms a decomposition which is hierarchal for the data objects. This process can be classified as any agglomeration or division, depending upon the basis of how it is formed. The agglomeration method is also refers as the bottom-up approach, in this particular object begins with creating a isolated group. Gradually joins the objects or groups that are close to each other before all groups are joined into one which is the highest level of the hierarchy, or before a terminating condition is maintained. Partitioning method also refers as the top-down method which begins with all objects in the same category. In continuous iteration group is divided into smaller groups, until finally each object is in a cluster or before a terminating condition is maintained. Hierarchical methods undergo from the reality that once a step is completed then it will never be incomplete at any cost. This type of severity is beneficial for the lower costs of computation. However, such an approach can not correct erroneous decisions
- **Density based method:** Multi-partition method grouped objects based on the distance between objects. The method of this type can be difficult to find the clusters and the spherical shape of any group found. New and different types of clustering methods have grown on the basis of the of density. Their natural concept is to continuous growth of the available cluster. Such kind of valuable method can refine the noise (outliers) and discover agglomerate arbitrarily. DBSCAN in [39] and its extension, the optical based methods are essential growing density clusters according to an analysis of the normal connectivity based on density.

- **Grid based:** In this it quantify the object space into a finite number of cells that form a grid structure. Grouping all operations are performed on the grid structure (ie quantized space). The total no processing time is the main advantage of this method, which is independent for the objects and depends totally on the quantized space for each dimension. Commonly used methods include network-based cluster wave [14]. This is a method which is based on transformations of wavelets, which is able to meet all the above concerns. By the use of this type of method it is easily search the shapes of arbitrary data.

2.2 Outlier Detection Method

Outlier detection is the finding of the inconsistent behavior of objects. This is a very big and crucial sector in data mining field with the number of different applications such as detecting credit card fraudulent, hacking discovery, types of criminal act etc. At some point, is as important as cluster detection. Most previous investigations of this method is in the statistics field [3] these proposed methods typically make inference about the distribution of data, distribution of statistical parameters and the type of outliers. Still, these types of parameters can not be easily decided, and these methods are difficult to apply. In [4] proposed a method based on depth for the detection of outliers. According to the depth value of point data items are arranged in layers in the data space. These are expected to be outliers in the shallower layers. These methods do not have the problem of adapting distributions. But for multidimensional spaces that may not apply. In [5] introduced a method of detecting outliers based on the conclusion after searching for same data and a disturbing element in the data series is found as outliers. This method needs a function which can produce the degree for the dissimilarity of the data set to increase. Looking for the subset of data that leads to reduction in Kolmogorov complexity to the amount of data discarded. In [6] proposed algorithms for detecting outliers of distance -based. O considered a data point in a set of data T, a DB (p, d) - outlier if at least a part p of the data points in T is greater than the distance D from O. Their algorithm based index performs a range search with radius D for each data point. If the number of data points in the D - neighborhood pass a threshold, the search stops and that data point is deleted as an outlier not, otherwise it is an outlier. In the cell approach, in which the entire data space and quantify the data points for the cells. By sniping most numbers of red blood cells that has many data points

and contains neighbors, evaluate the process which neglect unusual type of cells and accelerates the detection of outliers. The algorithm of the cell based process is very effective when the total dimension number is decreased to 4 in this type of experiments. For a larger dimensions (≥ 5), the cells growing rapidly for those loops which is included in the paper i.e. cell-based algorithm. Besides the form of identification for the detection of outliers, the same authors (Knorr and Ng, 1999) provides intentional ideas for explaining the exceptional type of outliers. In [7] proposed a guided discovery approach in the search for errors the guide by computed except that the various levels of detail in the data cube. They assessed a value of the cell in the data cube will be the exception and it is suffering or if it is better than the expectation value based on a statistical model. Thus, the system will be able to deal with hierarchies and dimensions, and the example of this option is to deal with the which is still a challenge. In [8] proposed a method for grouping in large multimedia database that this approach introduces the concept of influence functions and different notions of density groups depending on determination of density type attractor's functions. Also introduce local density function can be determined by using map data oriented representation. The generality of this approach also shows the simulation of a locality-based clustering and hierarchical based partitioning and the invariance of noise and error limits. In [9] proposed a new similarity search approximation based on hash fundamentals. What ever the memory of the data to the probability of collision is high because of what each other is far away . In [10] proposed cleaning a filter for detecting and removing outliers based on a moving window casual data, which is suitable for real time applications such as closed loop control of data. And also proposed optimization guidelines useful simple characterizations based on nominal variation seen in free portions of data outliers. In [11] proposed another algorithm for outlier detection based on distance which says that the main points n with maximum considered outliers D^K, where D^K (p) is distance of the k-th nearest neighbor of p. They used a clustering algorithm to a set of data divided into various groups. Cutting of the batch processing of a different groups can improve the performance of the outlier detection. In [12] proposed the concept of LOF (Local atypical factor), which indicates the degree of outlierness for each data point in a data set. An unusual strong is the one with a high LOF value. The basic idea is to compare the local density ρp of each data point p in a set of data with the local density ρp of each data point p from k-nearest neighbor set *NN*K of **p** (p) in the same set data **P** if local density is lower and its k-nearest neighbors local densities are highest , then the LOF for p is greater (i.e., p is more likely

to be an outlier). His definition of LOF has some advantages: (1) to the data points in a cluster, the LOF value approaches 1, for other data points, the lower and upper limits of their LOF values can be estimated easily, (2) intelligently choose k type of range, (3) his mark discovers lots of ideas detection algorithms based distribution outliers in, and (4) that can detect the data points are outliers relative to the density of their neighborhoods. In [13] Proposes an approach to outlier detection is Known as finding out as a byproduct of the wave Cluster [14], the basic concept of this approach Is that it Eliminates the original groups and then identify identity data outliers is performed by Applying signal processing techniques. In [15] proposed a new method for the detection of other high data which has fixed type of dimension works for exploring projections of lower dimensional locally limited and can not be discovered by applying the brute force method, the combined amount of space, and this mechanism can benefit more easily based outliers that can overcome the influence of the dimensionality. In [16] proposes a new and efficient approach for mining top n local outliers. The basic idea behind this approach is that it avoids the calculation of the local outcrop factors for most objects if n << N, where N is the size of the database. But this approach is a step back in the search of the local typical groups and finds strong local extreme value nested in several levels of granularity. In [17] proposes a clustering algorithm in two phases to detect outliers in the k means algorithm in phase 1 is modified using heuristics if a new input pattern is far from all groups assign it as a new shopping center the group, and the construction of a minimum spanning tree in Phase 2 and remove the longest edge. The tree containing the least number of nodes is selected as outliers. In [18] proposes a new distance-based outlier that look for each point the sum of the distances from its k nearest neighbors called ant those weight value of the largest weight is known as outliers. To calculate these weights are the k nearest neighbors of each point aligning the search space through the Hilbert space filling curve. In [19] proposes the technique for finding outlier from data: cluster-based on the local type of outliers which gives very specific reason for the uncommon data which is locally found in the data and a way for finding the real meaning of the outliers. In [20] proposed a comprehensive method called local correlation for the detection of outliers and the group of outliers. It offers several advantages: (1) providing an automatic cut given to determine if a point is an outlier. (2) can provide loci plot for each point. (3) Method of loci can be calculated as fast as the best previous method (4) carries practically linear approximate method that provides rapid detection atypical. (5) Expanded experiments on

synthetic and real data which shows that this method can naturally detect outliers and micro cluster. In [21] proposed reinforcement outlier detection based local mutual identifying local extreme values in the center instead of detecting local extreme values as noise. They are similar to some groups of objects on one side and are only on the other side of this approach is stable with similar but different user defined relationships. In [22] proposed a technique and approach for the detection of outliers which rely on the distance data streams. In this technique it uses the sliding window type of model in which outlier objection are achieved to uncover defects in the initial window. In these two algorithms it is seen that the first answer objection about the outliers, but require large space. The next algorithm directly comes from the exact but require little memory and gives a close outcome which rely on correct estimations. In [23] proposed a cluster based outlier detection which is based on k median. In this method a more attention is paid to the outliers instead of clustering. For the data stream, more points keep on flowing in. So they eliminate the safe region to free the physical memory for the coming data. But this approach only deals with the numerical data, so a great correction will be implemented for text mining also and on more complex datasets.

CHAPTER 3

A NOVEL APPROACH FOR CLUSTER OUTLIER DETECTION

In modern era there is lot of data mining algorithms which basically focuses on methods of clustering. There are many different types of approaches and techniques performed for outlier detection. Outliers are considered as those data objects which generally do not satisfied with the general and basic behavior of the data model. Many mining algorithms try to reduce the effect of outliers or eliminate them. This may lead to the loss of very valuable and secret information because if there will be a noise from one side could be the signal for another side. In different view, the outliers can be considered as specific importance, especially in the fraud detection cases, where outliers can better be helpful for finding fraudulent activity. So outlier detection and analysis is a striking data mining work which is referred to as outlier mining or outlier detection.

In this dissertation we explore and studied the above mentioned problems and hence introduced a cluster outlier detection algorithm which is capable of detecting the clusters and outliers in a different way for those data sets which hold noise. So in this algorithm the clusters are discovered and managed on the basis of intra-relationship within the clusters and on the basis of inter-relationship between the clusters and the outliers, and vice versa. The whole management, adjustment and correction of the clusters and outliers are done repeatedly before a termination. This new type of data algorithm can be applied and implemented in many areas of signal processing, clustering and pattern recognition.

3.1 Introduction

A large amount of multidimensional data be grouped and need to be analyzed first. Many types of clustering, and outlier detection methods were achieved in recent years. Today many real data sets contain a lot of noise, resulting in poor performance of the algorithms designed and also degrade the efficiency and working.

Cluster analysis process is used to discover similar and groups of separated objects in the sets of data and contribute its major role in data mining area, such as business and science.

Outlier detection is usually distressed with finding the inconsistent role of real and absolute type of objects from data. It is a very major section in the field or area of data mining with many different types of applications, mainly for detecting credit card fraudulent, computer intrusion and uncovering of criminal activities etc. Sometime it is as important as the detection of clusters. Several studies conducted in the detection of outliers. Most previous investigations of this method is in the statistics field [3] these proposed methods typically make inference about the distribution of data, distribution of statistical parameters and the type of outliers. Still, these types of parameters can not be easily decided, and these methods are difficult to apply. In [4] proposed a method based on depth for the detection of outliers. According to the depth value of point data items are arranged in layers in the data space. These are expected to be outliers in the shallower layers. These methods do not have the problem of adapting distributions. But for multidimensional spaces that may not apply. In [5] introduced a method of detecting outliers based on the conclusion after searching for same data and a disturbing element in the data series is found as outliers. This method needs a function which can produce the degree for the dissimilarity of the data set to increase. Looking for the subset of data that leads to reduction in Kolmogorov complexity to the amount of data discarded. In [6] proposed algorithms for detecting outliers of distance -based. O considered a data point in a set of data T, a DB (p, d) - outlier if at least a part p of the data points in T is greater than the distance D from O. Their algorithm based index performs a range search with radius D for each data point. If the number of data points in the D - neighborhood pass a threshold, the search stops and that data point is deleted as an outlier not, otherwise it is an outlier. In the cell approach, in which the entire data space and quantify the data points for the cells. By sniping most numbers of red blood cells that has many data points and contains neighbors, evaluate the process which neglect unusual type of cells and accelerates the detection of outliers. The algorithm of the cell based process is very effective when the total dimension number is decreased to 4 in this type of experiments. For a larger dimensions (≥ 5), the cells growing rapidly for those loops which is included in the paper i.e. cell-based algorithm. Besides the form of identification for the detection of outliers, the same authors (Knorr and Ng, 1999) provides intentional ideas for explaining the exceptional type of outliers. In [7] proposed a guided discovery approach in the search for errors the guide by computed except that the various levels of detail in the data cube. They assessed a value of the cell in the data cube will be the exception and it is suffering or if it is better than the expectation

value based on a statistical model. Thus, the system will be able to deal with hierarchies and dimensions, and the example of this option is to deal with the which is still a challenge. In [8] proposed a method for grouping in large multimedia database that this approach introduces the concept of influence functions and different notions of density groups depending on determination of density type attractor's functions. Also introduce local density function can be determined by using map data oriented representation. The generality of this approach also shows the simulation of a locality-based clustering and hierarchical based partitioning and the invariance of noise and error limits. In [9] proposed a new similarity search approximation based on hash fundamentals. What ever the memory of the data to the probability of collision is high because of what each other is far away . In [10] proposed cleaning a filter for detecting and removing outliers based on a moving window casual data, which is suitable for real time applications such as closed loop control of data. And also proposed optimization guidelines useful simple characterizations based on nominal variation seen in free portions of data outliers. In [11] proposed another algorithm for outlier detection based on distance which says that the main points n with maximum considered outliers D^K, where D^K (p) is distance of the k-th nearest neighbor of p. They used a clustering algorithm to a set of data divided into various groups. Cutting of the batch processing of a different groups can improve the performance of the outlier detection. In [12] proposed the concept of LOF (Local atypical factor), which indicates the degree of outlierness for each data point in a data set. An unusual strong is the one with a high LOF value. The basic idea is to compare the local density ρp of each data point p in a set of data with the local density ρp of each data point p from k- nearest neighbor set *NN*K of **p** (p) in the same set data **P** if local density is lower and its k- nearest neighbors local densities are highest , then the LOF for p is greater (i.e., p is more likely to be an outlier). His definition of LOF has some advantages: (1) to the data points in a cluster, the LOF value approaches 1, for other data points, the lower and upper limits of their LOF values can be estimated easily, (2) intelligently choose k type of range, (3) his mark discovers lots of ideas detection algorithms based distribution outliers in, and (4) that can detect the data points are outliers relative to the density of their neighborhoods. In [13] Proposes an approach to outlier detection is Known as finding out as a byproduct of the wave Cluster [14], the basic concept of this approach Is that it Eliminates the original groups and then identify identity data outliers is performed by Applying signal processing techniques. In [15] proposed a new method for the detection of other high data which

has fixed type of dimension works for exploring projections of lower dimensional locally limited and can not be discovered by applying the brute force method, the combined amount of space, and this mechanism can benefit more easily based outliers that can overcome the influence of the dimensionality. In [16] proposes a new and efficient approach for mining top n local outliers. The basic idea behind this approach is that it avoids the calculation of the local outcrop factors for most objects if n << N, where N is the size of the database. But this approach is a step back in the search of the local typical groups and finds strong local extreme value nested in several levels of granularity. In [17] proposes a clustering algorithm in two phases to detect outliers in the k means algorithm in phase 1 is modified using heuristics if a new input pattern is far from all groups assign it as a new shopping center the group, and the construction of a minimum spanning tree in Phase 2 and remove the longest edge. The tree containing the least number of nodes is selected as outliers. In [18] proposes a new distance-based outlier that look for each point the sum of the distances from its k nearest neighbors called ant those weight value of the largest weight is known as outliers. To calculate these weights are the k nearest neighbors of each point aligning the search space through the Hilbert space filling curve. In [19] proposes the technique for finding outlier from data: cluster-based on the local type of outliers which gives very specific reason for the uncommon data which is locally found in the data and a way for finding the real meaning of the outliers. In [20] proposed a comprehensive method called local correlation for the detection of outliers and the group of outliers. It offers several advantages: (1) providing an automatic cut given to determine if a point is an outlier. (2) can provide loci plot for each point. (3) Method of loci can be calculated as fast as the best previous method (4) carries practically linear approximate method that provides rapid detection atypical. (5) Expanded experiments on synthetic and real data which shows that this method can naturally detect outliers and micro cluster. In [21] proposed reinforcement outlier detection based local mutual identifying local extreme values in the center instead of detecting local extreme values as noise. They are similar to some groups of objects on one side and are only on the other side of this approach is stable with similar but different user defined relationships. In [22] proposed a technique and approach for the detection of outliers which rely on the distance data streams. In this technique it uses the sliding window type of model in which outlier objection are achieved to uncover defects in the initial window. In these two algorithms it is seen that the first answer objection about the outliers, but require large space. The next algorithm directly comes from the

exact but require little memory and gives a close outcome which rely on correct estimations. In [23] proposed a cluster based outlier detection which is based on k median. In this method a more attention is paid to the outliers instead of clustering. For the data stream, more points keep on flowing in. So they eliminate the safe region to free the physical memory for the coming data. But this approach only deals with the numerical data, so a great correction will be implemented for text mining also and on more complex datasets.

The approach used by us is very distinct and new from the above type of clustering approaches, and the methods of detecting outliers in this work we try to detect and arrange cluster and outliers groups on the basis of the relationship in the set of intra groups and a set of outliers and as ratio between the types of clusters and as well as different types of outliers.

There are many reference point for controlling the likeness or unlikeness of the clusters. [24, 25, 26]. ROCK [25] maps the likeness of the two groups by calculating the overall interconnectivity of the two types of clusters in behalf of users. Chameleon [26] it finds the likeness of two types of clusters which is totally rely on a vital type of model sequences. The other two clusters will be joined and merged if the inter connectivity and closeness of the two clusters are strongly attached to the internal inter connectivity and proximity groups of the elements within the clusters.

There are many different approaches have been proposed [27, 28] for calculating the clustering algorithm results. In [28] proposes a method for a group of information about the validity of indexes dispersion medium grade within the groups and the ordinary types of points. The indication takes routing method of compactness and density separation. In [27] proposes a valid function for measuring diffuse overall other ordinary solidity and dissolution of diffuse partition. Total ordinary solidity is assessed by moving of the data points from the center, and separating of the barrier is shown by the distance the centers. Clustering these types measurements assess validity of clustering calculation by the quality of clusters formed.

We investigated that in several different conditions that the clusters and the total no of outliers meanings are related and connected to each other, particularly for the data sets which contain lots of noise. So it is very important to deal with those types of clusters and those types of outliers as a perception of important factor in this field.

Other major drawback in the field of data mining is that clusters and outliers are discovered particularly on the available knowledgeable characteristics of data sets and the outcomes are

referred to background of the original clusters and outliers. In general the background of knowledgeable characteristics of the original well maintained data sets can't be compared with each other and better outcomes are difficult to be achieved even by using dimension reduction method. We will attempt to disclose the clusters and outliers in different manner by relying on the characteristics of the sets of data, and using the connection between the type of those clusters and the types of outliers through a measurable way.

So in this dissertation we explore and studied the above mentioned problems and hence introduced a cluster outlier detection algorithm which is capable of detecting the clusters and outliers in a different way for those sets of data which contains some noise. So in this algorithm the clusters are discovered and managed on the basis of intra-relationship within the clusters and on the basis of inter-relationship between the clusters and the outliers. The whole management, adjustment and correction of the clusters and outliers are done repeatedly before a termination. The next part of the work is classified as follows. Part 3.2 will show the descriptions of the difficulties faced. Part 3.3 will show the total working of about the cluster outlier detection algorithm. Part 3.4 will show the experiments, results and conclusions.

3.2 Problem Description

The basic design of both clusters and outliers are associated to each other. The data of real world consist of all natural cluster for those clusters that are not rare for few cases, actually data objects or data points in the data be a part of the natural cluster. Normally there are outliers in the existing data. One characteristic of the cluster and outliers shown that the amount of diversity. Clusters and outliers are views whose other significations are interrelated hence it is crucial for address those clusters and types of outliers as the meaning in the data mining process. The difficulties of outlier detection from clusters is described as follows. We will propose some codes and descriptions. So we will denote *n* as the total number of data points and *d* will be the dimension of the data space. **A** the input data set to be a *d* dimensional

$$\mathbf{A= \{}\boldsymbol{A_1, A_2,\ldots\ldots, A_n}\mathbf{\}}$$

Which is belonging in the hypercube $[0, 1]^d \subset R^d$. Now there will be a data point $\overrightarrow{A_i}$ is a d-dimensional vector.

(3.2.1) $$\overrightarrow{A_i} = [A_{i1}, A_{i2},\ldots\ldots, A_{id}].$$

So our major objective is to purify and boost the performance of algorithms. In this work, for the reasons of solidity and practical analysis of our method, we developed and modify an algorithm for clustering and make the process of refining and improving their performance and overall result. After all, we show how outlier detection cluster will be able to improve the outcome of other different algorithms in the experiment section. The results of outlier detection, the detection process of the other cluster until a reached termination.

On the basis of the divisional cluster outlier, we execute this type of algorithm repeatedly. For the next step we imagine that the instant cluster is F_c and the instant numbers of outliers is F_o. So we can take the sets of cluster as $C' = \{C', C',...., C'_{Fc}\}$, and outliers as $O' = \{O', O',....., O'_{Fo}\}$. In this we use the name compactness for measuring the cluster quality according to the data points closeness to centroid of the cluster.

3.2.1 Selection of Metric Distances

In this it is very important to describe an appropriate selection of metric distances for data mining problem in particular. We will say d (M_1, M_2) representing the data points distances M_1 and M_2 beneath a fixed metric distances. In a space high-dimensional data are generally scarce, and generally uses distance as an indicator of the Euclidean distance which may not perform as well as whenever there is increase in dimensionality. In [29] work that occurs in high-dimensional nearby neighbor turn into risky, the differences in the distances of the near and far points to a query point will not increase as rapidly as the least of two. In [30, 31] saw the curse of dimensionality from the mark of glimpse of distance measures for normally taken L_k, the drawback is unstable for K value: faster working will be degraded with the rapidly growing of K. So in this type of circumstances, the Manhattan metric (L1 norm) is considered strongly than the Euclidean metric (L2 norm). Research also oppressed the partial metric distances [30] that will overcome the effect of algorithms of clustering. Working on this research work we take $L_{0.1}$ to L_2 metric in this.

3.2.2 Cluster compactness

The type of cluster which is obtained from a data set is a subset in which the points shows a closer relationship with each other regarding the points belongs to the outside of the cluster. From [27, 28], the measurement of relationship is shown by solid compactness and the separation is shown by inter cluster relationship. The word compactness is taken as proportionate name to the object is called as solid as compact in nature with the relation to a flaccid enclosed environment.

For given clusters $C' = \{C', C',....,C'_{Fc}\}$ and outliers $O' = \{O', O',....., O'_{Fo}\}$, compactness (CP) is closest data points measurement in C_i.

(3.2.2) $$CP\ (C_i) = \frac{\sum_{p \in C_i} d(p, k_{C_i})}{|C_i|}$$

In this the centroid of the Cluster $C_{i,}$ is k_{C_i} data point in Cluster $C_{i,}$ is p. $|C_i|$ is the number of data points in C_i, and $d\ (p,\ k_{C_i})$ is the total distance of p and k_{C_i}. The centroid k_{C_i} is the algebraic average for points in the cluster: $k_{C_i} = \frac{\sum_{p \in C_i} P}{|C_i|}$.

This description does not adequately describe the distribution of the data and form the cluster, and doesn't shows the clusters and other clusters relation.

We will point this description which is:

Description: The S is called as the subset of V let MST (S) be the minimum spanning tree of sub graph contains S. The total ICD (S; G, w) is the internal connecting distance of S and the length of the edge of longest MST (S). The ECD (S; G, w) is external connecting distance, and length of the edge of shortest S and V – S. For compactness S is shown as (S; G, w), Now it will be taken as:

(3.2.3) Compactness of subset (S; G, w) = ECD (S; G, w)/ ICD (S; G, w)

3.2.3 Difference of data groups

The name difference is used to show the overall difference among the two clusters, between two extreme outliers and between the cluster and one of the outliers.

The purpose for allocating differences between them is the same for the difficulties of measuring the distance for a objection point from the cluster in the field of mining. From the literature [32] presents an increased range for a objection Q and a C uses a sequence of d_{min} and d_{avr} (the minimum and average distance) but also stated that neither d_{min} or d_{avr} has better rating in a original structure because the cluster don't show same data distribution.

$$d_{avr}(\mathbf{q}, \text{sphere}(\mathbf{c}, r)) = d(\mathbf{q}, \mathbf{c}),$$

(3.2.4)

$$d_{min}(\mathbf{q}, \text{sphere}(\mathbf{c}, r)) = \begin{cases} d(\mathbf{q}, \mathbf{c}) - r \text{ if } d(\mathbf{q}, \mathbf{c}) > r, \\ 0 \text{ otherwise} \end{cases}$$

In this r is the radius of S which is smaller and centered C which covers the data points in C. We will take the density ρ for : $\frac{\rho}{\rho+1}$ for d_{min} and $\frac{\rho}{\rho+1}$ for d_{avr}. The density ρ is as:

$$\rho = \frac{number\ of\ points\ in\ C}{volume\ of\ S} = \frac{number\ of\ points\ in\ C}{r^{logd}}$$

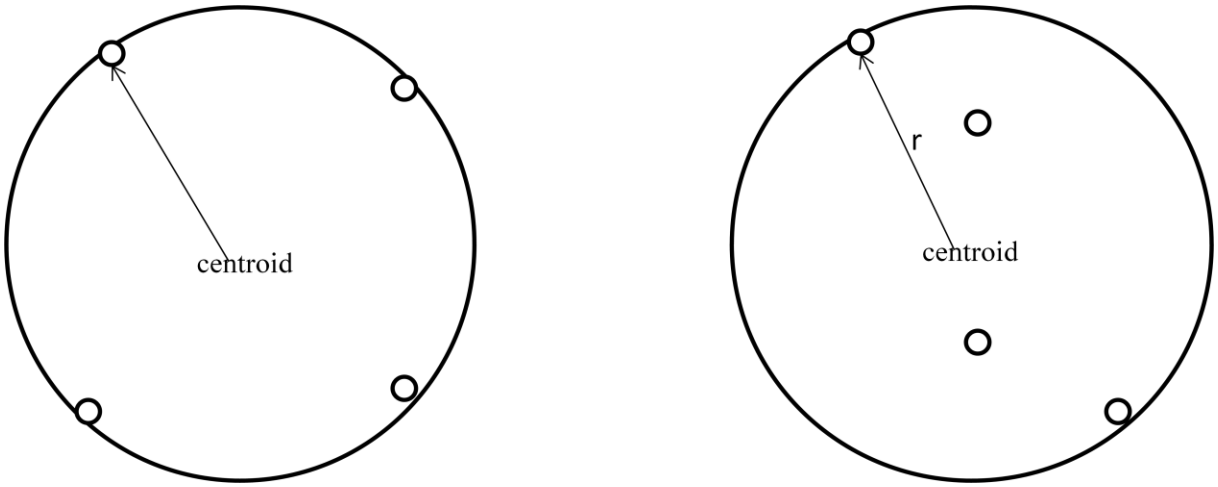


Figure 3.2.1: Two clusters showing different compactness and equal density

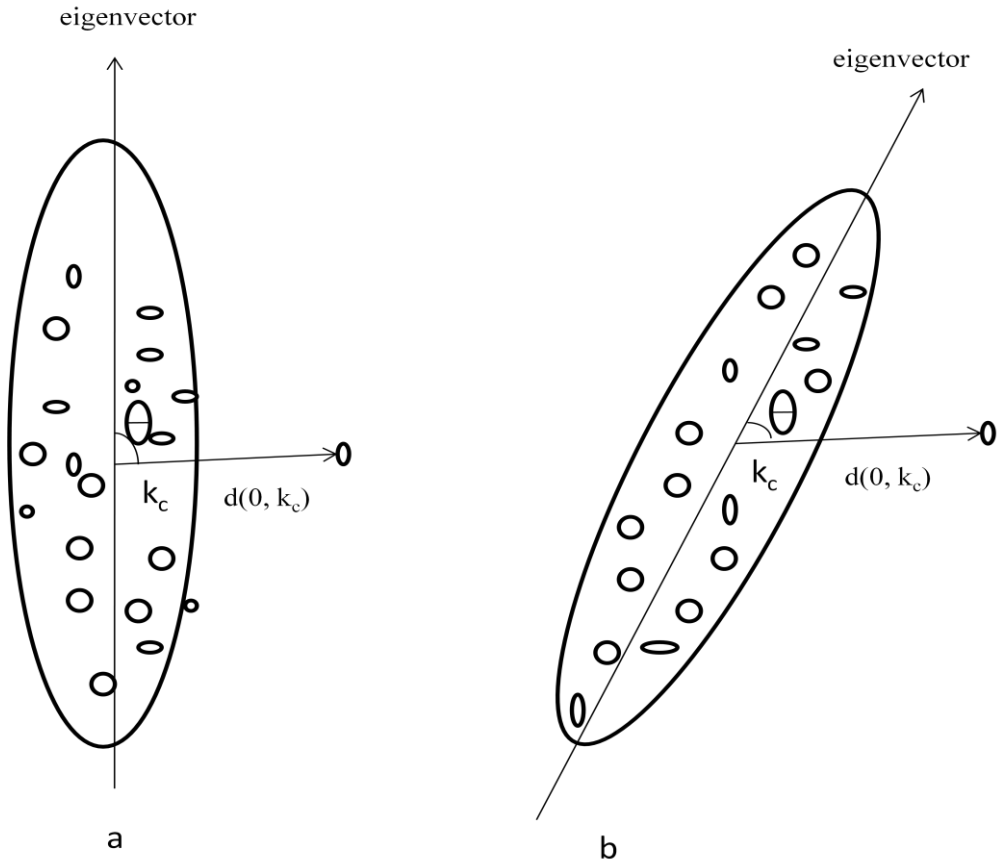


Figure 3.2.2: Two clusters showing different orientation and equal compactness

A definition of the density will not imply such internal design of a cluster of precision. Figure 3.2.1 shows that the two cluster of two dimensional form.

In Figure 3.2.1 density as defined these two clusters have the same density as shown, since it also contains the equal number of data points and the equal r. But by looking at the figure, we can see that the right side cluster is more tight and compact than the cluster on the left.

To solve these problems we can use the concept of compactness (CP) of the cluster rather than using the density for the creation of the weights of distance measurement. We will describe the difference of cluster C and an outlier O is as follows;

(3.2.5) $$\mathcal{D}_1 (C, O) = w_1.d_{min} (O, C) + w_2.d_{avr} (O, C)$$

Where $w_1 = \frac{1}{CP\ (C)+1}$, $w_2 = \frac{CP\ (C)}{CP\ (C)+1}$, d_{avr} (O,C) = d (O, k_c) and d_{min} (O,C) = max (d(O, k_c) – r_{max}, 0) in this r_{max} is the max observed distance.

In some cases this type of definition only help the spherical shapes clusters. Figure 3.2.2 shows that the two ellipse with same r_{max}, and d (O, k_c), d_{avr}(O, C) and d_{min}(O, C) with the outlier O. Therefore, the distances based on the definition above are equal. The orientation of the two cluster is different so that the distances to the outliers groups must also be different.

Actual data typically consist of unlike changes in the data points distribution, and also the true system is not good for distribution of data.

A well known technique is Singular Value Decomposition (SVD) [33] in which data is transformed into different system which minimizes correlations in the data. Thus constructed d*d matrix to derive the eigenvectors matrix defining an orthogonal system for the removal of the second order correlation of the data. We concerned with the achievement of the eigenvector with the highest expansion to capture the best knowledge that discriminates the data points together. Once acquired the eigenvector, we compute eigenvector angle and the vector linked with the centroid of the cluster and outlier O. Can assign the angle of (0, 90). If it is large so that the outlier also be away from the group assuming all other parameters are equal. Dot product is used

for calculation$|cos\theta|$. Consider two vectors $\overrightarrow{V_1}$ and$\overrightarrow{V_2}$, so the dot product can be defined as: $\overrightarrow{V_1} . \overrightarrow{V_2} = |\overrightarrow{V_1}| \, |\overrightarrow{V_2}| \, cos\theta$. So we compute $|cos\theta| = \frac{|\overrightarrow{V_1} . \overrightarrow{V_2}|}{|\overrightarrow{V_1}| \, |\overrightarrow{V_2}|}$.

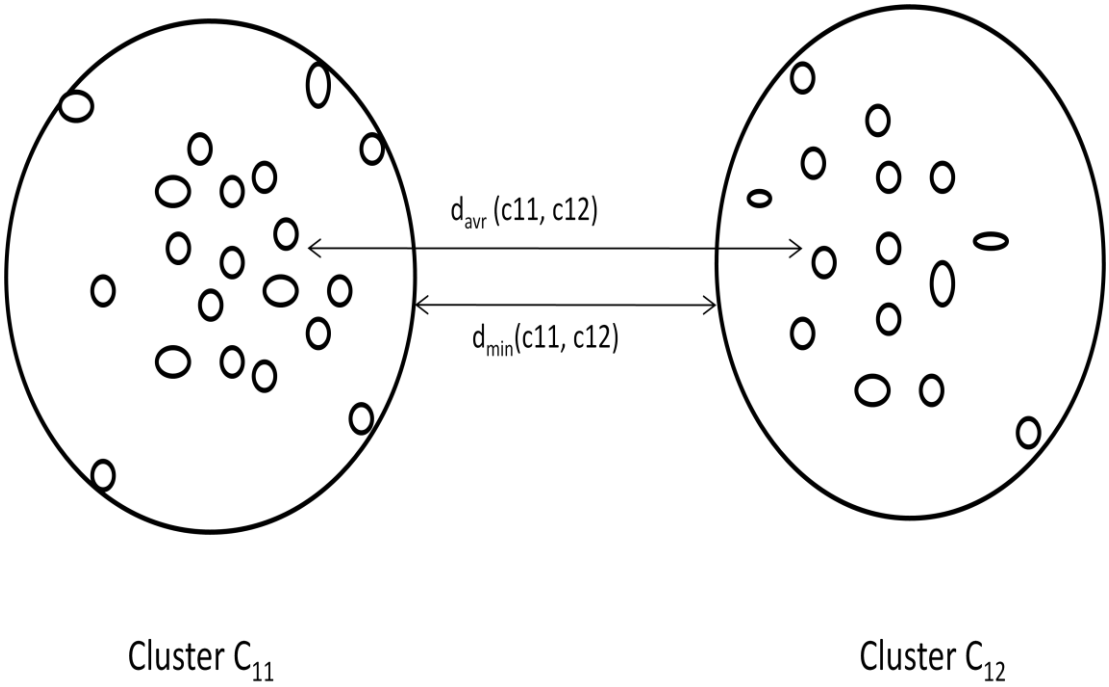
The modified description is.

Description : We say that difference of C and O as:

(3.2.6) $$\mathcal{D}_1(C, O) = \frac{w1.dmin\ (O,C) + w2.davr\ (O,C)}{1 + |cos\theta|}$$

Where $w_1 = \frac{1}{CP\ (C)+1}$, $w_2 = \frac{CP\ (C)}{CP(C)+1}$, d_{avr} (O,C) = d (O, k_c) and d_{min} (O,C) = max (d(O, k_c) – r_{max}, 0)

From the above description the distances for outlier and the cluster in Figure 3.2.2 has large value regarding b.



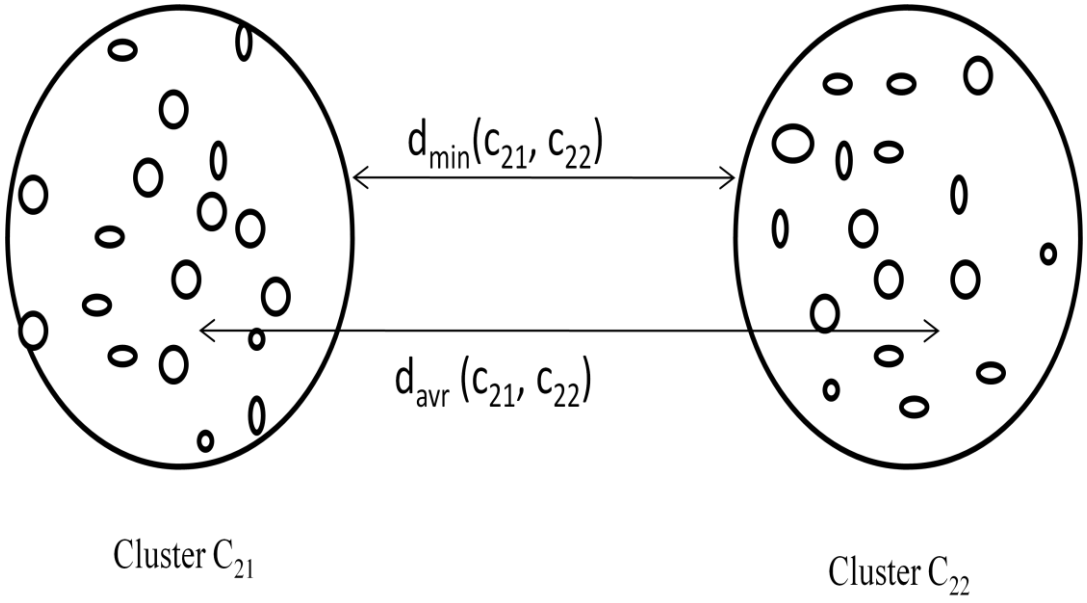


Figure 3.2.3: Two pairs of clusters having different diversity with equal average and minimum distance

Accurate difference may not be always represented by the distance between the two clusters. From figure 3.2.3 the C_{11} and C_{12} are very less than the C_{21} and C_{22}. As we are seeing the first pair (C_{11} and C_{12}) is expected to be joined in bigger. However there is no difference between the minimum distance and the average distance (d_{min} ($C_{11,}$ C_{12}) = d_{min}(C_{21}, C_{22}) and d_{avr}(C_{11}, C_{12}) = d_{avr}(C_{21}, C_{22})).

Some of these are cited in the past work. ROCK [25] proposed the concept of linking of two data points m_i and m_j for measuring the similarity between data points and increases sum of link (m_q, m_r) for data points pair m_q, m_r which belongs to cluster and also reduce the link (m_q, m_s) for m_q, m_s in other types of clusters instantly. For merging of the two clusters C_i and C_j it also defines goodness measure:

(3.2.7) $$g(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{(e_i+e_j)^{1+2f(\theta)} - e_i^{1+2f(\theta)} - e_j^{1+2f(\theta)}}$$

In which C_i, C_j is cluster i and j of size e_i and e_j, θ is the parameter which is used to control the points, f(θ) is function of θ, and link[C_i, C_j] is the number of cross links of the other no of clusters.

According to chameleon [26] shows that the pair of sub more similar groups considering both connectivity among others, and the proximity of groups. The connectivity relationship for a pair of clusters C_i and C_j is defined as the connection for EC_{ci}, C_i (sum of the edges connecting the vertices in vertex in C_i to C_j) for C_i and C_j regarding to the interlinked between inner and EC_{cj} & EC_{ci} (the sum of the edges cluster breakups into two approximately alike portion) of the two groups C_i and C_j:

(3.2.8) $$\text{RIC}\ (C_i, C_j) = \frac{\left|EC_{\{Ci,Cj\}}\right|}{\left|EC_{Ci}\right| + \left|EC_{Cj}\right|/2}$$

And the relative closeness between two groups C_i and C_j is defined as the absolute closeness between C_i and C_j that is normalized with respect to the inner proximity of the two groups C_i and C_j:

(3.2.9) $$\text{RCN}\ (C_i, C_j) = \frac{\bar{S}_{EC_{\{Ci,Cj\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{Ci}} + \frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{Cj}}}$$

Where $\bar{S}_{EC_{Ci}}$ and $\bar{S}_{EC_{Cj}}$ are the edges belonging to the C_i and C_j respectively groups and $\bar{S}_{EC_{\{Ci,Cj\}}}$ is the average weight of the C_i to the vertices in C_j.

However, the concept proposed in the ROCK [25] The suits for data sets of some useful values. Chameleon as in [26] which is on graphic K weighted data set of nearest attribute for each data point in its k nearest neighbors have to be collected, so that the cost of the total calculation time is very high, especially in sets of high dimensional data. And the calculation is also nontrivial bisection.

Here we have a simple measure that incorporates the concept of measuring compactness diversity of two groups, which is useful to show how compact the data points within the cluster.

Description: Difference of two clusters C_1 and C_2 is:

(3.2.10) $$\mathcal{D}_2(C_1, C_2) = \frac{d(C1,C2) * (1 + |cos\theta|)}{CP(C_1) + CP(C_2)}$$

Where d (C_1, C_2) is the average distance of the clusters or the minimum distance between them. This applies to the previous d (k_{c1}, k_{c2}). θ is the corresponding angle for the two eigenvectors. If the value of $\mathcal{D}_2(C_1, C_2)$ be sufficiently larger than the range between C_1 and C_2 groups also will be great.

From figure 3.2.3 ie the clusters C_{11} and C_{12} are selected for merging clusters besides C_{21} and C_{22} therefore that few of the latter.

Description: Difference of two outliers O_1 and O_2 is:

(3.2.11) $$\mathcal{D}_3(O_1, O_2) = d(O_1, O_2)$$

3.2.4 Qualities of data groups

Now, in this section it defines the cluster quality and the quality of outliers.

The overall quality of cluster C is shown not only by the difference between it and other clusters, but also between it and the outliers, it means that how far away from each other. Suppose if C is located to quality outliers will be affected by extreme values are to be away from any grouping. We account for testing of diversity both between groups and diversity between a group and typical case that will define the quality clusters.

Description: The quality of the cluster C is defined as:

(3.2.12) $$Q(C) = \frac{\frac{\sum_{C' \in C, C' \neq C} D_2(C,C')}{F_C - 1} + \frac{\sum_{O \in O} D_1(C,O)}{F_O}}{CP(C)}$$

Each time the large Q (C) is, the quality of the cluster C is better.

Similarly, the outlier quality O is affected not only by the difference between it and the groups but also by the difference among this and other outliers. If the distance will be away from other clusters outliers and then will get the best quality.

Description: The outlier O quality is defined as:

(3.2.13) $$Q(O) = \frac{\sum_{O' \in O, O' \neq O} D_3(O, O')}{F_o - 1} + \frac{\sum_{C \in c} D_1(C, O)}{F_c}$$

The larger Q (O) is, the outlier O will attain the better quality.

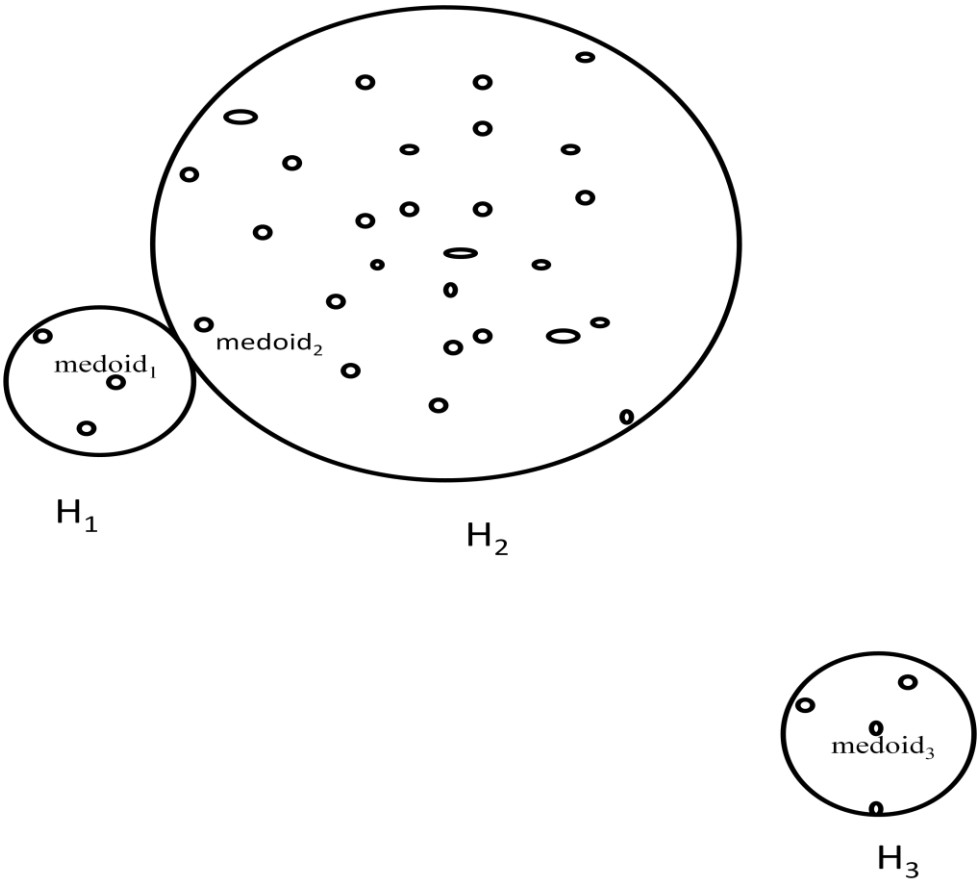


Figure 3.2.4: Association of data groups with different medoids

3.3 Cluster outlier detection algorithm

The main objective of this algorithm is to extract the optimal set of clusters and outliers of input dataset. Clusters and outliers are closely interlinked and also influence each other in some way. in this algorithm the clusters are discovered and managed on the basis of intra-relationship within the clusters and on the basis of inter-relationship between the clusters and the outliers, and vice versa. The whole management, adjustment and correction of the clusters and outliers are done repeatedly before a termination. This new type of data processing algorithm can be applied and implemented in many areas of signal processing, clustering and pattern recognition.

The algorithm is partitioned into two stages: In the first stage, we have to find the cluster's centres and the outliers locations. The difficulty of finding cluster centres has been widely examined. The authors in [34, 35] proposed method for calculating an open condition of an initial refining because it is an effective technique for the distribution modes. So here we apply the greedy technique proposed in [36] This technique attempts to pick up the alternative cluster centers known as medoids. In the second stage must refine and evaluate the cluster and outliers by transferring regularly best of few outliers and some data points from the boundaries of clusters.

Algorithm (K: Number of clusters)

Start

1. First stage

Repeat

A and B are the two equivalent constant to K, where A>B;

RS1 = A. K;

RS2 = B. K;

T_1 = it is the random set having the size of RS1;

T_2 = Finding K medoids from T_1 having size of RS2;

H ← Dispatch the data points (T_2);

C and O ← it is the cluster or the outlier ();

Until $|C| \geq K$

2. Second stage

$C' \leftarrow$ Merging of the Cluster (C');

Repeat

For outlier o $\in O'$ do

Start

Find cluster $\in C'$

Stop

Group of the current number of clusters and outliers in ascending order according to their qualities;

Now exchange the cluster and the outlier ();

O'' it is the set of outliers on the basis of bad quality;

BP is boundary data points according to the bad qualities;

U = O'' $\cup$ BP; outliers unites with boundary point

Until (U gives a constant value or the iteration $\geq \Omega$)

Stop.

3.3.1 First stage

At this stage we need to discover initial medoids set. In later step after finding the medoids, we have to send the data points closest to the medoids and subsets of data forms associated with medoids. Then we have to reach some kind of methods to conclude whether a subset of data is a set or cluster of outliers. We explain other step in detail as follows:

Acquiring medoids

It is crucial to discover medoids that may be near the centers of the various groups of our method. Because commonly outliers exist in the actual data set so which is negative in productive to directly use the greedy approach [36] to find the medoids K [37]. The method proposed in [37], from this, we will discover set of medoids K. Similarly from [37], we have to select a odd set of points T_1 which is set of original data, of size RS1, which is equivalent to the number of necessary data K cluster. Now we must use greedy approach [33] for getting random set of T_1&T_2 which is the size of RS2, which is also equivalent to the K and RS1 > RS2. After

the imposition of the RS2 greedy technique is highly improved efficiency of the algorithm, and there is a high reduction in the growth of the number of outliers produced by the algorithm.

Proc K Medoids: (Data set: D, and no of medoids: K)

Start

$T \leftarrow \emptyset$;

select a data point $medoid_1$ from D in a random way;

$T = T \cup medoid_1$;

Now for each data point $dp \in D - T$ do d' (dp) = d (dp, $medoid_1$);

For i= 2 to K do

Start

Now find $medoid_i \in D - T$ contains the largest d' ($medoid_i$);

$T = T \cup medoid_i$;

Now for each $dp \in D - T$ do

d' (dp) = minimum {d′ (dp), d (dp, $medoid_i$)}

Stop;

Return T

Stop

Dispatch the data points

After getting the set of medoids T_2, we have to check a technique to the conclusion that in some groups medoids are and what are the outliers. First we need to assign each data point for a specific medoids $dp \in T_2$ which is the closest to the dp. After completion each $medoid_i \in T2$ is the related data point.

Proc Dispatch the Data Points: (Initial Medoids set: T_2)

Start

$RS2 \leftarrow T_2$;

i = 1 to RS2 do $H_i \leftarrow \emptyset$;

For dp data point do

Start

Now $medoid_i \in T_2$ which have the smallest distance d (*dp*, $medoid_i$);

$H_i = H_i \cup dp$

Stop;

$H \leftarrow H_1, H_2,....H_{RS2}$;

Return H;

Stop

Data set division

Now, as the set H of the first data set **A**, now we will examine the associated medoid with the subset of data ∈ H. A $medoid_i$ that has the tiny subset of the associated data H_i perhaps one of the others. We have the initial medoids, several medoids are probably in the same natural group, and also contain very little data, but literally belonging to the group apart from outlier. 3.2.4 the figure shows three medoids which is taken at random and which also creates subsets (H_1, H_2 and H_3). The data subset H_1 has the small amount of size as per the figure. According to figure H1 is a part of the natural grouping. Thus the size is not sufficient to decide for medoids is an outlier or a cluster. Now we use some method to set the criteria for determining that a medoid is outlier or cluster. First we need to separate data subsets ∈ H with a certain size threshold t_s is oriented implementation throughout C me. For each subset s in O, suppose medoid consists $medoid_i$. which are taken as the possibility of being an outlier $medoid_i$. If $medoid_i$ creates a small subset of s data, but which is close to the other subsets of data, it is possible that the $medoid_i$ is the data point within the cluster apart from an outlier.

Proc data set division (initial data division set H, cluster size threshold $\mathbf{t_s}$)

Start

C← ∅;

O← ∅;

O← ∅;

Now each set H_i ∈ initial division of data H do

Start

If $|H_i| \geq \mathbf{t_s}$

$C' = C' \cup H$;

Else $O' = O' \cup H$;

Stop;

Now for each set $s \in O'$ do

Start

Find close set $H_j \in H$;

$d_t(s) = D_2(s, H_j)$;

Stop;

Group the subsets $\in O'$ in increasing order by $|s|/d_t(s)$; (where s is subset $\in O'$);

The set of list is l;

Find the upward point p' in l;

Now O' = data points $\in$ subsets before p' in l;

$C' = C' \cup$ data subsets

Return;

Stop

After the completion of the cluster or outlier, it should be unlike the cluster set size $C' < K$ if RS1 and RS2 initial sizes are large enough. If it happens again just run the initial setup to ensure that the overall cluster size is at least K of C'.

3.3.2 Second stage

In this second stage, we must first combine the current set in cluster K. And next step of grouping clusters and outliers in terms of quality and the bad selection of cluster and outliers is done. To check the quality of each group must be computed on the basis of intra- relationship of clusters and clusters and outliers inter relationship. In the next we will achieve some of the methods for the selection of contour data points to the worst qualities of clusters. And in the fourth and final step we will refine and extract the set of clusters and outliers continuously exchanging optimally selected data points limits and worst qualities of extreme values. Therefore steps two, three and four are performed several times just before termination. We explain each step in details as follows:

Merging of the clusters

Before the process of outlier detection first we have to combine the current set $\mathbb{C}$ cluster to cluster K. This process is an iterative one in each iteration stage, the nearest two or more cluster are in close $\mathbb{C}$ that are combined together. According to the measuring range $\mathbb{D}_2$ (C_1, C_2) of the two groups defined in part 3.2 calculate the distance between the groups C_1 and C_2. The iteration will be in the process continuously since the no of clusters in $\mathbb{C}$ is K. Now we have to calculate the centroid $C_i \in \mathbb{C}$.

Merging of cluster (initial set of cluster is $\mathbb{C}$, clusters: K)

Start

While $|\mathbb{C}|$ is greater

Start

Finding the two nearest or closest clusters;

$\mathbb{C} = \mathbb{C} \cup \mathbb{C}$;

$\mathbb{C} = \mathbb{C} - (\mathbb{C})$;

Stop;

Return $\mathbb{C}$

Stop

Grouping of clusters and outliers

Now outlier $\in$ O̓ we have to search closest cluster $\in \mathbb{C}$. According to the difference measurement $\mathbb{D}_1$ (C, O) defined in part 3.2 the distance between the cluster C and outlier O achieved. And according to the outliers in O̓ and the clusters in $\mathbb{C}$ the quality Q (O) (which is defined in the part 3.2) is calculated. The worst qualities of outliers are thrown in set O̓'.

Thus, for each cluster C $\in \mathbb{C}$. On the basis the clusters in $\mathbb{C}$ but also the outliers in O̓ the quality Q (C) (defined in the part 3.2) is calculated. The worst qualities of clusters thrown in set $\mathbb{C}$'.

Finding of boundary points

We will be needing data in the groups that are not only more distant than centroid group, and contain the smallest points such as data points in the clusters to its neighborhood. The remaining circumstances claim that this type of approach promote groups of different standard geometries, such as spherical hyper process. For cluster $C_i \in C'$, we have to find the contour data points according to the distance d_{ij} for data point and d_{pj} centroid C_i C_i to which it is a part of its neighbors $\xi (j)$. $\xi (j)$ is the points within the range of d_{pj} τ and τ is a set proportional to $| C_i |$. We grouped the points in Ci in decreasing order of $d_{ij} / \xi (j)$ and find the initial acute lower p point. The data point of after p are considered as data points C_i border. Hence the data point size of cluster contour C_i is beyond a certain proportion of size, we consider that the cluster has not C_i excellent contour data points and set their contour data points set as $\emptyset$ (null) .

Now we group all data points limits $\in C'$ groups ' in the whole of BP and also see the BP for its sizes. If the size of the BP is very great than the O', then we must decrease the BP size, and I do not think BP sizes and O' have much differences imbalance current partition of A in the considering for result of the current algorithms are certainly some clusters and outliers. So to reduce the size of BP perform other data points in the BP. The former type of points for every group $C_i \in C'$ single and the classification is done by combining of data points limits $\in C'$ for all groups. The basis for classification is different. For this the total no of distance covered for points of data for same boundary and the center of the cluster depending upon their association . This is done because $C' \in$ clusters' have different sizes and the classification is according to the distance calculation which is pushed to the data of the border for some clusters. So after we use $d_{ij} / (| C_i| * \xi (j))$ and the classification is used to remove the result of the different sizes between the groups. Then we will decrease the BP for its accurate size that only contains the |O'| points of data to the border in the list of grouped data points limits.

Exchange of data points

This step is used to exchange outliers and data points of contour features. Now, outlier O in | O'| add in the closest group. For every boundary *bp* in BP data points we changed to a new outlier.

Exchange data points in cluster and outlier

Start

$\mathbb{C}'$= it is clusters set which is containing bad qualities;

O' = it is the outliers which is containing bad qualities ;

Now setting BP for all clusters $\in \mathbb{C}'$;

If BP size is greater than $|O'|$ we will decrease it;

Changing the boundary points in BP into some outliers;

Now merge every outlier in O' in closest cluster to which it belongs'

We will compute quality Q (**A**) of divisional set of data;

Stop

There is a great cause for not changing the data points of the border between clusters is that every division degrades data quality if performed. There is a description below for supporting our conclusion.

Description 3.3.1 Suppose d_{p1} be the limit data point for the C_1 group, and d_{p2} be the limit data point for the cluster C_2. Let the new cluster be C_1^n and C_2^n respectively.

$$Q(C_1^n) + Q(C_2^n) \leq Q(C_1) + Q(C_2).$$

Proof. By the description and treatment partition of the initial dataset be achieved easily, if dp1 be the data point cluster boundaries C_1 and C_1^n is new but includes cluster without d_{p1} and with including d_{p2} then $Q(C_1^n) \leq Q(C1)$ and likewise the same goes for $Q(C_2^n) \leq Q(C2)$. So

$$Q(C_1^n) + Q(C_2^n) \leq Q(C_1) + Q(C_2).$$

Condition for termination of the process: By the completion of the step the states set U of the bad quality of all a typical |O'| and reduced data limits BP set point as $U = |O'| \cup BP$. The iteration is carried for either condition is reached: U $\in$ elements can not be changed more or desperately reaches a threshold of number of iterations.

3.3.3 Analysis of time and space

Now after that all the process has been completed and the total necessary type of work has been accomplished we will calculate the amount of time necessary to achieve the correct working of this algorithm which plays very important role in this dissertation.

Assume the data set size is n. Now, throughout the process we store information on all points of which have collectively the space O (n). For the second stage we require some certain space for configuration information for initial C' clusters and for outliers O' to the contour data points for each group, the bad qualities of the extreme values and clusters number in every performing iteration. So that the amount of space required is O(n). The total no of time needed for every iteration process is

$$O(n + |C'|log|C'| + |O'|log|O'|)$$

So that in particular for the calculation of several types of overall qualities and sorting process C' and O'.

Hence the total amount of computable time necessary for the proper working of this algorithm is

$$O(\Omega *(n + |C'|log|C'| + |O'|log|O'|))$$

in the above calculation we will take Ω as a threshold for number of iterations.

3.3.4 Workflow of the algorithm

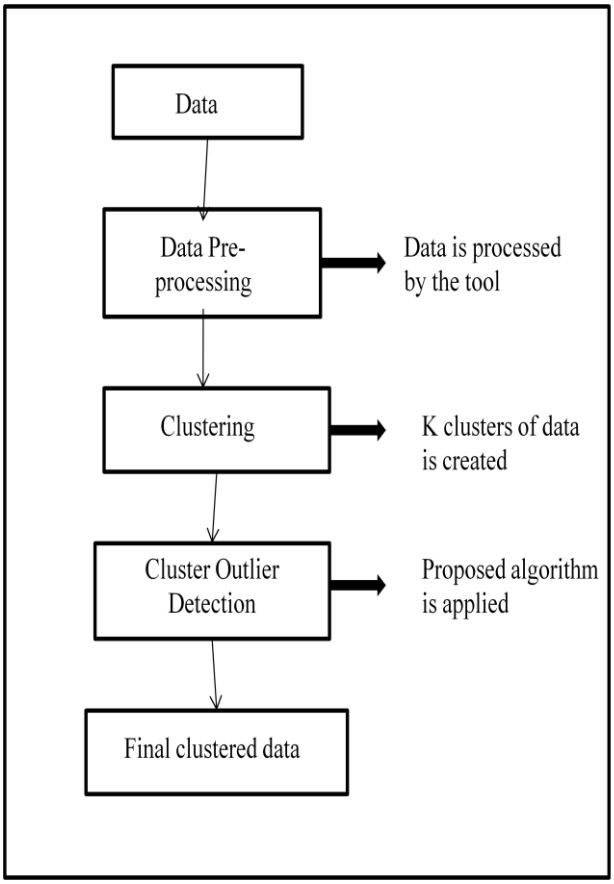


Figure 3.3.4: Workflow of the algorithm

From the above given figure 3.3.4 it shows the workflow chart of our proposed algorithm cluster outlier detection in high dimensional data. From the given figure we observe that data set is taken and loaded into the tool cluster 3.0 for pre-processing of the data and after the completion of the pre-processing step of the data we get our desired cluster after this step K number of clustered data is being created and we will check for the clusters as well as outliers in it after the completion of this step we observe that some of the data points still to be merged with their respective groups. Then we apply our proposed algorithm on the remaining set of data for cluster outlier detection and in this step we get some outliers so we will check the average value for those clusters to which that particular outlier to be merged before doing this step we will finds its quality and the diversity between the outliers as well as the clusters and also check the diversity between the outliers and clusters and check boundary points of the data. When this whole process completes then in the last we get the final clustered data.

CHAPTER 4
EXPERIMENTS

4.1 Experiments and results

We conducted experiments on real datasets to evaluate the effectiveness and accuracy of our proposed approach. All experiments were performed on the data sets of the actual absolute type of applications that are made compared to the different types of algorithms such as Cure.

To calculate the total no of accuracy of our proposed approach we will give some definition used efficiently for the calculation of our approach. The accuracy of the detected cluster is calculated in accordance with the accurate percentage value and the percentage of recall value. Now for the total no of detected and achieved cluster $C(D)$ and the real absolute no of cluster $C(R)$ we will define $C(D)$ accuracy in regard to $C(R)$ as $\frac{|C(D) \cap C(R)|}{C(D)}$ and we define the recall value as $\frac{|C(D) \cap C(R)|}{C(R)}$. Now $C(D)$ is called as good cluster of $C(R)$ for calculating the accuracy and the recall value of $C(D)$ with respect to $C(R)$ are high enough. So if the accurate percentage and the recall value percentage of the detected cluster with respect to the real cluster are high or large enough then we can say that the detected cluster is the corresponding cluster of the real cluster.

4.2 Tool used for experiment

We conducted the experiment at 3.0 clustering tool [41]. The real data set were imported into the tool in the text format. Cluster 3.0[41] The tool was originally developed by Michael Eisen at Stanford University and was later amended and updated by Michiel De Hoon. The cluster 3.0 is a software open source used to implement clustering method it is commonly uses with gene expression. Cluster 3.0[41], basically provides a graphical type of user interface routines that access the cluster. Also available for various operating systems. Cluster 3.0[41] can be run with command line. In this version of the group 3.0 has been modified K means [38] clustering

algorithm and also extends to the organization of the maps include two-dimensional rectangular grids.

4.2.1 Loading, Filtering and adjusting the data

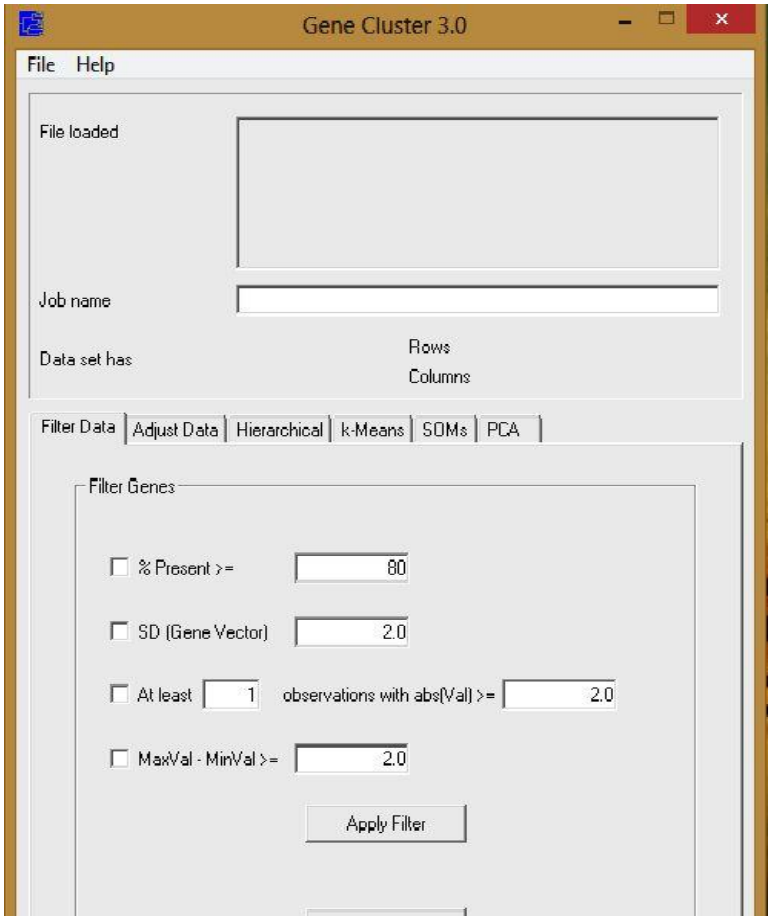


Figure 4.2.1 (a): cluster 3.0 tool [41].

Data can be feed into cluster 3.0 by selection of the load data file from the menu. Many options are available for adjusting and the process of filtering the data. These type of functions can be accessed with the help of filter data option and adjust data tabs options.

- **Loading the data**

  The first step in using the cluster data is imported. Cluster 3.0 only read tab, comma delimited text files in a specific format spaces. This tab, comma delimited text files can be made and taken in any spreadsheet program like MS Excel standard. In group 3.0 the ranks of the input table represent genes and columns represent observation.

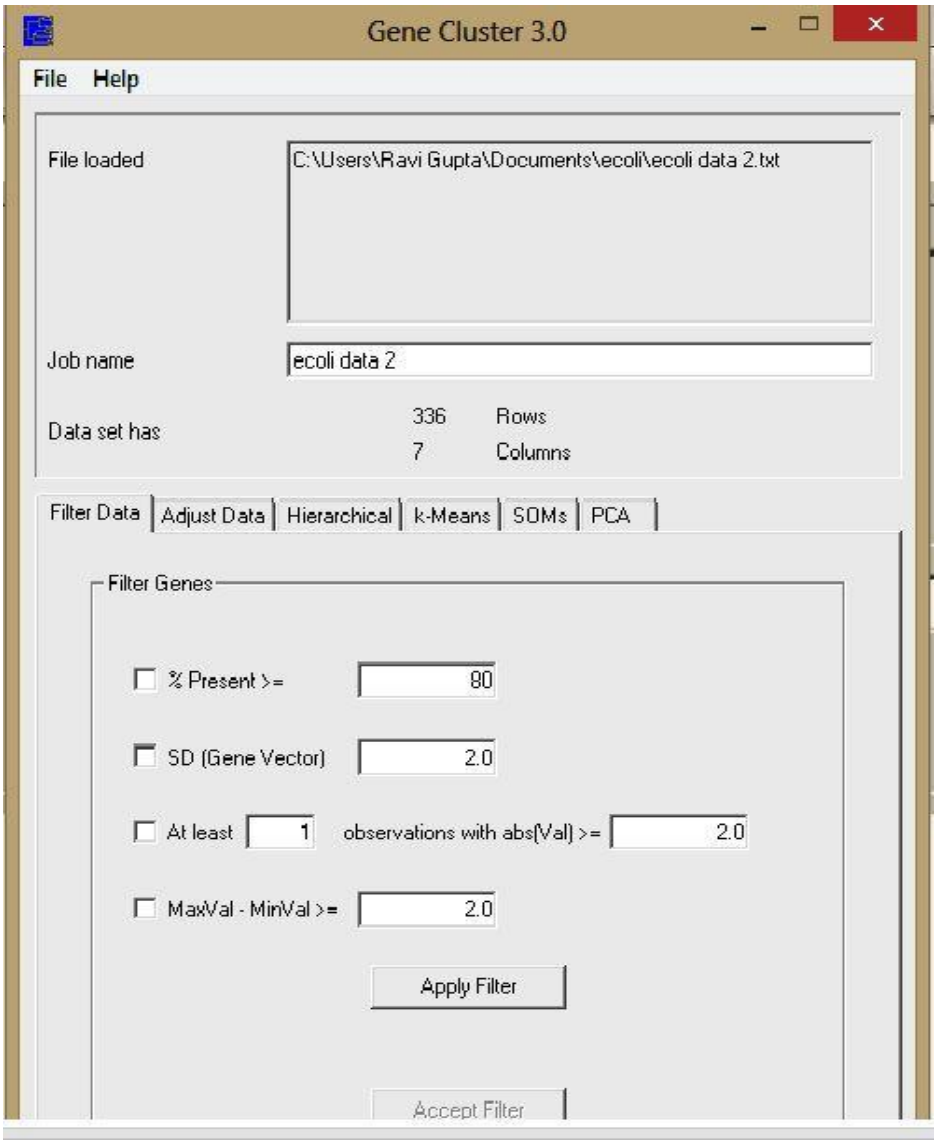


Figure 4.2.1 (b): Loading of data [41].

- **Filtering the data**

The filter data sheet allows us to eliminate genes that have desired properties of the dataset insured. Current properties available that may be used for data filtering are:

This%> = X: This removes all genes having missing values more than (100 - X) percent of the columns.

SD (Gene Vector)> = X: This delete all genes having standard deviations of the noticed values which is smaller than X.

At least X notification with abs (Val)> = Y: This destroys all genes that have small X notifications with absolute values greater than Y.

Max-MinVal > = X: This destroys all genes whose maximum minus minimum values are small than X.

Figure 4.2.1 (c): Filtering of data [41]

- **Adjusting the data**

  In the data sheet adjustments, we can perform a number of operations that alters the hidden data from table. These functions are:

  Log Transform : Used to replace all data values for x log2 (x) .

  Center genes [mean or median] : Subtract the average or median line of the values of each row of data, so that the mean or median value of each row is 0.

  Center matrices [mean or median] subtracting the average column mode or median of the values in each column of data, so that the mean or median result of every column is 0.

  Normalizing genes : used to multiply all values of each row of data by a margin value S so that the sum of the squares of the values corresponding to each row is 1.0. A separate S is taken for each row.

  Normalize arrays : Multiply all values in each column of data by a margin value S so that the sum of the squares of the values in each column is 1.0. A separate S is computed for each column.

4.2.2 K means clustering in the tool

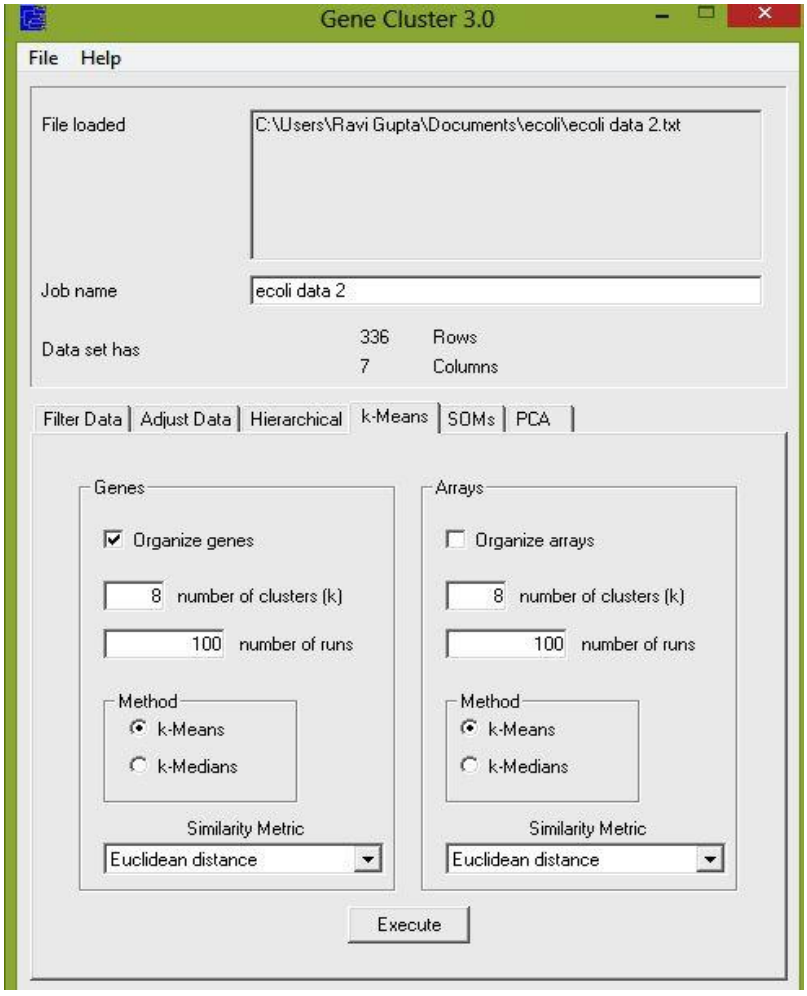


Figure 4.2.2: K means clustering in the tool [41].

The k -means [38] clustering algorithm is a simple and well- known type of cluster analysis. The ultimate goal is that we have to start with a group of elements, such as genes and some of the chosen Cluster number k we want to find. Articles are integers in the first randomly joined the cluster.

The k - means clustering application always starts recurring two-step process, such as:

1. Calculate the mean vector for all elements that belong to each group.

2. Items are reallocated to those groups whose center is closer to that element.

Therefore, initial cluster task is casual and most of the different runs of the algorithm k -means clustering can not produce the same end result clustering. So to deal with this type of problem, the k -means clustering is repeated several times and each time it is started from a different initial

clustering method. The sum of the distances within groups was used to compare the different clustering results. So the result grouping with the lowest sum of the distances within the cluster is assured. The number of runs to be made in particular depends on how problematic it is to find the optimal outcome for the group, which in turn is based on the number of genes included. Therefore, the group shown in the status bar the number of times it has produced the best results. If this will be one clustering key with an even smaller amount of intra cluster distances. Then the clustering algorithm k-means then be repeated again and again with other tests. If the optimal result is achieved in many cases, the result has been achieved is probably has the least amount of intra-group distances. Therefore, we can say that k -means clustering method has achieved full optimal clustering result .

Cluster 3.0 also implements a minor change in the k means clustering is called as k cluster median, in this the median instead of the mean of the elements of the nodes are taken. In general function, it is preferable to use the K mean with the Euclidean distance and K-median with Manhattan distance.

4.3 Description of data set

In this experiment we used the Ecoli dataset is obtained from the machine learning repository UCI [42]. Ecoli dataset contains protein localization sites. It is made of a total number of instances is 336 and each instance contains 7 functions. The Ecoli data set has 8 type of clusters which having total no of range sizes of (143, 77, 52, 35, 20, 5, 2, and 2).

The Ecoli set of data has 8 clusters, $C(R)$ for k (no of clusters) = 0, 1,...,7.

Now we will show how the whole experiment works on the Ecoli data set.

First, the data set is kept in the Excel sheet through which you can watch cluster and the number of outliers. Thus the data set containing a total of 336 instance and each instance contains seven characteristics with their respective classes.

- Ecoli data set before clustering

Figure 4.3: Graph before clustering

As shown in the previous figure Ecoli data before clustering. The entire data set is represented in a graph showing the number of cases in the axis of X and the attribute values on the axis of Y. Through observation we can easily say that there are many clusters and outliers present in the dataset Ecoli. Now the task is to detect the no of clusters and outliers in the sets of data. The entire set of data is provided in the Excel sheet tab and comma delimited space and then the data can be exported to a format. Txt Cluster 3.0 because the tool only reads. Txt file format. And after the data is loaded into the cluster, you must perform all the steps already mentioned in (4.2).

4.4 Experiment on data set

After the whole Ecoli data set loaded into the cluster 3.0, and when we apply the K means clustering of the data and set all the parameters required for the clustering result was obtained from the tool. For example the output obtained from the cluster 3.0 is arranged as follows:

Sequence name	mcg	gvh	lip	chg	aac	alm1	alm2	Cluster
ACKA_ECOLI	0.59	0.49	0.48	0.5	0.52	0.45	0.36	0
ALKH_ECOLI	0.67	0.39	0.48	0.5	0.36	0.38	0.46	0
CYNR_ECOLI	0.51	0.54	0.48	0.5	0.41	0.34	0.43	0
DLDH_ECOLI	0.56	0.51	0.48	0.5	0.34	0.37	0.46	0
FABB_ECOLI	0.65	0.47	0.48	0.5	0.59	0.3	0.4	0
GLPD_ECOLI	0.61	0.45	0.48	0.5	0.48	0.35	0.41	0
METR_ECOLI	0.52	0.57	0.48	0.5	0.42	0.47	0.54	0
PTKB_ECOLI	0.64	0.76	0.48	0.5	0.45	0.35	0.38	0
PTWB_ECOLI	0.57	0.54	0.48	0.5	0.37	0.28	0.33	0
PTWX_ECOLI	0.65	0.55	0.48	0.5	0.34	0.37	0.28	0

Figure 4.4(a): Clustered Ecoli data set

Due to space limitation we are able to show only 10 instances from the cluster 0 so in this way we get the whole 8 cluster and their instances. We arranged all the clustered data in similar manner as shown in the above figure 4.4 (a). And after making cluster of the data set by using cluster 3.0 we will have to check for the detection of cluster and outliers in the data set. After the clusters which are created by the clustering tool cluster 3.0 we have to arrange all the 8 clusters according to the above given figure. The output which is obtained from the tool is in .txt format so we will arrange all the instances and features as well as values of those instances according to the output which was obtained by the tool. The output which was obtained for the 0 cluster belongs to that clusters only and all the instances as well as their features and values should be arranged according to the obtained cluster result and similarly in this way we will arrange all the other 7 clusters.

Now according to the output from the tool we get:

138 instances for the cluster k = 0.

80 instances for the cluster k = 1.

73 instances for the cluster k = 2.

4 instances for the cluster k = 5.

No result for the rest clusters k = 3, 4, 6, 7.

So on the basis of basic conception data are of only 8 clusters Ecoli currently in it. So let's set the parameter number K cluster to 5 to calculate the average accurate value and the recall value retirement respectively.

- Ecoli data set after clustering

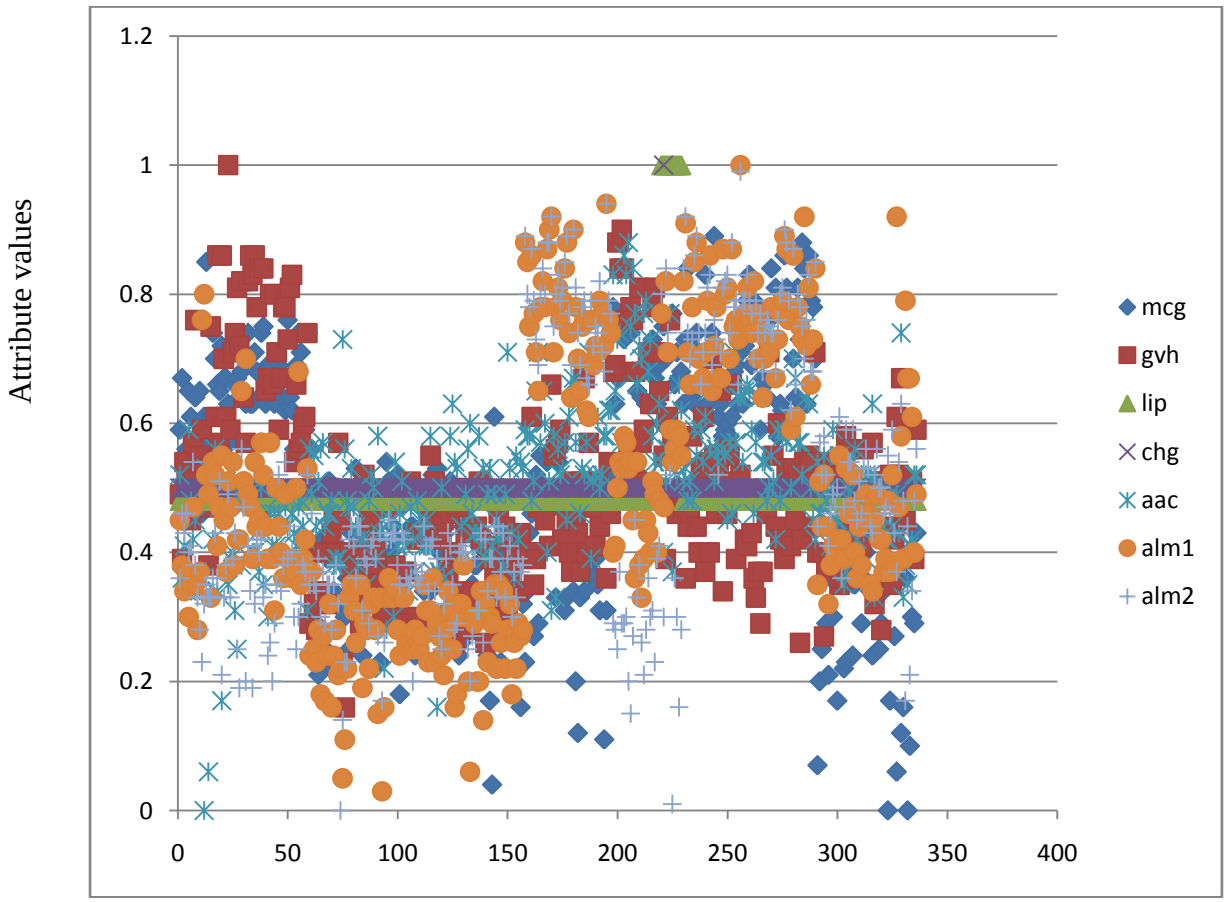


Figure 4.4(b): Ecoli data set after clustering

As we are seeing from the above figure 4.4(b), the sets of Ecoli data after clustering. The whole set of data is represented on a graph which shows the Number of instances on its axis X and the Attribute values on the axis Y. Now after the cluster 3.0 tool implementation on the data set we can easily observe that there are still few data points left which doesn't belong to any of the cluster so from this observation and by applying our proposed approach we will try to detect clusters as well as outliers from the data set.

4.5 Results

We use our proposed algorithm to cluster detection of outliers in the data set Ecoli. The results obtained after applying the algorithm to the data set shows some changes in the grouped data and was different from the above, as a result after application of the cluster tool in the data set that is not capable of groups correctly detect and outliers. Now, when we imposed our proposed approach proposed in the pooled data is able to detect those groups and outliers that were left by the tool during the clustering process .

Our proposed approach to the detection of other cluster is completely rely on the algorithm as explained in detail in the previous section of work (Section 3.3). First medoids suitable for natural or real groups are examined and after the initial groups were detected and outliers and the no of clusters and the no of outliers are adjusted and filtered in accordance with the link between them.

The experiment was taken on the data set for Ecoli cluster detection and outlier. Ecoli data set is used to find the site of localization of proteins. The data set contains a total of 336 cases (objects) that each attribute (1 Name and 7 input functions). Now in accordance with the reality of data set that contains 8 groups Ecoli that i.e. number of clusters k = 0, 1, ... , 7. Although the three groups in the data set are too small to neglect and we set the parameter of the number of clusters in K-5 .

Now our algorithm will be performed on the data set Ecoli. 4.5.1 table shows the result of clustering of our proposed algorithm and after the experiment we observed that our proposed algorithm has the greatest amount of information about the first three larger clusters. This means that it produces most of the information about the first three largest clusters given input data set after applying the algorithm on it.

	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7
$C'(R)$	143	77	52	35	20	5	2	2
$C'(D)$	138	80	73	N	N	4	N	N
$C'(D) \cap C'(R)$	132	54	50	N	N	4	N	N
Accuracy %	95.65	67.50	68.49	N	N	100	N	N
Recall value %	92.3	70.12	96.15	N	N	80	N	N

Table 4.5.1: Result of the algorithm for Ecoli data

Now from the result table we see data set contains only 8 clusters $C'(R)$ for K = (0 to 7). There are three clusters observed which are small in size so we take clusters parameter K to 5. And accuracy of the detected cluster judged on the basis of the accuracy % and the recall value %. Now for $C'(D)$ detected cluster and $C'(R)$ for the real cluster we count the accuracy of $C'(D)$ with respect to $C'(R)$ as $\frac{C'(D) \cap C'(R)}{C'(D)}$ and the recall value is $\frac{C'(D) \cap C'(R)}{C'(R)}$. Hence $C'(D)$ is called as good cluster regarding of $C'(R)$ if the accuracy and the recall value of $C'(D)$ and $C'(R)$ are high.

Now from the given table 4.5.1: we observe that there are 8 clusters, for real cluster $C'(R)$ for k = (0 to 7) and it contains the instances values given (0 = 143, 1 = 77, 2 = 52, 3 = 35, 4 = 20, 5 = 5, 6 = 2, 7 = 2). And for detected cluster $C'(D)$ the instances values obtained (0 = 138, 1 = 80, 2 = 73, 3 = N, 4 = N, 5 = 4, 6 = N, 7 = N). And for $C'(D) \cap C'(R)$ the instances values obtained (0 = 132, 1 = 54, 2 = 50, 3 = N, 4 = N, 5 = 4, 6 = N, 7 = N).

The accurate value of $C'(D)$ with respect to $C'(R)$ is defined as $\frac{C'(D) \cap C'(R)}{C'(D)}$ and we obtain the accurate value for the cluster k = 0 is 95.65%, k= 1 is 67.50%, k= 2 is 68.49%, and k= 5 is 100% and for clusters k = 3, 4, 6, and 7 is N (not available).

Similarly the recall value of $C'(D)$ with respect to $C'(R)$ is defined as $\frac{C'(D) \cap C'(R)}{C'(R)}$ and we obtain the recall value for the cluster k = 0 is 92.3%, k = 1 is 70.12%, k = 2 is 96.15% and for clusters k = 3, 4, 6, and 7 is N (not available).

The average accurate value is 77.21% (which is the total average no of the accurate value of the result from table 4.5.1: 95.65, 67.50, and 68.49). The average recall value is 86.19% (average recall of the clustering is from the table 4.5.1: 92.30, 70.12, and 96.15).

We can clearly observe that from the result table that our proposed algorithm contains information mostly about the first three largest natural or real clusters.

- Ecoli data set after cluster outlier detection algorithm

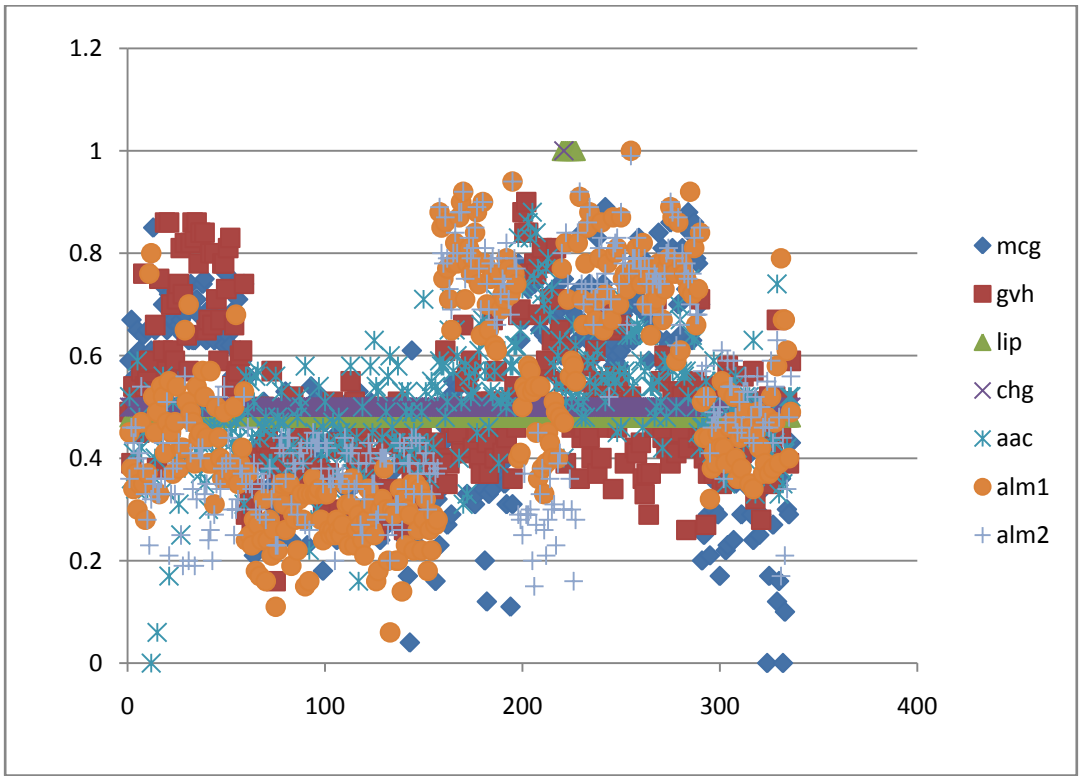


Figure 4.5: Ecoli data set after cluster outlier detection algorithm

From the above given figure 4.5 we can observe that after imposing our proposed algorithm work on the set of Ecoli data the remaining clusters and outliers were arranged on the relationship. All the left outliers were easily detected by our algorithm but few data points still to be merged with some of the clusters as well as outliers but as the ground truth that every real world data sets have some outliers or noise so we cannot remove all the outliers completely. But our proposed algorithm has managed to detect the clusters and as well as outliers at its best possible way.

4.6 Comparison with existing algorithms

We compare our algorithm cluster outlier detection with other existing algorithm to show its efficiency and effectiveness. In this step we include the same data set for its comparison with existing algorithm and also perform the same experiment using the tool cluster 3.0.

We used the implementation of COID algorithm [40]. COID[40] is used to compare our algorithm applies different parameter values widely and approved the best results according to the group . This algorithm makes a medoid for some groups, and sets of those type of clusters and the no of outliers are searched and then these are modified, adjusts and refined which is based on the relationship between them. Applying this algorithm dataset Ecoli. We created the cluster K to 5 according to the field reality Ecoli data set because we already have three small groups in the data set so that the mean value and the exact value depending on the parameter set retirement .

When COID[40] algorithm is carried out on the proposed Ecoli data we observe. 4.6.1 table shows the result of clustering algorithm COID[40] and after the experiment shows that the existing algorithm also has the largest amount of information about the first three larger groups.

From the comparison part we can easily check that the total no of values and their outcomes when applied properly on the basis of algorithm work and it will give the overall scenario that how this algorithm will work in the coming future and the absolute world applications.

	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7
$C'(R)$	143	77	52	35	20	5	2	2
$C'(D)$	135	85	68	N	N	4	N	N
$C'(D) \cap C'(R)$	128	52	45	N	N	4	N	N
Accuracy %	94.81	61.17	66.17	N	N	100	N	N
Recall value %	89.51	67.53	86.53	N	N	80	N	N

Table 4.6.1: Clustering result of COID algorithm

Now from the result table we see data set contains 8 clusters $C'(R)$ for K = (0 to 7). There are three clusters which have the least information so we will set clusters parameter number K to 5. And the total accuracy value of the detected cluster calculated on the accuracy % and the recall value %. Now for $C'(D)$ for the total no of detected cluster and $C'(R)$ for the real cluster we calculate the accuracy of $C'(D)$ with respect to $C'(R)$ as $\frac{C'(D) \cap C'(R)}{C'(D)}$ and the recall value is $\frac{C'(D) \cap C'(R)}{C'(R)}$. Hence $C'(D)$ is called as good cluster regarding of $C'(R)$ if the accuracy and the recall value of $C'(D)$ and $C'(R)$ are high.

Now from the given table 4.6.1: we observe that there are 8 clusters, for real cluster $C'(R)$ for k = (0 to 7) and it contains the instances values given (0 = 143, 1 = 77, 2 = 52, 3 = 35, 4 = 20, 5 = 5, 6 = 2, 7 = 2). And for detected cluster $C'(D)$ the instances values obtained (0 = 135, 1 = 85, 2 = 68, 3 = N, 4 = N, 5 = 4, 6 = N, 7 = N). And for $C'(D) \cap C'(R)$ the instances values obtained (0 = 128, 1 = 52, 2 = 45, 3 = N, 4 = N, 5 = 4, 6 = N, 7 = N). The accurate value of $C'(D)$ with respect to $C'(R)$ is defined as $\frac{C'(D) \cap C'(R)}{C'(D)}$ and we obtain the accurate value for the cluster k = 0 is 94.81%,

k= 1 is 61.17%, k= 2 is 66.17%, and k= 5 is 100% and for clusters k = 3, 4, 6, and 7 is N (not available).

Similarly the recall value of $C'(D)$ with respect to $C'(R)$ is defined as $\frac{C'(D) \cap C'(R)}{C'(R)}$ and we obtain the recall value for the cluster k = 0 is 89.51%, k = 1 is 67.53%, k = 2 is 86.53% and for clusters k = 3, 4, 6, and 7 is N (not available).

Now, to calculate the average accurate value and the recall value we will consider only the first three groups because the algorithm has more information about the first three clusters. The average accurate value is 74.05% (the total no of accurate value of the result of Table 4.6.1: 94.81, 61.17, and 66.17). The average recall value is 81.19% (as a result of the average recall of the clustering table comes 5.6.1: 89.51, 67.53 and 86.53).

Therefore we have the comparison between our proposal and the existing algorithm and then to calculate the exact value and the average value of recovery is observed that the average accurate value and the recall value of our proposal is much more than the existing working algorithm process.

4.7 Conclusion and analysis

Now in this achieved work, we later introduced a novel approach for cluster outlier detection in high dimensional data. This approach is used to increase the total no of clusters and the total no of outliers qualities for those high dimensional type of data which contains noise. in this algorithm the clusters are discovered and managed on the basis of intra-relationship within the clusters and on the basis of inter-relationship between the clusters and the outliers, and vice versa. The whole management, adjustment and correction of the clusters and outliers are done repeatedly before a termination. This new type of algorithm working process can be applied and implemented in many areas of signal processing, clustering and pattern recognition.

We also worry about washing the difficulties of the deficiency of the correspondence between the reality on the ground of the actual data and characteristics obtainable.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

This part will show our achieved work and also presents the upcoming future working process. In the chapter we had studied about the cluster outlier detection in high-dimensional type of data. in this working algorithm the clusters are discovered and managed on the basis of intra-relationship within the clusters and on the basis of inter-relationship between the clusters and the outliers, and vice versa. The whole management, process of adjustment and correction of the clusters and outliers are done repeatedly before a process of termination. So in the coming future we will try to develop our work regarding the cluster outlier detection in subspace.

5.1 Conclusion

The rapid growth of huge data has far exceeded the limits of our human ability to understand apart from capable tools. It is literally necessary to develop tools which are used to uncover the important information which is established in the very large data. So this dissertation mainly target on efficient and effective mining search of novel, exceptional and meaningful design obtained from the sets of real data. We had study the drawbacks of the cluster outlier detection in high dimensional data types.

In modern era there is lot of data mining algorithms which basically focuses on methods of clustering. There are also different types of approaches and techniques created for outlier detection. Outliers are considered as those data objects which generally do not satisfied with the general and basic behavior of the data model. In several different conditions we had investigated that the clusters and outliers meanings are related and connected, commonly for the noisy type of sets of data. So it is important to deal with those type of clusters and those types of outliers as a perception of the same important factor in analysis field of data.

So on the these types of drawbacks in chapter 3 we introduced a novel approach for cluster outlier detection in high dimensional data and hence introduced a cluster outlier detection algorithm which is capable of detecting the clusters and outliers in a different way for those data sets which contains some noise. So in this algorithm the clusters are discovered and managed on

the basis of intra-relationship for clusters and on the basis of inter-relationship for the clusters and the outliers, and vice versa. The whole management, process of adjustment and correction of the clusters and outliers are done repeatedly before a termination. Then we examine a new procedure for determining the distances for the two clusters, two outliers and the distance of the cluster and the outlier. We implement some new formulas to describe and define the clusters and outliers qualities as well as the whole data set division.

We implemented our proposal on the real world dataset. We show that gradually how outliers are exchanged and some of the boundary points groupings for the best qualities of both outliers and clusters. Therefore, it is more the groups and outliers as a concept of the same importance in the analysis. We also worry about washing the difficulties of the deficiency of the correspondence between the reality on the ground of the actual data and characteristics obtainable.

5.2 Future Scope

In this work we study the problem of exceptional cluster outlier detection in high-dimensional types of data. So in the near upcoming future we will study further and develop our this work towards the detection of cluster outlier in subspace.

In chapter 3, we introduced and investigated a novel approach for cluster outlier detection in high dimensional data and hence introduced a cluster outlier detection algorithm which is capable of detecting those clusters as well as those outliers in a different form for sets of data which contains some noise. So it is important to deal with clusters and outliers as a perception of the same important factor in data analysis field. So we proposed a novel approach for cluster outlier detection in high dimensional data. Although cluster approaches to outlier detection are not always effective and efficient provided they are exercised in full data space. Therefore, for a given set of high-dimensional spaces not all of them is relevant to the same. The data that have unnecessary type of worthless features also creates a large decrement in the precision of a few working algorithms process. So we develop new type of cluster outlier detection approach in subspace likely to investigate the approach outlier detection in subspace cluster. So in this new work, each group will links with dimensions of subsets and each outlier is also linked to its particular dimensions subsets. So we will search and study few instant irregular types of sets of cluster and outliers. Now, depending according to the regular instant sets regularly going to boost

the results obtained from the detection of the other cluster. So every step continuous can adjust the degree of compaction for each group, and bounded subsets of separation and handled for intra relationships of the those types of clusters and for the relationships between the cluster and type of outliers. The size and quality of each outcrop subset is linked to is set and managed which depends on the intra relationship for outliers, and the inter relationships for that type of clusters and also for the those types of outliers .

The main overall objective of the new work is to extract best sets of good type of clusters and also for the good type of outliers for the sets of data set in the subspaces associated to cluster and outliers. As we discover the initial set of cluster as well as initial outliers sets, we can find the total no of subspace captured by these two which is linked to all clusters and outliers.

CHAPTER 6
REFERENCES

[1] Jiawei Han, Micheline Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2001.

[2] A. k. Jain, M. N. Murthy, and P. J. Flynn. Data clustering: A review. ACM Computing Surveys, 31 (3), 1999.

[3] Barnett V, Lewis T (1994) Outliers in statistical data (3rd edition). Wiley, New York

[4] Ruts I, Rousseeuw PJ. Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis 23: 153-168.1999.

[5] Arning, A., Agrawal et al. A linear method for deviation detection in large databases. In proceedings of KDD' 96, Portland, USA, pages. 164-169, 1996

[6] Knorr E, Ng R (1998) Algorithms for mining distance-based outliers in large datasets. In the Proceedings of the 24th VLDB conference, New York, August, pages 392-403. 1998

[7] Sarawagi S, Agrawal R, Megiddo N. Discovery-driven exploration of OLAP data cubes. In Proceedings of the sixth international conference on extending database technology (EDBT), Valencia, Spain, 1998.

[8] Hinneburg, A. and Keim, D. (1998). An efficient approach to clustering in large multimedia databases with noise, Proceedings of 4th International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 58–65.

[9] Aristides Gionis et al. Similarity search in high dimensions via hashing. In the proceeding of VLDB 1999, pages 518-528, 1999.

[10] P. H. Menold. Online outlier detection and removal. In the proceedings of the 7th Mediterranean Conference on Control and automation(MED99) Haifa, Israel, June, pages 28-30, 1999.

[11] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In Proceedings of the ACM SIGMOD CONFERENCE on management of data, Dallas, TEXAS, 16-18 May, pages 427-438.2000

[12] MM Breuning et al. LOF: identifying density based local outliers. In proceedings of the ACM SIGMOID CONFERENCE on management of data, Dallas, TEXAS, 16-18 may, pages 93-104, 2000

[13] D. Yu, G. Sheikholeslami, and A. Zhang. Findout: Finding outliers in very large datasets. The Knowledge and Information Systems (KAIS), (4), October 2000.

[14] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wave cluster: A multi-resolution clustering approach for very large spatial databases. In Proceedings of the 24th International Conference on Very Large Data Bases, 1998.

[15] Charu C. Aggarwal and Philip S. Yu. Outlier detection for high dimensional data. In SIGMOD Conference, 2001.

[16] W. Jin, A. Tung and J. Ha. Mining top n local outliers in large database. In the proceedings of KDD, pages 293-298, 2001.

[17] M.F. Jiang et al. Two phase clustering process for outlier detection. Pattern Recognition Letters 22 pages 691-700. 2001.

[18] Angiuli et al. Fast outlier detection in high dimensional spaces. In the proceedings of KDD. 2002.

[19] Zengyou He, Xiaofei Xu, Shengchun Deng. Discovering Cluster Based Local Outliers, Pattern Recognition Letters. 2003.

[20] Papadimitriou et al. LOCI: fast outlier detection using the local correlation integral, Data Engineering, 2003. In the Proceedings of the 19th International Conference on 5-8 march, pages 315-326, 2003.

[21] J. Xu Yu et al. Finding centric local outliers in categorical/numerical spaces. Knowledge and information Systems March 2006, Volume 9, Issue 3, pages 309-338. 2006.

[22] Angiulli, F. and Fassetti, F. Detecting Distance-based Outliers in Streams of Data. In Proc. of the Sixteenth ACM Conf. on information and Knowledge Management (Lisbon, Portugal, November 2007).CIKM ’07.

[23] Parneeta Dhaliwal, MPS Bhatia and P Bansal. A cluster based approach for outlier detection in dynamic data streams, Journal of computing, volume 2, issue 2, February 2010.

[24] Pavel Berkhin, Survey of Clustering Data mining techniques, Accrue software, San Jose, CA, 2002.

[25] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In the Proceedings of the IEEE Conference on Data Engineering, 1999.

[26] George Karypis, Euo-Hong Han, and V Kumar NEWS. Chameleon: Hierarchical clustering using dynamic modelling. Computer, 32(8):68-75, 1999.

[27] Chi-Farn Chen, Jyh-Ming Lee. The validity measurement of Fuzzy C-Means Classifier for Remotely Sensed Images. In the Proceedings of ACRS 2001 - 22nd Asian Conference on Remote Sensing, 2001.

[28] Maria Halkidi, Michalis Vazirgiannis. A data set oriented Approach for clustering Algorithm Selection. In the PKDD, 2001.

[29] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In International Conference on Database Theory 99, pages 217–235, Jerusalem, Israel, 1999.

[30] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. Lecture Notes in Computer Science, 1973, 2001.

[31] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In The VLDB Journal, pages 506–515, 2000.

[32] Dantong Yu and Aidong Zhang. ClusterTree: Integration of Cluster Representation and Nearest Neighbor Search for Large Datasets with High Dimensionality. IEEE Transactions on Knowledge and Data Engineering (TKDE), 14(3), May/June 2003.

[33] K.V. Ravi Kanth, Divyakant Agrawal, and Ambuj Singh. Dimensionality reduction for similarity searching in dynamic databases. In Proceedings of the ACM SIGMOD CONFERENCE on Management of Data, pages 166–176, Seattle, WA, 1998.

[34] Usama Fayyad, Cory Reina and P.S. Bradley. Initialization of the iterative refinement clustering algorithms. In the proceedings of fourth International Conference on Knowledge Discovery and Data Mining, pages 194-198, New York, August 1998.

[35] Paul S. Bradley and Usama M. Fayyad. Refining initial points for K Means clustering. In the proceedings of 15th International conference on machine learning, pages 91-99. Morgan Kaufmann, San Francisco, CA, 1998.

[36] T. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:311-322, 1985.

[37] Charu C. Aggarwal, C. Procopiuc, J.L. Wolf, P. Yu, and J.S. Park. Fast algorithms for projected clustering. In Proceedings of the ACM SIGMOD CONFERENCE on Management of Data, pages 61–72, Philadelphia, PA, 1999.

[38] J. MacQueen. Some methods for classification and analysis of multivariate observations. In the Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, 1967.

[39] M ester et al. A density based algorithm for discovering clusters in large spatial Databases with noise. In the proceedings of 2nd International Conference on KDD, page 226 - 231, 1996.

[40] Yong Shi. Cluster outlier iterative detection approach to multidimensional data analysis. Knowledge and Information system, volume 28, Issue 3, pages 709 – 733, Springer, 2011

[41] Michiel de Hoon. [http://bonsai.ims.u-tokyo.ac.jp], Open source clustering software. Institute of medical Science University of Tokyo.

[42] The UCI KDD Archive [http://Kdd.ics.uci.edu]. University of California, Irvine, Department of Information and Computer Science.