Unsupervised Learning Method Based on Partitioning in Data Mining

Kelechi Churchill Onyejiaka

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
May 2015
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Serhan Çiftçioğlu
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazım Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Ersin Kuset Bodur
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Asst. Prof. Dr. Arif Akkeleş
3. Asst. Prof. Dr. Ersin Kuset Bodur

ABSTRACT

This study provides the introduction of some basic definitions about clustering method of data mining. For this purpose, it is given the methods of data mining, some algorithms of clustering method. Meanwhile, the k-Means clustering and Hierarchical clustering algorithms are defined.

The aim of this study is to cluster the dataset into two clusters using Hierarchical clustering algorithm and k-Means algorithm. In order to achieve our target, two distance formulas are used to measure the distance between the vectors in the algorithms: the Euclidean distance and k-Nearest neighborhood distance.to compare two methods.

Keywords: Data mining, data mining algorithms, data mining applications

ÖZ

Bu çalışma veri madenciliği kümeleme yönteminin bazı temel tanımlarını sunar. Bu amaçla, veri madenciliği yöntemleri, veri madenciliğinin bazı kümeleme yöntemleri algoritmaları veriliyor. Bunun yanında, K-ortalama ve Hiyerarşik kümeleme algoritmaları tanımlanır.

Bu çalışmanın amacı, Hirerarşik ve K-ortalama algoritmalarını kullanıp veri kümesini iki kümeye ayırmaktır. Amacımıza ulaşmak için, vektörler arasındaki uzaklığı ölçmek için iki tane tanım kullanılır: Öklit uzaklık ve en yakın K komşu bağıntıları.

Anahtar kelimeler: Veri madenciliği teknikleri, veri madenciliği algorimaları, veri madenciliği uygulamaları

To my family

ACKNOWLEDGMENT

I would like to thank the most high for His mercies and guidance during the period of writing this thesis and my studied.

I would also like to recognize the input of my supervisor, Asst. Prof. Dr. Ersin Kuset Bodur who has shown me the best of guidance and have also made remarkable input to my thesis in particular and to my studies in general. Her patience, inspiration and passion throughout the period of writing this thesis have been overwhelming, her in-depth understanding of the concept of data mining has been invaluable to my research and I will forever see her as an outstanding mentor.

I would also like to thank all my friends and colleagues whom are too numerous to mention, Miss Nora Anaso and Mr George Ike most especially and my academics guidance counsellor Prof. Dr. Rashad Aliyev, who was there during my most trying period and all the academic and non-academic staff of EMU for through their remarkable effort and perseverance that the Eastern Mediterranean University has attained the heights it has attained.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Data mining is a process of extracting new ideas or knowledge from raw data. Because of the big data involve nowadays, it makes the scientists and engineers to look into big data transformation and processing model in different aspects of fields like medicine, engineering, physical, and biomedical sciences etc. The big data transformation and data processing model in data mining show us how information can be extracted from data mining and survey, how people attention can be modeled, safety and privacy concerns [1].

The data mining process can assist the organizations to carry out their planning in a data mining project and get quicker results from their data [3]. The problem of clustering is to collect the similar elements of data into the same groups. Clustering is an unsupervised learning tool used in data mining to determine a structure in collection of untagged data. We have different types of clustering techniques such as k-Means clustering, hierarchical Clustering and partition clustering, [4].

Data mining can be also used in cloud computing to gain new ideas or knowledge over the computing resources on the internet service to retrieve useful information from the database server. The applying of data mining techniques and applications in cloud computing will assist the users to extract useful information from a practical integrated data warehouse that minimizes the cost of storing data and infrastructure

[2]. Data mining is also used in CRISP-DM (Crossing –Industry Standard for Data mining). The CRISP-DM is an industry,-tool and application-neutral model. The model helps the organization to achieve their structure necessary to realize better and urgent outcomes from data mining. CRISP-DM is carried out in six stages in data mining which are data preparation, modelling, evaluation, business understanding, data understanding and deployment, [3].

In data mining today, because of large data incurred in our observations and the evolution of technology we have, data mining is used to extract knowledge from data or prediction of hidden information from data. Data mining was introduced in 1990s and it has found in three fields of areas which are machine learning, artificial intelligence and classical statistics.

The classical statistics is used to study data or get valued information from data and it is one of the techniques in data mining to establish or build a pattern, for example: regression analysis, cluster analysis, standard distribution, discriminant analysis, and standard variance.

Artificial intelligence is another techniques used in stimulating the human intelligence or human thought process in statistical problems. The artificial intelligence can be used in a heuristic to oppose the use of statistics and also adopted by some of high-end commercial products such as query optimization modules for relational database management systems.

Machine learning is the intersection of artificial intelligence and classical statistics and considered to be the change of artificial intelligence. The use of machine

learning allows the computer program to study the data, and makes use of the quality of studied data to make better decisions based on the statistical original resolution and ensuring the use of artificial intelligence heuristics and algorithms to achieve its targets. Data mining was introduced or brought out because of the hidden knowledge discovery from large or huge data sets. This is the different kinds of data analysis tools to discover patterns and relationships that many be used to make valid predictions.

In recent years, the data mining was seen or essential to a great data analysis tools used to discovered knowledge and to enhance or increase revenue and reduced cost. Because of the knowledge discovery from data, data mining was a misnomer and it was called ‘data mining’ because of its popularity than to use knowledge mined from data. Data mining can have slight different or same idea with the data pattern analysis, knowledge extraction, knowledge mining from data and archaeology and dredging. It can be used as a KDD or knowledge discovery from data.

These are some steps to discover the knowledge from data such as:
Data Cleaning is used to avoid unused or unwanted data from the large or huge data. Data integration is known as the process of bringing different data sources together as a schema or information. Data selection is bringing out the important or essential data for the analysis from the database. Data transformation is the essential part where data are being transformed into essential form by taking final operations.

Data mining is the critical part where best methods are applied to explore knowledge from data or data patterns.

Pattern evaluation is also the important part to identify the appropriate pattern use for knowledge discovery based on some essential measures.

Knowledge presentation is the technique used to represent graphically or visualized representation of knowledge mine from data to show the user the knowledge discovery from data (KDD).

Recently, data mining has become more famous than the usual term knowledge extraction from data because of weather forecast, industries uses, business enterprise, stock exchange, and media etc. Because of the popularity of data mining today, it has adopted a broad view of functionality in knowledge discovery. Data mining has been a means of discovering attractive or essential knowledge or data patterns from huge amounts of data stored in data warehouse, database and information repositories.

Meanwhile, the data warehouse, database, spreadsheets or other information sources are kept for retrieved or fetched from the database or data warehouse server depends on the user's data pattern request or knowledge mine from data.

Database or data warehouse server is a place where all data are kept for retrieve or fetch from the database or data warehouse server depends on the user's data pattern request or knowledge mine from data.

Data mining engine is the important part of the data mining system and it has a set of functional modules task which are classification, prediction, association, cluster analysis and evolution analysis. Knowledge base is the dominion idea normally used

to control the process or analysis at the knowledge discovery from data. Example of actual knowledge are extra interesting's challenges or thresholds and metadata.

Graphical User Interface is an important part in the architecture of the data mining system; it helps the user to communicate with system by allowing the users to retrieve data from database query and bringing information to assist the search and performing the intermediate data mining results. The graphical user's interface helps the users to look into directory of data and data warehouse architectures or data frames; analysis knowledge mined data patterns and visualized the patterns in the different ways or forms.

The kind of knowledge discovery from data can be used in applications of telecommunication and stock broker, market analysis and relationship, to production control and science exploration. Because of the recent development, the found changes in the database which are data collection, database and information technology have been revolved from file processing to a better and delicate database system.

This query can be viewed as read-only as for the efficient methods for On-line transaction processing (OLAP) have much contributed to the changes and overall receipt of relational technology as a significant tool for resourceful storage, retrieval and management of huge amounts of data, [17].

In Data Mining Task, we have two types of data mining task which are descriptive data mining task and predictive data mining task. Predictive data mining task uses some variables to predict unknown or future values of other variables.
And, descriptive data mining task uses human-interpretable patterns that describe the data. Types of data mining techniques are Classification, Clustering, Predication, Association methods and neural network.

Classification is the method of discovering the model that recount and distinguishes data classes in order to make use of the model to predict the class of objects that the classes label is uncertain. The derived model depends on the evaluation of the set of training data and it is presented in the various forms such as neural networks, decision tress, classification (IF-THEN) rules and mathematical formulae.

Classification is used in the set of records at big amount. Examples of classification applications are credit risk and fraud detection etc. It is used to estimate the correct classification rules by using the classification test data and when the correctness is adopted the rules is used to the new data tuples.

Clustering can be defined as the recognition of related classes of objects. This can be used in identifying compressed and spare regions in object area and to find out the general distribution pattern and relations among data features. Classification is supervised training data while clustering is an unsupervised training data and because of that the clustering is a pre-processing method for attribute subset selection and classification.

Types of clustering methods can be categorized as:

a. Partition method
b. Hierarchical method
c. Density based method
d. Grid-based method
e. Model-based method

The prediction can be used for regression analysis. Regression is defined as the way of modelling the relationship between one or more dependent and independent variables. Today data mining, independent variables are normally known attributes of class labels and dependent variables are normally what we used to predict.

For example, stock prices, sales volumes and product failures rates are very hard to predict unless using the complex techniques that uses on the multiple predictor variables.

Regression methods can be gathered as:

a. Linear regression
b. Multivariate linear regression
c. Nonlinear regression
d. Multivariate nonlinear regression
e. Logistic regression

Association rule is defined as the method of finding frequent item set in discovering among the big data sets. It assists in the business or organizations to identify items that are bought together by sufficiently many customers and make certain decisions.

Association rule algorithms is used normally to make rules with confidence values to a reduced amount of one and the amount of association rules for a given dataset is large and have a minor value of great proportion of rules.

Types of association rule are

a. Multidimensional association
b. Multilevel association rule
c. Qualitative association rule

Neural Network is defined as a set of linked of input and output components and each linkage has a weight present on it. The neural network is used to adjust the weights during the learning process to predict the exact class labels of the input tuples and to mine patterns and to find trends that are many applications in industries and real world business difficulties and best of finding patterns that is good for prediction.

Types of Neural network are defined as the following, [21]

a. Back propagation
b. Multilayer perception
c. Recurrent Neural Networks
d. Clustering Algorithms and self-organizing
e. Wavelet Neutral Network
f. Radial Basis function Network.

Normally, the cluster analysis is illustrated on the bases of field of data mining; where data mining is define as the knowledge extraction or knowledge discovery from data (KDD), non-important, data patterns in large collections of data.

Cluster analysis is divided into a useful groups or meaningful data. If an important data can be achieved, the cluster analysis should be captured the resulting "created or natural" structure of data. Cluster analysis has been used to group related data that have closer functionality, documents for browsing, machine learning, marketing, business organization, data mining and whether forecast. For example, in business today, it also assists the companies to find out the definite groups to identify their customers buying pattern in their various products and also uses in some of the universities to group students for admission according to academics scores in their post exams. We have typical pattern clustering activities which are: Data abstraction, Assessment of output, Clustering, Pattern representation.

Definition of pattern proximity measures appropriate to the data domain. Clustering analysis can also be investigative discovery process which can be used to find out pattern in data without giving an interpretation [23].

There are two major aspects of cluster analysis which are; Clustering and Clustering validation. Clustering is the partitioning of objects into clusters according to its basis or standard for the data set. In clustering analysis, the applications of clustering can be achieved in different ways because of the large number of clustering algorithms have been found [22], [23], [20].

Cluster Validation is the quality estimate process of clustering results that will assist to fit the best clustering algorithms for the data problem. Clustering is already known as unsupervised classification task [23].

In clustering, the most significant problem is to divide the data into clusters for which the data points in the same cluster is similar than the data points in the other cluster by given basis or standard. We have two classes of clustering techniques which are partition clustering techniques and hierarchical clustering techniques.

Partitioning method is the process of allocating the data points into k clusters and iteratively reapportions data points to enhance the quality of clustering results. Partitioning clustering algorithms which uses these algorithms to assign the data into k clusters are k-Means, CLARANS, k-Medoids, Pam and CLARA. Hierarchical method is the process of allocating the data points in diagram structure groups. It can be classified into two categories which are agglomerative and divisive clustering.

The Problems of Clustering Analysis are given in [24] as:

i. It is the time wasting to assess the quality of clustering results in a large database using statistics –based methods.
ii. It is the ineffective of clustering algorithms on working with arbitrary shaped data distributing of data.

Features of Good Clustering are given as the following

i. The intra-class/cluster: The closer of a set of objects in the same group or cluster is high.
ii. The inter–class/cluster: This is the closeness of the set of objects in the same group or cluster is low.
iii. The value of clustering relies on both the similarity measure usually for the method and its implementation.

Some Applications of Clustering: Clustering has wide applications which are Image Processing, Economic Science, Pattern Recognition, and Spatial Data Analysis.

This study has five chapters, as well as these five chapters of this study are ordered as follows. Chapter 1 is the review part of our study. Basic definitions and related concepts are presented in Chapter 2. Hierarchical clustering algorithm and its application are given in Chapter 3. Finally, k-Means algorithm and the data which is solved by k-Means algorithm are presented in Chapter 4. Chapter 5 consists of Conclusion.

Chapter 2

REVIEW OF CLUSTERING CONCEPT

2.1 Definition of Clustering Concept

Clustering is defined as the subset of objects which are “similar or closer” or it can be defined as the process of sectioning a set of data or objects into a set of important or meaningful sub-classes. Clustering is always unsupervised learning and finds “natural” grouping of cases given un-labelled data. Clustering is the means of attaching objects of same into a group (Cluster) of different distribution of patterns in the data set. Clustering is mostly used in data mining process to initialize new clusters (groups) for objects of same distribution of patterns.

To resolve clustering difficulty is by separating a set of data into clusters of similar data points in a cluster to the one of different clusters of different data points. For example, in a supermarket, we have different people that come into the store to buy goods (items) the need, for instances, we can have ten customers that like buying milk, five customers like buying chocolate and sweet, eight customers like buying perfumes and deodorants, all buy goods every week and we can group the customers accordingly to their goods the buy every week into each cluster.

The clustering process could cluster the customers of same or similar buying patterns into the same cluster. Clustering process has brought a significant approach in clustering of objects of patterns into reasonable groups which shows the similarity

and dissimilarity of data points of data set in such a way that gives a useful conclusion of result.

Clustering can be used in many fields like medicine, pattern recognition, Biology, Artificial intelligent and engineering etc. and also can be named different names in different terms of many useful fields (Theodoridis and Koutroubas, 1999). For instance, in the clustering procedure or process, it shows that there are no cases that will use to indicate the actual relations of the valid relationship of data and there are no predefined classes among the data to be clustered (Berry and Linoff, 1996).

Clustering can be defined as unsupervised learning because there are no predefined classes and no example or experienced to cluster object. On the other hand, classification is a means of making a way of giving a data set to a predefined set of classes. And clustering gives initial class in which our significant of our data set are grouped during the classifying process.

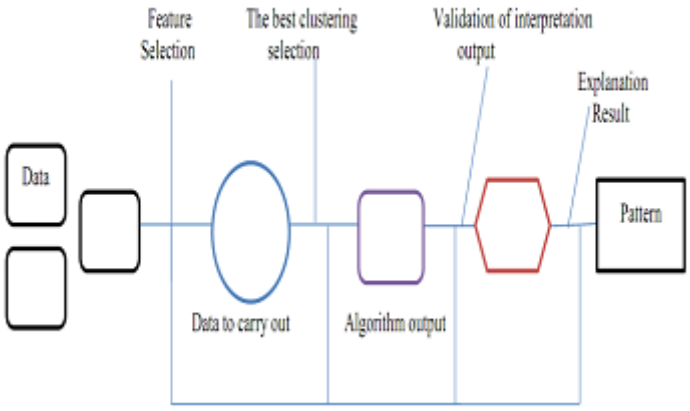

Figure 2.1: Steps of clustering process

In clustering process, different distributions of patterns are been clusters in the data set depend on the actual clustering criterion is being used. For instance, we can be

told to clusters living things of frog, eagle (bird), goat, shark, crocodile, and wheel according to environment of their living.

We cluster frog and crocodile as same group because both live in land and water, cluster goat, eagle, sheep as another group of same cluster while shark and wheel as another group of same cluster. In clustering process, we have the basic steps which are the following, [7]:

Feature Selection comprises of features that can be selected adequately for which the clustering can be actually implemented so as to bring more useful ideas or information in the task of our view.

Clustering Algorithm, this is to select the best clustering algorithm for the data set to be clustered. Using the best clustering algorithm will give a useful result for our data set. And also, proximity measure and clustering criterion are the best tools to characterize the clustering algorithm to bring out the useful pattern that suitable for the data set.

Proximity Measure is the measure that determine the value of ‘’closeness’’ of data points in the data set. In the proximity measure, we have to ensure that all the features are been carried out in computation and no data points is better than the other.

Clustering Criterion depends on the explanation the specialist or expert wants to give a term to represent the type of clusters considering the data set. For instance,

elongated cluster may be useful according to another while compact cluster of data points in the R-dimensional space may be useful according to another.

Validation of the Results tests the results of clustering algorithm is proved by using actual criteria and techniques to determine the correct result of the data set used.

The normally use the clustering algorithms to cluster objects that are unknown despite the clustering methods applied and it is required in the final separating of data set in some of the applications to emphasize clusters.

In this case, according to the interpretation of the results, the experts in a particular field have to explain the clustering results with other experiment prove and analysis in a situation to have a useful conclusion from the result.

2.2 Some Applications of Clustering

Data Reduction is the process by which the data are being reduced in the large data set. Nowadays, because of the large data that we have in the database, the process becomes necessary for the data set to be implemented. The clustering can be used to separate the data into many useful clusters than to process the data as a single cluster.

Hypothesis Generation is a way by which the clustering process is used to deduce hypotheses from the interesting data being used. For example, we can find out from the data in airplane and we have two categories of passengers that board plane base on their age and days the normally travel. And then, from the data, we deduce the hypotheses that the old people board airplane much on weekdays while the young

people board airplane much on weekend. It shows the data can be clustered from the information deduce from the data set.

Hypothesis testing is used to prove of the actual hypotheses being processed in the data set. For instance, we have the old people that board airplane much on weekdays. To verify if the hypothesis is true, we can apply the representative set of airplanes, and each airplane is represented by passengers' information (the day's board by passengers and the age). By clustering analysis, it observed that the old people that board much airplane on the weekdays are determined; it means that the hypothesis is verified by cluster analysis.

Prediction Based on Groups is the cluster analysis that determines the result of the cluster by characterized the attributes of the patterns that clusters. In this case, the unknowing patterns can be clustered by knowing the similarity and the cluster concept to the clusters attributes.

Clustering can be used in Biology to group the genes that have same similarity and also in Ecology and also, indicate a useful insight structure of population. Also in business, clustering analysis is important in business aspect and it assists the marketers to gain useful information from the customers' database which assist to cluster the buying patterns of customers.

One of the types of clustering can be applied to web mining to gain useful information from the group of documents on the large amount of documents that is clustered similarly in the web.

Spatial Data Analysis is the type of clustering application that the use to classify and gain a useful information and patterns that occurred in the spatial databases. It uses normally in Geographical information systems, image database exploration and satellite images because of large amount of spatial data occur in the database, [6].

Clustering analysis is taken to be the most significant type of unsupervised learning [7]. In cluster, objects are grouped according to their similarity in nature and differences in distance objects are grouped according to different clusters. Objects of the same cluster is not normally group according to their similarity measure but they are cluster (group) according to their strength of illustrative ideas which the objects of the same cluster has a common concept to all objects.

Euclidean distance measure and Mahalanobis distance measure are two main distance measures usually used to measure the similarity between two data points in a data set while Kullback-Leibler is used to measure the dissimilarity distance of data point in a data set, [8].

Clustering can be used in direct communication to form a cluster in a device that is close to another device in order to communicate with each other, and also, it assists in the sharing of radio resources for devices to have to and fro in a base station and providing a mixed network for devices. The clustering can be used to achieve good performance in communication when resolving the cellular network and interference limited system, [9]. Clustering is most significant in word categories of similar words or context. Clustering of word categories can be evaluated by using context similarity to cluster words of same similarity.

In grammar rules of the word categories, by knowing the grammar categories can assist the grammar rules to generate a useful way of information from the cluster [10].

2.3 Definition of Similarity Measure

This is the process of determining the similarity of two documents. Similarity measure can be used in text documents and clustering to show the similarity of two documents having the corresponding feature value. A document is a vector which each constituent or element represents the value of the feature in the document. The worth of the feature can be represented by frequency (the number of repeating of a term in a document). In text document, it has played a significant role in knowledge discovery, web search and gain more useful information from text processing, [11], [12] and [13]. Every word model is generally been used in text processing [14], [15], while relative term frequency is the ratio between the number of appearing term in a document and the total number of appearing terms in the document [16].

If the size of vector is big then, the corresponding vector is sparse. There are different methods of computing the similarity of two vectors which are Kullback-Leibler divergence, Euclidean distance measure, Manhattan distance measure, Cosine similarity, Bray-Curtis dissimilarity measure and Jaccard coefficient. It is a symmetric measure that usually used for computing the similarity measure, its importance is to consider the present and absent of a feature in a document than the difference between the two values in the two documents. The similarity will be rise when there is difference between two worth of a feature in the document decrease while the similarity will decrease when the number of present and absent features

increases. When there is absent value of a feature in the document and there will be no similarity.

It is useful and efficient to use in many text applications to measure the similarity like multi-label classification-means in clustering, single-label classification, and hierarchical agglomerative clustering, [17]. The Euclidean distance measure is the square root difference between the corresponding coordinates.

Let $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ be arbitrary two vectors in R^n.

The Euclidean distance is defined by

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{2.1}$$

Manhattan distance measure is the distance between two points measure along axes at right angles

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2}$$

Cosine similarity is a measure of similarity between given two vectors u and v in R^n (n-dimensional space) defined by inner product and magnitude of the given vectors that measures the cosine of the angle between u and v.

$$S_{cos\theta} = \frac{u \cdot v}{\|u\| \|v\|} \tag{2.3}$$

Jaccard coefficient is a statistic used for comparing the similarity and diversity of sample sets which is given by

$$J_{\mu}(A, B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} \ . \tag{2.4}$$

The Euclidean distance measure is the most usually used in the calculation between the distances from one point to another in a cluster.

For instance, when having the vectors A, B and C in different clusters in Table 2.1, all distance calculations are done using Manhattan distance in order to cluster the given vectors.

Table 2.1: Vectors of the example

	x	y
A	7	8
B	3	5
C	2	9
D	6	1

This is the calculation of the distance from one vector to another to compare the similarity of vectors in order to cluster them. In fact, the equation 2.2 is used to calculate the distance between the vectors.

Step 1 First of all, the distance between each of the vectors is calculated as the following calculations to find the lowest distance of the vectors.

$d(A,A)=0$, $d(A,B)=|7-3|+|8-5|=4+3=7$ and

$d(C,D)=|2-6|+|9-1|=4+8=12$.

All distance calculations are shown in the Table 2.2 and the minimum distance is 5 between the vectors B,C so, we suppose to know that those vectors are in the same cluster.

Table 2.2: Step 1 of the example

	A	B	C	D
A	0	7	6	8
B	7	0	5	7
C	6	5	0	12
D	8	7	12	0

Step 2 At the end of step 1, we have the clusters, $C_1 = \{\mathrm{B},\mathrm{C}\}$, $C_2 = \{A\}$ and $C_3 = \{D\}$, we continue to find the distance between these clusters. Then, Table 2.3 is obtained by the following calculations:

$$d(C_1, A) = \min\{d(B, A), d(C, A)\} = min\{7, 6\} = 6$$

$$d(C_1, D) = \min\{d(B, D), d(C, D)\} = min\{7, 12\} = 7$$

Table 2.3: Distance table of step 2

	$C_1 = \{A\}$	$C_2 = \{B, C\}$	$C_3 = \{D\}$
$C_1 = \{A\}$	0	6	8
$C_2 = \{B, C\}$	6	0	7
$C_3 = \{D\}$	8	7	0

Now, we join two clusters C_1 and C_2 because of the minimum distance. Then we update the cluster elements so new clusters are $C_1 = \{A, B, C\}$ and $C_2 = \{D\}$. In figure 2.2, the dendrogram for above calculation is shown, there are four clusters at level 1, at level two the number of clusters is 3 and finally the number of clusters is reduced to two at step three.

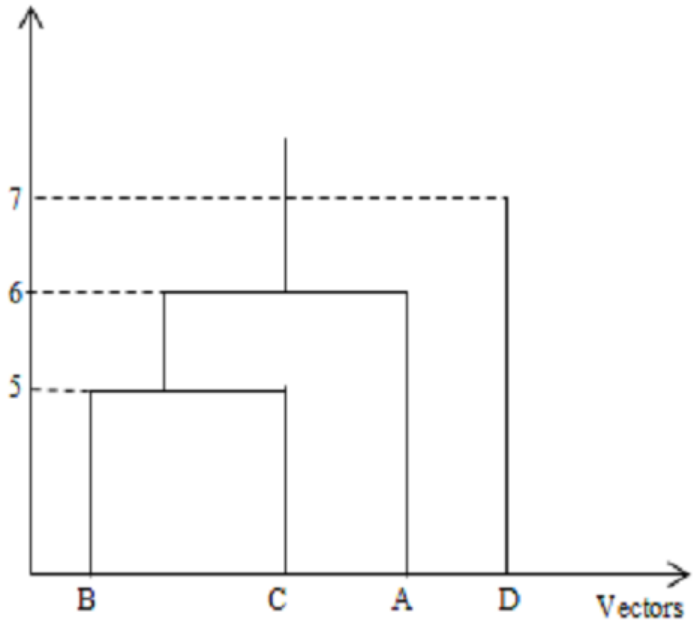

Figure 2.2: The dendogram of clusters

Chapter 3

HIERARCHICAL CLUSTERING CONCEPT

3.1 Hierarchical Clustering and Experiment

Hierarchical clustering is the way of clustering objects into a tree of cluster [18]. It is also the method of arranging things into groups so that their objects are similar. And the hierarchical clustering of spike events is a process of clustering events that are alike or similar in topology, morphology or both. For instance, one might imagine a shop keeper receiving 80 cases of toothpaste, only to know that exchange for a marvelous price, the distributor had put together the various kinds of white and red toothpaste in each of the cases.

The task of arranging all of the toothpaste into meaningful groups (Ipana in one group, Close–up in one group, Maclean in another, etc.) and someone would have made a list of what is in each packet of toothpaste to know what the received, then decide how to group the toothpaste [19].

Hierarchical clustering methods can be categorized into two which are agglomerative and divisive hierarchical clustering. It can be depend on the hierarchical decomposition is found in top-down (splitting) or bottom-up (merging). The quality of a good hierarchical clustering methods have the problem of not adjusting the split or merge decision once. It has been carried out and it cannot be reversed or correct.

There are two types of hierarchical clustering methods which are:

i. Agglomerative hierarchical clustering (Bottom-up).
ii. Divisive hierarchical clustering (Top-down).

Agglomerative Hierarchical Clustering:

This is a bottom-up method that begins by positioning each object in its own larger clusters, until it either becomes a single cluster or it reaches discontinuous conditions. Many of this hierarchical clustering belongs to this class and they are different in the inter-cluster similarity.

Divisive Hierarchical Clustering:

This is a top-down method that begins with the splitting of the objects from a single cluster to different clusters of each object formed. The divisive hierarchical clustering splits the cluster in bit forms, until each object forms a clusters of its kind or reaches a discontinue conditions are satisfied in such a way that a desired amount of clusters are generated or within a certain threshold of diameter of each cluster is obtained.

Opposition between Agglomerative and Divisive Hierarchical Clustering:

In the use of application for AGNES (Agglomerative Nesting), the agglomerative hierarchical clustering method: Originally, AGNES keeps each data point into a cluster of its own, combining them gradually according to some principle while the divisive hierarchical clustering method using the application of DIANA (Divisive Analysis).

However, DIANA has a position of all the data points to form one initial cluster. The splitting of the cluster is carried out by the Maximum Euclidean Distance between the closest marching data points in the cluster. The cluster subdivides till it reaches each new cluster that remains only a single data point on it.

Either of the divisive or agglomerative hierarchical clustering, it can be used to identify the amount of clusters as a termination condition or the specify amount of cluster as an end of merge or split cluster. This algorithm takes minimum distance to estimate the distance that links clusters which is called the nearest-neighbor clustering algorithm. The single-linkage algorithm is when the clustering process is stopped at which the distance between closest clusters more than arbitrary threshold.

3.2 Problems of Hierarchical Clustering

a. The selection of split or combined clusters: This is a problem to determine the types of hierarchical clustering methods (agglomerative hierarchical clustering or divisive hierarchical clustering) to be chosen or use to cluster objects.

b. It cannot return what it has done previously in the clusters: This is a problem in agglomerative hierarchical clustering when it cannot start the clustering from the beginning because of numerous merging of the clusters.

c. It causes a low quality if not properly taking the steps well: This affects the agglomerative or divisive hierarchical clustering when the similarity of distance measure is not calculated properly.

The methods that are used to improve hierarchical clustering by using other clustering techniques are BIRCH, ROCK and CHAMELEON. The Birch is used to cluster a large size of numerical data by combining hierarchical clustering.
The Birch overwhelm the challenges of agglomerative cluster methods which are scalability and unable to do what it has done in the past steps.

There are two Birch conceptions which are clustering feature and clustering feature tree usually use to review cluster image. This concepts enables the clustering method to achieve good scalability and speed in large databases and also enhances the successful dynamic and incremental clustering of incoming data point.

Rock is a hierarchical clustering for categorical attributes. Rock is defined as a hierarchical clustering algorithm that investigating the amount of mutual neighbors between two objects of the data with categorical attributes. In clustering categorical data, the distance measure cannot be used for high quality cluster. Moreover, the similarity between points that is most clustering algorithms and it leads an error. Rock normally consider the neighbors of individual pairs of objects, in Rock, when two similar objects have similar neighborhoods and it can be cluster same and merged in the same cluster.

Chameleon is a hierarchical clustering algorithm using dynamic modelling. Chameleon is defined as similarity between pairs of clusters that uses dynamic modelling of a hierarchical clustering algorithm. The Chameleon shows how the interconnectivity and close to each other. Chameleon is a hierarchical clustering algorithm that makes use of k-nearest neighbor graph to create a spare graph and

one of the vertices of the graph that indicate the data point. It occurs at the edge between two vertices when one object is among the k - most alike data point of the other.

It is also used in agglomerative hierarchical clustering algorithm and graph partitioning algorithm for partition the k -nearest neighbor graph into a big number of little sub cluster and merges sub cluster of similar data points. The neighborhood radius of a data points is found by the density of the region in which the data points are placed, [17].

3.3 Experiment Using Hierarchical Method

The aim of this experiment is to cluster the fifteen clients (vectors) which shows the two clusters where the clients grouped into two clusters (cluster one and cluster two) for which one of the two clusters represents the clients` loan granted and the other cluster are not granted by using two methods of clustering; hierarchical method and k-means method which are used to solve the data problem in this experiment.

We have five attributes for each of the vectors and the attributes are represented by notation A_i, $i = 1,\ldots,5$ where $A_1 = \text{net income (USD)}$, $A_2 = \text{Age (years)}$, $A_3 = \text{last employment period (years)}$, $A_4 = \text{requested loan amount}$ and $A_5 = \text{loan maturity (years)}$. Each of the customers is denoted by the notation C_i, $i = 1,2,\ldots,15$ and the Microsoft Excel program is used for numerical and statistical calculations in both experiments in Chapter 3 and Chapter 4 in this study.

We have the raw data and normalized data which are shown in Table 3.1 and Table 3.2 below. The normalized data assist the clients to get the actual distance between the clients in order to cluster the clients with its attributes. A dendrogram is a tree

diagram or tree structured graph usually used to show the presentation of clusters or represents the relationship of alike among a group of clusters in hierarchical clustering. Here in Chapter 3, Euclidean distance measure is used in order to evaluate the distance between each of clients and then, the clusters of the clients are represented in dendogram for the experiment.

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \tag{3.1}$$

Using this formula to calculate the mean and standard deviation of the actual distance of the clients in order to normalize the data.

$$\frac{x-\bar{x}}{\text{standard d}\textit{eviation}} \tag{3.2}$$

Where $\bar{x}$ is the mean of x.

Table 3.1: The raw data for the experiment

	A_1	A_2	A_3	A_4	A_5
C_1	1073	29	3	3000	36
C_2	893	32	4	3000	36
C_3	664	25	2	6000	36
C_4	1348	34	2	8000	36
C_5	250	20	0.5	2000	24
C_6	400	24	3	2500	12
C_7	140	25	1	1500	30
C_8	524	39	5	5000	36
C_9	662	32	4	6500	36
C_{10}	1695	37	7	15000	24
C_{11}	1743	47	9	20000	36
C_{12}	231	26	2	3000	36
C_{13}	1543	48	6	16000	36
C_{14}	359	27	2	2000	36
C_{15}	944	33	5	10000	24

By using the Table 3.1 and Table 3.2 for the raw and normalized data for the experiment, respectively, the distance of each vector with its attributes from one

vector of a client to another using Euclidean distance measure to determine the distance value is calculated.

At the beginning of the experiment, all the vectors are in different clusters since the used clustering is Agglomerative clustering, so, we suppose to know that there are fifteen clusters each contains one vector. The distance value of the clusters is shown in step 1 below.

Table 3.2: The normalized data for the experiment

	A_1	A_2	A_3	A_4	A_5
C_1	0.44	-0.35	-0.29	-0.67	0.60
C_2	0.11	0.02	0.13	-0.67	0.60
C_3	-0.31	-0.84	-0.71	-0.15	0.60
C_4	0.95	0.26	-0.71	0.19	0.60
C_5	-1.07	-1.45	-1.33	-0.84	-1.04
C_6	-0.79	-0.96	-0.29	-0.75	-2.67
C_7	-1.27	-0.84	-1.13	-0.92	-0.22
C_8	-0.56	0.87	0.54	-0.33	0.60
C_9	-0.31	0.02	0.13	-0.07	0.60
C_{10}	1.58	0.63	1.38	1.39	-1.04
C_{11}	1.67	1.85	2.21	2.24	0.60
C_{12}	-1.10	-0.72	-0.71	-0.67	0.60
C_{13}	1.31	1.97	0.96	1.56	0.60
C_{14}	-0.87	-0.60	-0.71	-0.84	0.60
C_{15}	0.21	0.14	0.54	0.53	-1.04

Step 1: Table 3.3 which is symmetric shows the distance values of clusters in matrix form and its minimum distance value of a cluster. The clusters $\{C_{12}\}$ and $\{C_{14}\}$ are the most similar with the minimum distance value of 0.32 and they are merged together in the same cluster. Now, the vectors C_{12}, C_{14} are in the same cluster and the others are in different clusters.

In step 2: The distance values of the clusters from one to another is calculated using Euclidean distance and the nearest neighbor method is used to determine the minimum distance of a cluster and the calculation is shown below. Therefore, the number of clusters is being reduced as the cluster is being merged.

$$d\{(C_1,(C_{12},C_{14})\} = \min\{d(C_1,C_{12}), d(C_1,C_{14})\} = min\{1.64, 2.26\} = 1.64$$

$$d\{(C_2,(C_{12},C_{14})\} = \min\{d(C_2,C_{12}), d(C_2,C_{14})\} = min\{1.65, 1.43\} = 1.43$$

$$d\{(C_3,(C_{12},C_{14})\} = \min\{d(C_3,C_{12}), d(C_3,C_{14})\} = min\{0.95, 0.92\} = 0.92$$

Table 3.3: The distance values in step 1

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}
C_1	0	0.65	1.11	1.24	2.7	3.55	2.14	1.82	1.11	3.45	4.59	1.64	3.56	1.41	2.26
C_2	0.65	0	1.37	1.48	2.89	3.56	2.22	1.22	0.73	3.32	4.31	1.65	3.3	1.43	2.08
C_3	1.1	1.37	0	1.7	2.12	3.39	1.54	2.14	1.2	3.89	5.04	0.95	4.03	0.92	2.43
C_4	1.24	1.48	1.7	0	3.33	4.03	2.86	2.12	1.55	3	3.97	2.43	2.78	2.25	2.22
C_5	2.7	2.89	2.12	3.33	0	2.02	1.06	3.48	2.85	4.86	6.57	1.91	5.57	1.96	3.09
C_6	3.55	3.56	3.39	4.03	2.0	0	2.64	3.87	3.54	4.27	6.32	3.32	5.53	3.32	2.69
C_7	2.14	2.22	1.54	2.86	1.0	2.64	0	2.69	2.15	4.75	6.14	0.97	5.07	1.03	2.95
C_8	1.82	1.22	2.14	2.12	3.4	3.87	2.69	0	1.02	3.31	3.92	2.12	2.9	2.02	2.13
C_9	1.11	0.73	1.2	1.55	2.8	3.54	2.15	1.02	0	3.21	4.12	1.49	3.13	1.4	1.87
C_{10}	3.45	3.32	3.89	3	4.86	4.27	4.75	3.31	3.21	0	2.37	4.5	2.18	4.41	1.89
C_{11}	4.59	4.31	5.04	3.97	6.57	6.32	6.14	3.92	4.12	2.37	0	5.59	1.48	5.22	3.67
C_{12}	1.64	1.65	0.95	2.43	1.91	3.32	0.97	2.12	1.49	4.5	5.59	0	4.56	0.32	2.85
C_{13}	3.56	3.3	4.03	2.78	5.57	5.53	5.07	2.9	3.13	2.18	1.48	4.456	0	4.45	2.91
C_{14}	1.41	1.43	0.92	2.25	1.96	3.32	1.03	2.02	1.4	4.41	5.22	0.32	4.45	0	2.79
C_{15}	2.26	2.08	2.43	2.22	3.09	2.69	2.95	2.13	1.87	1.89	3.67	2.85	2.91	2.79	0

The following distance values of clusters above which are calculated using Euclidean distance and nearest neighbor method is shown in the Table 3.4. The distance values

of the clusters from one to another are calculated and the distance between the cluster with vector C_1 and the cluster with vector C_2 is the minimum distance with value of 0.65 and they are merged together in same cluster. Now, the vectors C_1 and C_2 are in the same cluster.

In step 3: the distance values of the clusters from one to another are calculated and the distance between the clusters $\{C_1, C_2\}$ and $\{C_9\}$ is the minimum with distance value of 0.73 and they are merged together in the same cluster.

Step 4: The distance value between the clusters $\{C_{12}, C_{14}\}$ and $\{C_3\}$ is the minimum so, they are merged to obtain new cluster in this step.

In step 5: the distance value of the clusters from one to another is calculated and the clusters $\{C_3, C_{12}, C_{14}\}$ and $\{C_7\}$ are the most similar with minimum distance of 0.97 and they are merged together.

In step 6: the distance values of the clusters from one to another are calculated and the clusters $\{C_1, C_2, C_9\}$ and $\{C_8\}$ are the most similar with minimum distance of 1.02 and they are merged together in same cluster.

Table 3.4: The distance values in step 2

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	$C_{12,14}$	C_{13}	C_{15}
C_1	0	0.65	1.11	1.24	2.7	3.55	2.14	1.82	1.11	3.45	4.59	1.64	3.56	2.26
C_2	0.65	0	1.37	1.48	2.89	3.56	2.22	1.22	0.73	3.32	4.31	1.43	3.3	2.08
C_3	1.11	1.37	0	1.7	2.12	3.39	1.54	2.14	1.2	3.89	5.04	0.92	4.03	2.43
C_4	1.24	1.48	1.7	0	3.33	4.03	2.86	2.12	1.55	3	3.97	2.25	2.78	2.22
C_5	2.7	2.89	2.12	3.33	0	2.02	1.06	3.48	2.85	4.86	6.57	1.91	5.57	3.09
C_6	3.55	3.56	3.39	4.03	2.0	0	2.64	3.87	3.54	4.27	6.32	3.32	5.53	2.69
C_7	2.14	2.22	1.54	2.86	1.0	2.64	0	2.69	2.15	4.75	6.14	0.97	5.07	2.95
C_8	1.82	1.22	2.14	2.12	3.4	3.87	2.69	0	1.02	3.31	3.92	2.02	2.9	2.13
C_9	1.11	0.73	1.2	1.55	2.8	3.54	2.15	1.02	0	3.21	4.12	1.4	3.13	1.87
C_{10}	3.45	3.32	3.89	3	4.86	4.27	4.75	3.31	3.21	0	2.37	4.41	2.18	1.89
C_{11}	4.59	4.31	5.04	3.97	6.57	6.32	6.14	3.92	4.12	2.37	0	5.52	1.48	3.67
$C_{12,14}$	1.64	1.43	0.92	2.25	1.91	3.32	0.97	2.02	1.4	4.41	5.52	0	4.45	2.79
C_{13}	3.56	3.3	4.03	2.78	5.57	5.53	5.07	2.9	3.13	2.18	1.48	4.45	0	2.91
C_{15}	2.26	2.08	2.43	2.22	3.09	2.69	2.95	2.13	1.87	1.89	3.67	2.79	2.91	0

In step 7: the distance values of the clusters from one to another are calculated and the clusters $\{C_3, C_7, C_{12}, C_{14}\}$ and $\{C_5\}$ are the most similar with the minimum distance value of 1.06 and they are merged together in the same cluster.

In step 8: the distance values of the clusters from one to another are calculated and the clusters $\{C_3, C_5, C_7, C_{12}, C_{14}\}$ and $\{C_1, C_2, C_8, C_9\}$ are the most similar with minimum distance of 1.11 and they are merged together in the same cluster.

In step 9: the distance values of the clusters from one to another are calculated and the clusters $\{C_1, C_2, C_3, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$ and $\{C_4\}$ are the most similar with minimum distance value of 1.24 and they are merged together in the same cluster.

In step 10: the distance values of the clusters from one to another are calculated and the distance between the cluster with vector C_{11} and the cluster with vector C_{13} is the minimum value of 1.48 and they are merged together in the same cluster. All the calculations are presented in Table 3.5

Table 3.5: The distance values in step 10

	$\{C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$	$\{C_6\}$	$\{C_{10}\}$	$\{C_{11}, C_{13}\}$	$\{C_{15}\}$
$\{C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$	0	2.02	3	2.78	2.08
$\{C_6\}$	2.02	0	4.27	5.53	2.69
$\{C_{10}\}$	3	4.27	0	2.18	1.89
$\{C_{11}, C_{13}\}$	2.78	5.53	2.18	0	2.91
$\{C_{15}\}$	2.08	2.69	1.89	2.91	0

In step 11: the distance values of the clusters from one to another are calculated and the distance from the cluster with the vector C_{10} to the cluster with the vector C_{15} is the minimum of 1.89 and they are combined together.

In step 12: the distance values of the clusters from one to another are calculated and the clusters $\{C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$ and $\{C_6\}$ are the most similar with minimum distance value of 2.02 and they are put in the same cluster.

In step 13: the distance values of the clusters from one to another are calculated and the clusters $\{C_{10}, C_{15}\}$ and $\{C_{11}, C_{13}\}$ are the most similar with the minimum distance value of 1.89 and it is combined together as shown below.

Table 3.6: The distance values in step 12

	$\{C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$	$\{C_6\}$	$\{C_{11}, C_{13}\}$	$\{C_{10}, C_{15}\}$
$\{C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$	0	2.02	2.78	2.08
$\{C_6\}$	2.02	0	5.53	2.69
$\{C_{11}, C_{13}\}$	2.78	5.53	0	2.91
$\{C_{10}, C_{15}\}$	2.08	2.69	2.91	0

In this Hierarchical experiment, the cluster $\{C_1, C_2, C_3, C_5, C_7, C_8, C_9, C_{12}, C_{14}\}$ is combined with the cluster $\{C_6\}$ to form a new cluster while the cluster $\{C_{11}, C_{13}\}$ is combined with $\{C_{10}, C_{15}\}$ to form another new cluster, and the results are given in Table 3.7.

Hence, we formed two new clusters in order to cluster the clients that loan will be granted or not in one of the two new clusters. The dendogram of this experiment is shown in Figure 3.1 below to represent how the customers have been clustered.

In the Table 3.7 below, the clusters with vectors are given in every step of the experiment. As a result, we say that the dataset is divided into two clusters. That means customers can get loan in the first cluster, there are eleven vectors in this cluster and customers cannot get loan in the second cluster with four vectors.

Table 3.7: The distance values in step 13

	$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$	$\{C_{11}, C_{13}\}$	$\{C_{10}, C_{15}\}$
$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$	0	2.78	2.08
$\{C_{11}, C_{13}\}$	2.78	0	1.87
$\{C_{10}, C_{15}\}$	2.08	1.87	0

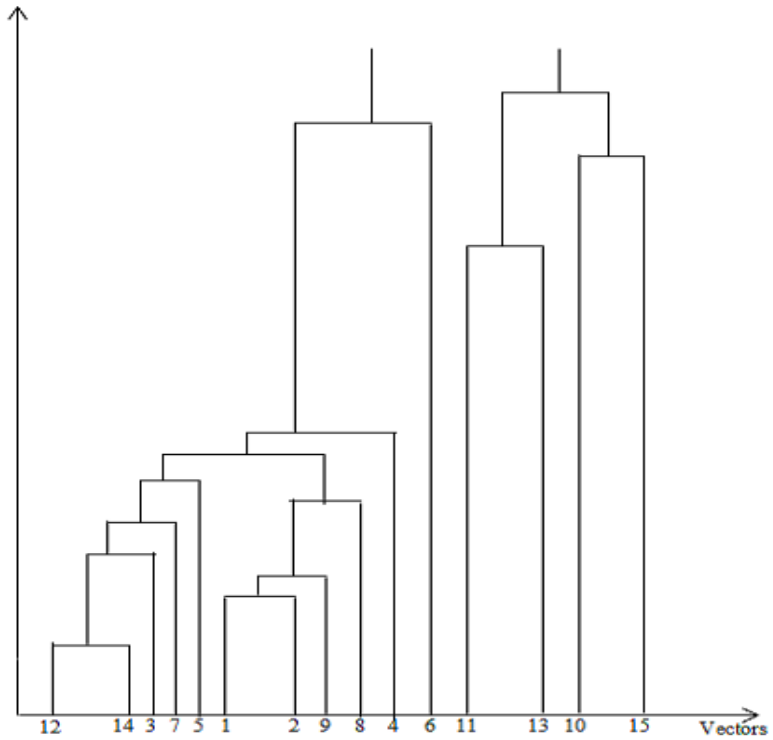

Figure 3.1: The dendogram for experiment 1

In Table 3.7, the clusters with vectors can be seen for every step. In step 1, there are 15 clusters with a single vector, but at the end there are just two clusters which are $\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$ and $\{C_{10}, C_{11}, C_{13}, C_{15}\}$.

Table 3.8: Clusters in every step

Steps	Clusters
1	$\{C_1\}$ $\{C_2\}$ $\{C_3\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_7\}$ $\{C_8\}$ $\{C_9\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_{12}\}$ $\{C_{13}\}$ $\{C_{14}\}$ $\{C_{15}\}$
2	$\{C_1,C_2\}$ $\{C_3\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_7\}$ $\{C_8\}$ $\{C_9\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
3	$\{C_1,C_2,C_9\}$ $\{C_3\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_7\}$ $\{C_8\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
4	$\{C_1,C_2,C_9\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_7\}$ $\{C_8\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_3,C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
5	$\{C_1,C_2,C_9\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_8\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_3,C_7,C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
6	$\{C_1,C_2,C_8,C_9\}$ $\{C_4\}$ $\{C_5\}$ $\{C_6\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_3,C_7,C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
7	$\{C_1,C_2,C_8,C_9\}$ $\{C_4\}$ $\{C_6\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_3,C_5,C_7,C_{12},C_{16}\}$ $\{C_{13}\}$ $\{C_{15}\}$
8	$\{C_4\}$ $\{C_6\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_{13}\}$ $\{C_{15}\}$ $\{C_1,C_2,C_3,C_5,C_7,C_8,C_9,C_{12},C_{16}\}$
9	$\{C_6\}$ $\{C_{10}\}$ $\{C_{11}\}$ $\{C_{13}\}$ $\{C_{15}\}$ $\{C_1,C_2,C_3,C_4,C_5,C_7,C_8,C_9,C_{12},C_{16}\}$
10	$\{C_{10}\}$ $\{C_{11}\}$ $\{C_{13}\}$ $\{C_{15}\}$ $\{C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8,C_9,C_{12},C_{16}\}$
11	$\{C_{10}\}$ $\{C_{11},C_{15}\}$ $\{C_{13}\}$ $\{C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8,C_9,C_{12},C_{16}\}$
12	$\{C_{10},C_{15}\}$ $\{C_{11},C_{13}\}$ $\{C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8,C_9,C_{12},C_{16}\}$
13	$\{C_{10},C_{11},C_{13},C_{15}\}$ $\{C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8,C_9,C_{12},C_{16}\}$

Chapter 4

k-MEANS CLUSTERING AND EXPERIMENT PEFORMS ON DATA

4.1 Partitioning Clustering

Partitioning clustering is the process of splitting the objects or data points into k partition in order that, each partition corresponds to a cluster. This partition is used to minimizing square error criterion which is evaluated as;

$$E = \sum\sum \left\| u - m_i^2 \right\| \tag{4.1}$$

where u is the vector or point in a cluster and m_i is the average of the vector (or cluster).

The partitioning cluster which review two properties are as follows:

i. Each cluster must have at least one data point or object.

ii. Each data point or object must belong to absolutely one cluster or group.

This gives the drawback of this algorithm by having a data point close to the center of another cluster which gives a low result due to overlapping of objects or data points, [20].

Partition clustering is the way of dividing the n objects of data set into k clusters $k \leq n$ which represents each cluster.

This is used to optimize the aim of partitioning criterion of dissimilarity function rely on distance which indicates data points within a cluster are ‘’alike’’ and the n data points of different clusters are dislike in the data set attributes.

There are two classical partition methods which are:

i. k-Means Algorithm

ii. k-Medoids.

4.2 The k-Means Method

The k-Means (centroid–based technique) algorithm is the process of splitting the n objects into k clusters and takes m as input parameter in which the inter-cluster similarity is low but the intra-cluster similarity is high. The cluster similarity is used in k-Means algorithm and this is used as the mean value of the objects in a cluster. The clustering similarity can be represented as clusters centroid or center of gravity.

Steps of k-Means algorithms:

a. Arbitrarily selects k clusters of objects randomly.

b. Each of the k clusters represents a cluster mean or center.

c. Objects are cluster according to the similarity of the distance between the cluster mean and object.

d. Update the cluster means to calculate the new average for each cluster.

e. It iterates the process until the similarity clusters ends (no change for the clusters).

4.3 The Square-Error Criteria

The square-error criterion uses to compressed the k clusters and split up the clusters. The totality of the square error for all the data points or objects in the data set is given by

$$E = \sum_{i=1}^{k} \sum_{p \grave{o} c_i} \left\| u - m_i \right\|^2 \tag{4.1}$$

where m_i is the average of cluster, C_i and u is the point in spacing indicating a given object.

The k-Means method is comparatively scalable and efficient in the process of big data sets and the computational complexity of the algorithm is $O(nkt)$ and the method is applied when the average of a cluster is defined. Some of the applications of data are not used in k-Means method that has categorical attributes and it is not moral for finding clusters of very different size and pale to noise and outlier data points.

The K-Medoids method is the process of using each representative object per cluster and each remaining objects is clustered by comparing with the nearest similar representative object. This partition method is carried out based on the rule of minimizing the sum of the dissimilarities between reference point and each object.

This is absolute error criterion formula for all objects or data points used as:

$$E = \sum_{J=1}^{K} \sum_{p \grave{o} c_j} \left\| u - o_j \right\| \tag{4.2}$$

where u represents point in a space of a given object in cluster, C_j, and o_j is the representative object of C_j. The error calculation ends until no change occurs.

According to the study in [17], the K-means algorithm is a method that is better than K-Medoids in terms of cost of processing. The K-Medoids has little effect of noise or other extreme values and also has more costly process than the k-Means algorithm.

The k-Means algorithm is used to cluster of the client's events in the following problem.

4.4 Experiment by using k-Means Algorithm

In step 1, using the normalized data in Table 3.2, arbitrarily it is divided into two clusters. The client's events of vector from one to vector seven are clustered in Cluster 1 while the client's events of vector eight to vector fifteen are clustered in Cluster 2.

The centroid or mean of the vectors in Cluster 1 and Cluster 2 are calculated as below. Centroid is the average or mean of the points in the cluster while the clusters are represented by centroids.

Using Euclidean distance measure, the distance of each vector with the centroid of each attributes of Cluster 1 and Cluster 2 are calculated and also the distance is evaluated to the remaining vectors of the clusters to compare the similarity or the minimum distance of the Cluster 1 and Cluster 2 where D_1 is the distance from the

centroid of Cluster 1 to each point in dataset and D_2 is the distance from the centroid of Cluster 2 to each point in dataset.

For Cluster 1:

Table 4.1: Centroid of Cluster 1 in step 1

	A_1	A_2	A_3	A_4	A_5
C_1	0.44	-0.35	-0.29	-0.67	0.60
C_2	0.11	0.02	0.13	-0.67	0.60
C_3	-0.31	-0.84	-0.71	-0.15	0.60
C_4	0.95	0.26	-0.71	0.19	0.60
C_5	-1.07	-1.45	-1.33	-0.84	-1.04
C_6	-0.79	-0.96	-0.29	-0.75	-2.67
C_7	-1.27	-0.84	-1.13	-0.92	-0.22
Centroid	**-0.28**	**-0.60**	**-0.62**	**-0.95**	**-0.22**

For Cluster 2:

Table 4.2: Centroid of Cluster 2 in step 1

	A_1	A_2	A_3	A_4	A_5
C_8	-0.56	0.87	0.54	-0.33	0.60
C_9	-0.31	0.02	0.13	-0.07	0.60
C_{10}	1.58	0.63	1.38	1.39	-1.04
C_{11}	1.67	1.85	2.21	2.24	0.60
C_{12}	-1.10	-0.72	-0.71	-0.67	0.60
C_{13}	1.31	1.97	0.96	1.56	0.60
C_{14}	-0.87	-0.60	-0.71	-0.84	0.60
C_{15}	0.21	0.14	0.54	0.53	-1.04
Centroid	**0.21**	**0.14**	**0.54**	**0.53**	**-1.04**

Table 4.3: Distance from centroids in step 1

	D_1	D_2
C_1	1.17	1.73
C_2	1.33	1.39
C_3	0.94	2.07
C_4	1.86	1.54
C_5	1.62	3.51
C_6	2.56	3.70
C_7	1.20	3.01
C_8	2.07	1.26
C_9	1.35	1.09
C_{10}	3.65	2.20
C_{11}	5.12	3.14
C_{12}	1.18	2.52
C_{13}	4.08	2.18
C_{14}	1.05	2.44
C_{15}	1.99	1.28

In the Table 4.3, we then compare each distance vector in between Cluster 1 and Cluster 2 to remaining of the vectors that have the minimum distance values in Cluster 1 and the vectors with maximum distance values in Cluster 2 to cluster the vectors that has the same similarity by evaluating the distance between them.

In step 2, the vectors $C_1, C_2, C_3, C_5, C_6, C_7, C_{12}, C_{14}$ are in Cluster 1 and the following vectors $C_4, C_8, C_9, C_{10}, C_{11}, C_{13}, C_{15}$ are in Cluster 2 and we calculate the centroid of the Cluster 1 and Cluster 2. Then we determine the distance from Cluster 1 and Cluster 2 to the centroids, respectively by using Euclidean distance formula.

And we then compare each distance vector in between Cluster 1 and Cluster 2 to the remaining of the vectors that have the minimum distance values in Cluster 1 and the vectors with maximum distance value in Cluster 2.

In step 3, the vectors $C_1, C_2, C_3, C_5, C_6, C_7, C_9, C_{12}, C_{14}$ are in Cluster 1 which has a new vector C_9 is added in cluster 1, and the resulting vectors $C_4, C_8, C_{10}, C_{13}, C_{15}$ are in Cluster 2.

Also, we calculate the centroid of the cluster 1 and cluster 2 and determine the distance from Cluster 1 and Cluster 2 to the centroids of each cluster, respectively, by the Euclidean distance formula to calculate.

We compare each distance vector in between Cluster 1 and Cluster 2 to the remaining of the vectors that have the minimum distance value in Cluster 1 and the vectors with maximum distance value in Cluster 2.

In step 4, the vectors $C_1, C_2, C_3, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}$ are in Cluster 1 which a new vector C_8 is added to the pervious step in Cluster 1 and the following vectors $C_4, C_{10}, C_{11}, C_{13}, C_{15}$ are in Cluster 2. We calculate the centroid of the Cluster 1 and Cluster 2 and determine the distance from Cluster 1 and Cluster 2 to the centroids of each cluster using Euclidean distance formula to calculate.

We compare each distance vector in between Cluster 1 and Cluster 2 to the remaining of the vectors that have the minimum distance value in Cluster 1 and the vectors with maximum distance value in Cluster 2.

In step 5, the vectors $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}$ such that a new vector 4 is added to the previous step in Cluster 1 and the following vectors $C_{10}, C_{11}, C_{13}, C_{15}$ are in Cluster 2.

We calculate the centroid of the Cluster 1 and Cluster 2 and determine the distance from Cluster 1 and Cluster 2 to the centroids of the clusters by using Euclidean distance formula.

In step 6, there are the same vectors in Cluster 1 and Cluster 2 as previous step; there are no changes in the cluster. Therefore, the Cluster 1 and Cluster 2 in step 5 are same with step 6, and by using k-Means method to cluster the clients' events that has the same similarity features and extract information from the data. The tables below are the clustering of the clients with its attributes. Centroid calculations are presented for step 6 for Cluster 1 and Cluster 2, respectively, in the following:

For Cluster 1:

Table 4.4: The centroid of Cluster 1 in step 6

	A_1	A_2	A_3	A_4	A_5
C_1	0.44	-0.35	-0.29	-0.67	0.60
C_2	0.11	0.02	0.13	-0.67	0.60
C_3	-0.31	-0.84	-0.71	0.15	0.60
C_4	0.95	0.26	-0.71	0.19	0.60
C_5	-1.07	-1.45	-1.33	-0.84	-1.04
C_6	-0.79	-0.96	-0.29	-0.75	-2.67
C_7	-1.27	-0.84	-1.13	-0.92	-0.22
C_8	-0.56	0.87	0.54	-0.33	0.60
C_9	-0.31	0.02	0.13	-0.07	0.60
C_{12}	-1.10	-0.72	-0.71	-0.67	0.60
C_{14}	-0.87	-0.60	-0.71	-0.84	0.60
Centroid	**-0.43**	**-0.42**	**-0.46**	**-0.52**	**0.08**

For Cluster 2:

Table 4.5: The centroid of Cluster 2 in step 6

	A_1	A_2	A_3	A_4	A_5
C_{10}	1.58	0.63	1.38	1.39	-1.04
C_{11}	1.67	1.85	2.21	2.24	0.60
C_{13}	1.31	1.97	0.96	1.56	0.60
C_{15}	0.21	0.14	0.54	0.53	-1.04
Centroid	**1.19**	**1.15**	**1.27**	**1.43**	**-0.22**

Table 4.6: Distance from centroids in step 6

	D_1	D_2
C_1	1.05	3.21
C_2	1.06	2.97
C_3	0.81	3.65
C_4	1.79	2.64
C_5	1.89	4.94
C_6	2.84	4.65
C_7	1.25	4.61
C_8	1.73	2.73
C_9	1.01	2.79
C_{10}	3.66	1.05
C_{11}	4.96	1.71
C_{12}	0.94	4.21
C_{13}	3.92	1.21
C_{14}	0.81	4.12
C_{15}	2.02	2.00

Table 4.6 belongs to step 6, it shows the distance calculations from each of the centroids of each vectors in step 6. These calculations are the same as in step 5 distance calculations.

The clustering terminates at this stage since distance calculations are the same as in step 5 and step 6. Two clusters with vectors are obtained in step 5 but algorithm terminates in step 6.

Like the first experiment, there are two clusters with exactly same vectors in this second experiment. But comparing two algorithms we have seen that the k-Means algorithm is better than Hierarchical algorithm since we need huge matrices in Hierarchical algorithm by the way we spend more time in the calculations.

Table 4.7: Clusters for every step

Steps	Clusters
1	$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7\}$ $\{C_8, C_9, C_{12}, C_{14}, C_{15}\}$
2	$\{C_1, C_2, C_3, C_5, C_6, C_7, C_{12}, C_{14}\}$ $\{C_8, C_9, C_{10}, C_{11}, C_{13}, C_{15}\}$
3	$\{C_1, C_2, C_3, C_5, C_6, C_7, C_9, C_{12}, C_{14}\}$ $\{C_4, C_8, C_{10}, C_{11}, C_{13}, C_{15}\}$
4	$\{C_1, C_2, C_3, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$ $\{C_4, C_{10}, C_{11}, C_{13}, C_{15}\}$
5	$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$ $\{C_{10}, C_{11}, C_{13}, C_{15}\}$
6	$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{12}, C_{14}\}$ $\{C_{10}, C_{11}, C_{13}, C_{15}\}$

4.5 Validity of Clusters

We calculate the square error of the vectors of cluster 1 and cluster 2 in every step by using this formula below with the given set of N samples that has been partition into k cluster $\{C_1, C_2, \ldots, C_k\}$. Each sample is a good one cluster and each cluster C_k has n vectors such that $\sum_{i=1}^{k} n_i = N, i = 1, \ldots..k.$

$$e_k^{\,2} = \sum_{i=1}^{n_k} \left(u_{ik} - M_k\right)^2 \tag{4.3}$$

The square error for the whole clustering space having k cluster is shown below

$$E_k^{\,2} = \sum_{K=1}^{K} e_k^{\,2} \tag{4.4}$$

where the final clusters of the vectors of the k-Means experiment is optimum. The numbers e_1^2 and e_2^2 represent the error values in cluster 1 and cluster 2 of each of the steps, respectively while the total error is the summation of the two errors of each of the steps. As the each of the steps has been calculated, as more efficient is the cluster and the error is being reduced after each of the steps.

The following calculation below is the square error and the total error calculations for the last two steps of k-Means experiment.

In step 4, the centroid of the cluster in cluster 1 is $\{-0.57, -0.49, -0.44, -0.59, 0.03\}$ and also the centroid results in cluster 2 are $\{1.14, 0.97, 0.88, 1.18, -0.05\}$.

$e_1^{\,2} = 2.78 + 3.84 + 3.13 + 0.81 + 10.76 = 21.32$ and the square error for cluster 2 is calculated as follows:

$$e_2^{\,2} = \{(0.95-1.14)^2 + (0.26-0.97)^2 + (-0.71-0.88)^2 + (0.19-1.18)^2$$
$$+(0.60+0.05)^2 + (1.58-1.14)^2 + (0.63-0.97)^2 + (1.38-0.88)^2 +$$
$$(1.39-1.18)^2 + (-1.04+0.05)^2 + (1.67-1.14)^2 + (1.85-0.97)^2$$
$$+(2.21-0.88)^2 + (2.24-1.18)^2 + (0.60+0.05)^2 + (1.31-1.14)^2$$
$$+(1.97-0.97)^2 + (0.96-0.88)^2 + (1.56-1.18)^2 + (0.60+0.05)^2$$
$$+(0.21-1.14)^2 + (0.14-0.97)^2 + (0.14-0.97)^2 + (0.54-0.88)^2$$
$$+(0.53-1.18)^2 + (-1.04+0.05)^2\} = 1.42+3.09+4.65+2.72+3.21 = 15.09.$$

Total error of step 4, is $E^2 = 21.32 + 15.09 = 36.41$.

Then, in step 6, the centroids of vectors in cluster 1 and in cluster 2 are

$Centroid_1$ = (- 0.43,- 0.42,- 0.46,- 0.52, 0.08)

and

$Centroid_2$ = (1.19, 1.15, 1.27, 1.43,- 0.22) , respectively.

And, the square error of cluster 1 is calculated as follows:

$$e_1^{\,2} = \{4.88+4.34+3.20+1.36+11.06\} = 24.84$$

$$e_2^{\,2} = (1.58-1.19)^2 + (0.63-1.15)^2 + (1.38-1.27)^2 + (1.39-1.43)^2$$
$$+(-1.04+0.22)^2 + (1.67-1.19)^2 + (1.85-1.15)^2 + (2.21-1.27)^2$$
$$+(2.24-1.43)^2 + (0.60+0.22)^2 + (1.3-1.19)^2 + (1.97-1.15)^2$$
$$+(0.96-1.27)^2 + (1.56-1.43)^2$$
$$+(0.60+0.22)^2 + (0.21-1.19)^2 + (0.14-1.15)^2 + (0.54-1.27)^2$$
$$+(0.533-1.43)^2 + (-0.14+0.22)^2$$
$$= 1.37+2.46+1.52+1.49+2.67 = 9.51.$$

Total error of step 6 is $E^2 = 9.51 + 24.84 = 34.35$.

The total error in step 5 is less than the total error in step 4 and the clusters reaches its peak and better than the clusters in the previous steps.

In the Table 4.8, the step 5 and step 6 have the same square error values so the algorithm terminates at step 5.

Table 4.8: The error calculations

	The calculation of square error	Total error
Step 1	e_1^2=18.08, e_2^2=36.20.	54.28.
Step 2	e_1^2=18.40, e_2^2=26.3	44.70.
Step 3	e_1^2=17.75, e_2^2=19.86	37.61.
Step 4	e_1^2=21.32, e_2^2=15.09	36.41.
Step 5	e_1^2=24.84, e_2^2=9.51	34.35.
Step 6	e_1^2=24.84, e_2^2=9.51	34.35

Chapter 5

CONCLUSION

Recently, because of the large data involves in our daily activities and the new innovations in technology, data mining was introduced to extract knowledge or patterns from data.

Knowledge discovery is used in data mining to assist in the lives of individuals and organizations to gain ideas and make life easy to cluster pattern extract from data. Clustering in data mining has assist in grouping of data elements of same class or similar into one cluster and also helps in business, companies and individuals to gain a pattern from data and assist to cluster different data according to their similarities into various clusters. Clustering is an unsupervised learning while classification is a supervised learning and unsupervised learning is normally used to discover structure or pattern from data without explanation of label data.

In this thesis, Hierarchical clustering method and the k-Means clustering method were used to cluster fifteen vectors (Clients) with five attributes into two clusters to represent the clients loan granted or not in each of the clusters in order to assist in clustering the new clients of same attributes to know where it belongs in various clusters.

REFERENCES

[1] Wu, X., Zhu, X., Wu, G.-Q. & Ding, W. (2014). Data Mining with Big data. *IEEE Transaction on Knowledge and Data Engineering*. 26, 97-107.

[2] Ruxandra - Stefania, P. (2012). Data Mining in Cloud Computing. *Database System Journal*. 3, 67-71.

[3] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for data mining. *Journal of Data Warehousing*. 5, 4-12.

[4] Joshi, A., Rajneet, K. (2013). Comparative Study of Various Clustering Techniques in Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*. ISSN: 2277 128X. 3, 55-57.

[5] Rafsanjani, M., Kuchaki, V., Z. Asghari & Chukanlo, N. E. (2012). A Survey of Hierarchical Clustering Algorithms. *The Journal of Mathematics and Computer Science*. Oxford University press. 5, 229-240.

[6] Elvarasi, A. S., Akilandeswari, J. & Sathiyabhama, B. (2011). A Survey on Partition Clustering Algorithms. *International Journal of Enterprise Computing and Business System*, (online). 1, 1-11.

[7] Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems.* 17, 107-145.

[8] http;//en.wikipedia .org/wiki/unsupervised_learning.

[9] Zhang, W., Qin, Z. (2011). Clustering Data and Imprecise Concepts. *IEEE International Conference on Fuzzy Systems*. Tapei, Taiwan. 603-608.

[10] Koskela, T., Hakola, S., Chen, T. & Lehtomaki, J. (2010) Clustering Concept Using Device–to-Device Communication in Cellular System. *IEEE Communications Society, WCNC 2010 Proceedings.* 3-6.

[11] Kovacs, L., Repasi, T., Baksa-Varga, E. & Barabas, P. (2008). Clustering Based on Context Similarity. *The First International Conference on Complexity and Intelligence of the Artificial and Natural Complex Systems*. ISBN: 978-0-7695-3621-7, 25, 157-165.

[12] Joachims, T. (2007). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Proc. 14th International of Mach. Learn, San Francisco, CA, USA. 143-151.

[13] Knight, K. (1999). Mining Online Text. *Communications of the ACM*. 42, 58-61.

[14] Fabrizio, S. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR).* 34, 1-47.

[15] Kim, H., Howland, P. & Park, H. (2005). Dimension Reduction in Text Classification with Support Vector Machines. *Journal of Machine Learning Research*. 6, 37-53.

[16] Salton, G., Gill, M. J. (1993). Introduction to Modern Information Retrieval. *McGraw-Hill Book Co.* 1-9.

[17] Jan, J., Kamber, M. (2006). Data Mining Concepts and Techniques. *2nd Ed. San Francisco, CA, USA Morgan Kaufmann, Boston, Ma, USA: Elsevier*. 2-28.

[18] Lin, Y.-S., Jiang, J.-Y. & Lee, S.-J. (2013). A Similarity Measure for Text Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering* 26. 1575-1590.

[19] Guess, M. J. & Wilson, S. B. (2002). Introduction to Hierarchical Clustering. *Journal of Clinical Neurophysiology*. 19, 144-151.

[20] Pavel, B. (2002). Survey of Clustering Data Mining Techniques. *Technical report, Accrue Software, San Jose, California*. 25-72.

[21] Bharati, M. R. (2011). Data Mining Techniques and Applications. *Indian of Computer Science and Engineering.* 1, 301-305.

[22] Jain, A. K & Dubes, R. C. (1988). Algorithms for Clustering Data. Englewood Cliffs: *Prentice Hall*. 6, 1-8.

[23] Jain, A., Murty, M. N. & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys.* 31, 264-323.

[24] Pande, S. R, Sambare, S. S & Thakre, V. M. (2012). Data Clustering Using Data Mining Techniques. *International Journal of Advanced Research in Computer and Communication Engineering.* 1, 494-499.