International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015
ISSN: 2395-3470
www.ijseas.com

A Comparative Study on Clustering and Classification Algorithms

Jyotismita Goswami

Assistant Professor, Department of Computer Science ,Arunachal University of Studies Namsai , Arunachal Pradesh, India

Abstract: --
The abnormal rise in data is a matter of great concern in todays world .Recent studies emerged with the fact that data increases twice every year. In order to gain more precise information from this huge database mining is needed. Data mining is defined as the method of extracting hidden relationships from large databases. It binds together disciplines like machine learning, statistics, information retrieval and visualization techniques helping users in predictive analysis of raw data. Clustering deals with grouping objects into different classes based on their similar habits whereas classification classifies objects based on predefined classes. It finds its application in statistical data analysis, pattern recognition, image analysis, information retrieval etc .This paper makes a comparative study of a few classification and clustering algorithms using WEKA .

Keywords: clustering , classification, hierarchical,partitional,soft clustering etc.

I. INTRODUCTION

Data mining [6] in simple terms can be stated as the process which automates the detection of relevant patterns in a database, using well defined approaches and algorithms in analyzing historical data as well as predicting future trends. Its use is extensively found in business in making timely decisions which otherwise would have consumed more time . Clustering and classification forms the basic 2 divisions of data mining.

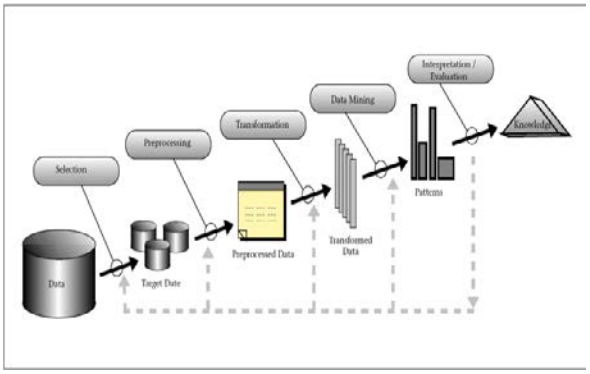


Fig 1:Data Mining as a Part of KDD Process.

Clustering technique is considered as self learning. It learns by taking into account the objects sharing similar behavior patterns thus separating them into different groups whereas classification deals with grouping based on some predefined knowledge of classes. Since the goal of clustering is to discover a new set of categories thus helping users

A. Criteria for Good Clustering.

A good clustering method needs to produce high quality clusters in which:

i. The similarity between objects of the same cluster is high.
ii. The similarity between different clusters is low.
iii. The clustering quality is dependent on both the similarity measure used by the method and its implementation.
iv. The quality is measured by its ability to discover some new information from those hidden patterns.

B. Main Categories of Clustering Methods

i. Partitioning algorithms: Construct various partitions followed by their evaluation based on some criterion.
ii. Hierarchy algorithms: Creating a hierarchical decomposition of the data set using some criterion.

iii. Density-based: Clustering based on connectivity and density functions
iv. Grid-based: Based on a multiple-level granularity structure
v. Model-based: A model is made for each of the clusters and the idea is to find the best fit of that model to each other.

C. *Distance Measures*

The standard measures generally adopted for measuring the distance are as follows:

i. Euclidean distance
ii. Manhattan distance
iii. Mahalonobis distance

II. ALGORITHMS

A. *Partitioning Method*

Non hierarchical or partitioning creates clusters at one go. Partitioning methods relocate instances by moving them from one cluster to another, starting from a particular point. The number of clusters should be pre-set by the user for this method. A relocation method iteratively reshuffles the points between the *k* clusters ending up with creation of definite 'k' clusters where the value of 'k' given as input plays a major in cluster formation. To determine the goodness of any proposed algorithm some standard metrics are used. However the common criterion function used is the squared error metric(measures the squared distance from each point to the centroid for the associated cluster).The algorithms falling under this category are discussed below in detail.

i. k-means :The k-means algorithm [7] is a simple iterative method to partition a given dataset into a user specified number of clusters, *k*. This algorithm has been discovered mainly by 5 researchers notably lloyd , forgey , friedman , rubin and mcqueen . A detailed history of k-means along with variants descriptions are given in [5]. (Fig 2) discusses about its merits and demerits and Fig 3 demonstrates its working.

ADVANTAGES	DISADVANTAGES
1) Simple, understandable 2) Items automatically assigned to clusters	1) Must pick number of clusters before hand 2) Often terminates at a local optimum. 3) All items forced into a cluster 4) Too sensitive to outliers

Fig 2:Merits and Demerits

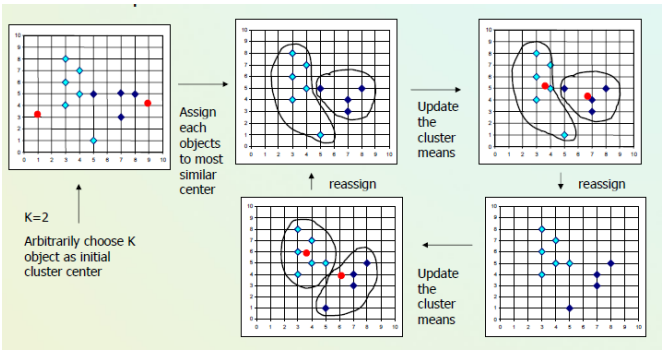


Fig 3: Working of K-Means

Variations of k-means

The k-medoid algorithm(PAM)

PAM is one of those derived from the standard K-Means .It is almost similar to k-means except that the centroids have to belong to the data set being clustered. PAM stands for "partition around medoids". The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into a set S of selected objects. If O is the set of objects that the set U = O − S is the set of unselected objects.The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object.

The algorithm has two phases:
(i) In the first phase, BUILD, a collection of k objects are selected for an initial set S.
(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.
Another version of K-means is the Kernel K-means.

ADVANTAGES	**DISADVANTAGES**
1)Algorithm is able to identify the non-linear structures. 2) Algorithm is best suited for real life data set.	1) Here number of cluster centers need to be predefined. 2) Algorithm time complexity is large.

Fig 3:Merits and Demerits of Kernel k- means.

ii. Minimum Spanning Tree Algorithm

A spanning tree is an acyclic subgraph of a graph *g*, which contains all the vertices from *g*. the minimum spanning tree (mst) of a weighted graph is the minimum weight spanning tree of that graph. zahn [8] proposes to construct an mst of a point set and delete inconsistent edges — the edges, whose weights are significantly larger than the average weight of the nearby edges in the tree. Zahn's inconsistency measure definition states - Let e denote an edge in the mst of the point set, and v1 and v2 be the end nodes of e, w be the weight of e.A depth d neighborhood n of an end node v of an edge e is defined as a set of all edges that belong to all the paths of length d originating from the node v, excluding the paths that include the edge e. Let n1 and n2 be the depth d neighborhoods of the end nodes v1 and v2. Let wn1 be the average weight of edges in n1 and σn1 be its standard deviation with similar definitions for wn2 . Zahn relies on the concept of depth-d neighborhoods, n1 and n2, for each incident point, v1 and v2, of an edge e. The neighborhood of v1 is the set of edges on paths from v1 having length no greater than d, and excluding the edge e.

iii. k- Nearest Neighbour Algorithm

K-NN [7] is derived from k-means with a little modification. Here each of k clusters are segregated based on its mean or weighted average of its points, the centroid which is the point representing the mean of the coordinates of all the points in the clusters .Here an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k *n*earest neighbors (*k* is a positive integer, typically small).

iv. Bond Energy Algorithm

The bond energy algorithm was developed for database design to determine how to group data and how to physically place data on a disk.With BEA the affinity between database attributes is based on common usage.This bond is used by the clustering algorithm as a similarity measure.

The basic steps of this algorithm is

1.Create an attribute affinity matrix in which each entry indicates the affinity between 2 associate attributes.The entries in the similarity matrix are based on the frequency of common usage of attribute pairs.

2.The BEA then converts this similarity matrix to a BOND matrix in which the entries represent a type of nearest neighbour bonding based on probability of co-access.The BEA rearranges rows or columns so that similar attributes appear close together in the matrix.

3.Finally the designer draws boxes around regions in the matrix with high similarity.

v. GA Based Clustering

Genetic algorithms (GAs)[3] are search and optimization procedures that are motivated by the principles of natural selection and natural genetics . In GAs, the role of selection and recombination operators is ver crucial.Selection operator controls the direction of search whereas recombination operator generates new regions for search. GAs perform search in complex, large and multimodal landscapes, and provide near-optimal solutions for objective or fitness function of an optimization problem. The parameters in the search space are represented in the form of strings (chromosome), encoded, by a combination of cluster centroids. A collection of such chromosomes is called a population. Initially, a random population is created, which represents different solutions in the search space.Based on the principle of survival of the fittest, a few of chromosomes are selected and each is assigned into the next generation.Biologically inspired operators like crossover and mutation are applied to chromosomes to yield new child chromosomes.The operator of selection,crossover and mutation continues several generations till the termination criterion is satisfied. The fittest chromosome seen in the last generation provides the best solution to the clustering problem.

Variations

In **genetic Kmeans algorithm (GKA)[**1], K-means operator was defined and used as a search operator instead of crossover. GKA also define a biased mutation operator specific to clustering called distance-based mutation. Using finite Markov chain theory, it was proved that the GKA converges to the global optimum. GKA searches faster than some of the other approaches.

Fast Genetic K-means Algorithm (FGKA) [2] was inspired by (GKA) with several improvements over GKA. Experiments indicate that, while K-means algorithm might converge to a local optimum, both FGKA and GKA always converge to the global optimum eventually but FGKA runs much faster than GKA.

Incremental Genetic K-means Algorithm (IGKA) [3] was an extended version of FGKA. IGKA outperforms FGKA when the mutation probability was small. IGKA was

developed with the aim to calculate the objective value Total Within-Cluster Variation (TWCV) and to cluster centroids incrementally whenever the mutation probability was small. It also converges to the global optimum.

GGA (Genetically Guided Algorithm) [4] describes a genetically guided approach to optimize the fuzzy and hard c-means (FCM/HCM respectively) functional using fitness functions. On data sets with several local extrema, the GA approach always avoids the less desirable solutions. Degenerate partitions were always avoided by the GA approach.

A hybrid genetic based clustering algorithm, called **HGA-clustering** was proposed in [9] to explore the proper clustering of data sets. This algorithm, with the cooperation of tabu list and aspiration criteria, has achieved harmony between population diversity and convergence speed. A genetic algorithm was proposed for designing the dissimilarity measure, termed Genetic Distance Measure (GDM) such that the performance of the K-modes algorithm is improved [13].

A semi-supervised clustering algorithm was proposed in [10] that combines the benefits of supervised and unsupervised learning methods. The approach allows unlabeled data with no known class to be used to improve classification accuracy. The objective function of an unsupervised technique, e.g. K-means clustering, was modified to minimize both the cluster dispersion of the input attributes and a measure of cluster impurity based on the class labels.

The **K-means Fast Learning Artificial Neural Network (KFLANN**)in [11] was a small neural network bearing two types of parameters, the tolerance, δ and the vigilance, μ. In KFLANN GA was introduced as a possible solution for searching through the parameter space to effectively and efficiently extract suitable values to δ and μ.

SPMD (Single Program Multiple Data) algorithm presented in [21] combines GA with local searching algorithm – uphill. The hybrid parallel method not only improves the convergence of GA but also accelerates the convergence speed of GA. The SPMD algorithm exploits the parallelism of GA overcomes the poor convergence properties of GA.

Genetic Weighted K-means Algorithm (GWKMA), which was a hybridization of a genetic algorithm (GA)
and a weighted K-means algorithm (WKMA),proposedby Fang-Xiang et al. GWKMA encodes each individual by a partitioning table which uniquely determines a clustering, and employs three genetic operators (selection, crossover, Mutation) and a WKMA operator.

Though possessing a number of merits with a variety of progress there exists a major problem with GAs which is their sensitivity to the selection of various parameters such as population size, crossover and mutation probabilities, etc.

vi. Neural Networks**.**

This type of algorithm[12] represents each cluster by a neuron or "prototype". The input data is also represented by neurons, which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptively during learning.Neural networks that useunsupervised learning attempt to find features in tha data that characterizes the desired output.These types of NNs are called self organising neural networks.
There are 2 types

Competitive: It is based on Hebb's Law which states that if two neurons on either side of a connection are activated synchronously, then the weight of that connection is increased.

Non competitive: With competitive learning nodes are allowed to compete and winner takes all.It has a 2 layer structure in which all nodes from one layer to other are connected.As training occurs nodes in the output layer become associated with certain tuples in the input dataset.Thus this provides grouping of these tuplesWhen a tuple is input to the NN all output nodes produce an output valueWeights are adjusted and winner is the one whose weight is similar to the input.
In the late 1980s, Teuvo Kohonen introduced a special class of artificial neural networks called **self organizing feature maps** based on competitive learning. The Kohonen model provides a topological mapping. It places a fixed number of input patterns from the input layer into a higher- dimensional output or Kohonen layer.

Training in the Kohonen network(Fig 4) begins with the winner's neighborhood of a fairly large size. Then, as training proceeds, the neighborhood size gradually decreases.The lateral connections are used to create a competition between neurons. The neuron with the largest activation level among all neurons in the output layer becomes the winner.

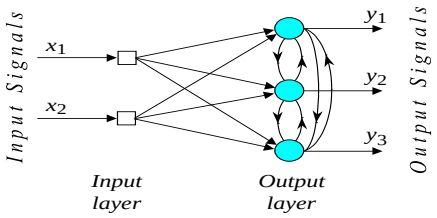


Fig 4 : Kohonen Layer Network

B. Clustering Large Data Sets

Recent research has found efficient methods for clustering large databases.

1. Read a subset of the database into main memory.
2. Apply clustering technique to data in memory.
3. Combine results with those of prior samples.
4. The in memory data are then divided ionto 3 different types:those items that will always be needed even when the next sample is brought in those that can be discarded with appropriate updates to data being kept in order to answer the problem and those that will be saved in a compressed format.
5. If termination criteria are not met then repeat from step 1.

C. Density-Based Methods

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution. The aim of these methods is to identify the clusters and their distribution parameters and continue growing the given cluster as long as the density of data points in the neighbourhood exceeds some thresholdFor that the neighbourhood of a given radius has to contain at least a minimum number of objects. An acceptable solution in this case is to use the maximum likelihood principle.DBSCAN falls under this category

i. The DBSCAN algorithm (density-based spatial clustering of applications with noise) discovers clusters of arbitrary shapes and is efficient for large spatial databases. The algorithm searches for clusters by searching the neighbourhood of each object in the database and checks if it contains more than the minimum number of objects. . It applies the directly density reachable technique.The first part ensures that the second point is close enough to the first point while second part ensures that there are enough core points close to each other.A point is said to be density reachable from another if there is a chain from one to the other containing only points that are density reachable from the previous point.

ii. EXPECTATION MAXIMIZATION:

The EM algorithm iteration takes place in 2 steps-

The Expectation Step: Using the current best guess for the parameters of the data model, we construct an expression for the log-likelihood for all observations. This expression is marginalized over the unobserved data and later will be shown to depend on both the current best guess for the model parameters treated as variables in the log-likelihood function.

The Maximization Step: Given the expression resulting from the previous step, for the next guess we choose those values for the model parameters that maximize the expectation expression. These constitute our best new guess for the model parameters.

The output of the Expectation Step codifies our expectation with regard to what model parameters are most consistent with the data actually observed and with the current guess for the parameters provided we maximize the expression yielded by this step.We stop iterating through the two steps when any further change in the log-likelihood of the observed data falls below some small threshold.

D. Simulated Annealing for Clustering.

Another general-purpose stochastic search technique that can be used for clustering is simulated annealing (SA), which is a sequential stochastic search technique designed to avoid local optima. This is accomplished by accepting with some probability a new solution for the next iteration of lower quality (as measured by the criterion function). The probability of acceptance is governed by a critical parameter called the temperature (by analogy with annealing in metals), which is typically specified in terms of a starting (first iteration) and final temperature value. Selim and Al-Sultan (1991) studied the effects of control parameters on the performance of the algorithm. SA is statistically guaranteed to find the global optimal solution.

E. Hierarchical Methods

These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These following algorithms falls under this class.

i. Agglomerative Hierarchical Clustering—Each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.It has the following subdivisions.

Single-link clustering (also called the connectedness, the minimum method or the nearest neighbor method) — method

clusters based on the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster .

Complete-link clustering (also called the diameter, the maximum method or the furthest neighbor method) forms clusters based on the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Average-link clustering (also called minimum variance method) considers the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Fig 6 shows the dendograms built by all the 3 variants of aglomerative clustering.

ii. Divisive hierarchical clustering — All objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues untilthe desired cluster structure is obtained. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion.

F. Soft Computing Approaches

i Fuzzy c-means clustering algorithm-This works by giving membership values to its data points

ADVANTAGES	DISADVANTAGES
1) Gives best result for overlapped data set and comparatively better then k-means algorithm. 2) Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more then one cluster center.	1) Apriori specification of the number of clusters. 2) With lower value of β we get the better result but at the expense of more number of iteration. 3) Euclidean distance measures can unequally weight underlying factors.

G. Applications of Clustering

i. Economic science (especially market research) and in documer
ii. Cluster weblog data to discover groups of similar access patter
iii. Pattern recognition including spatial data analysis
iv. Image processing.

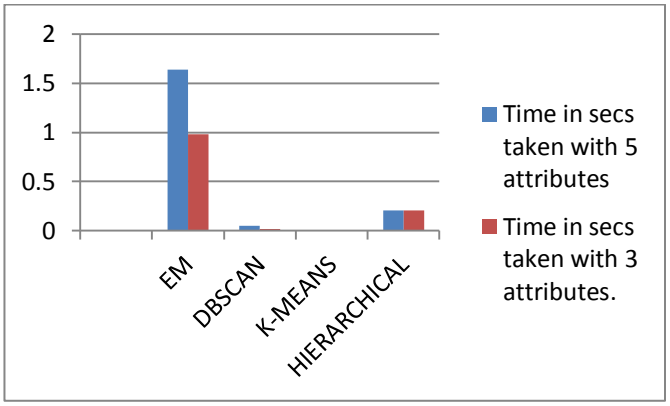
Comparison of Clustering algorithms.

Experiments and results:

We have chosen a few clustering algorithms for our experiment in order to compare their efficiency. For this purpose we have used WEKA tool to evaluate their performance. The dataset for the experiment has been collected from the UCI repository having 5 attributes.The experiment has been carried out in 2 phases.First the clustering was carried out having all the 5 attributes and their respective time to form clusters is noted. Secondly their number of attributes are reduced to three and same process has been carried out. The results state that when the number of attributes are reduced then the efficiency increases.

The results have been tabulated.

Algorithm	Time in secs taken with 5 attributes	Time in secs taken with 3 attributes.
EM	1.64	0.98
DBSCAN	0.05	0.02
K-MEANS	0	0
HIERARCHICAL	0.21	0.21



Thus we can conclude from the results that clustering time is reduced with usage of less attributes.

Classification

A. *Some classification algorithms*

Naïve Bayes : The Naïve Bayes Classifier technique [16] is based on theBayesian theorem and is particularly suited inputs which has dimension high. Despite of its simplicity, It can often outperform more sophisticated classification methods. The Naive Bayes continuous which works similarly with continuous variable as input [17].

Multilayer Perceptron
It is the most popular network architecture in today world. The units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are arranged in a layered feed forward topology. The network has a simple input-output model, with the weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptrons is the design specification of the number of hidden layers and the number of units in these layers [17].

Random Forest Tree(Rnd Tree)
A Random Tree consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. For regression problems, the tree response is estimated for the dependent variable given by the predictors [17].

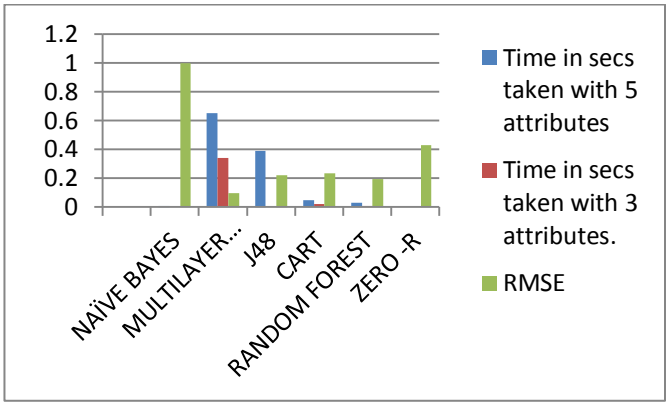
CART
The CART method under Tanagra is a very popular classification tree learning algorithm. CART builds a decision tree by splitting the records at each node, according to the function of a single attribute it uses the gini index for determining the best split. The CS-CRT is similar to CART but with cost sensitive classification [18].

J48
J48 is an open source version of C4.5 developed by Ross Quinlan. C4.5 is a software extension and thus improvement of the basic ID3 algorithm designed by Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [19]. For inducing classification rules in the form of Decision Trees from a set of given examples C4.5 algorithm was introduced by Quinlan. C4.5 is an evolution and refinement of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation etc.

Comparison of Classification algorithms.

Along with clustering we have made the same experiment with some of the selected classification algorithms.The findings have been discussed here with proper diagrams.

ALGORITHM	Time in secs taken with 5 attributes	Time in secs taken with 3 attributes.	RMSE
NAÏVE BAYES	0.01	0	0.996
MULTILAYER PERCEPTRON	0.65	0.34	0.0989
J48	0.39	0	0.219
CART	0.05	0.02	0.2345
RANDOM FOREST	0.03	0.01	0.1952
ZERO –R	0	0	0.4309



Form the experiments we have seen that naïve bayes takes the minimum time in both experiments but with highest RMSE value whereas multilayer takes the minimum RMSE value.

III. CONCLUSION AND FUTURE WORK

The study made presents a tutorial overview of the main clustering methods used in Data Mining along with concepts of a few classification algorithms. We have also carried out an experiment checking the performance change based on attribute selection.Thus as future work we can extend it to more algorithms for experiments as a few have been taken for this purpose.The algorithms discuused in this study can be improved further through thorough research made in this field so that its application gets more intense.

References

1. K. Krishna and M. N. Murty, "Genetic K-Means Algorithm", IEEE Transaction On Systems, Man, And Cybernetics Part B:CYBERNETICS, Vol. 29, No. 3, June 1999.
2. Yi Lu, Shiyong Lu, Farshad Fotouhi ,"FGKA: A Fast Genetic K-means Clustering Algorithm", SAC'04 Nicosia,Cyprus. , March 2004 ACM 1-58113-812-1/03/04.
3. Yi Lu1, Shiyong Lu1, Farshad Fotouhi1, Youping Deng, d.Susan, J. Brown," an Incremental genetic K-means algorithmand its application in gene expression data analysis", BMCBioinformatics 2004.
4. Hall, L.O., Ozyurt I. B., and Bezdek, J.C, "Clustering with a genetically optimized approach". IEEE Trans. On Evolutionary Computation,1999.
5. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs.
6. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", second edition, Morgan Kaufmann Publishers an imprint of Elsevier.
7. XindongWu · Vipin Kumar · J. Ross Quinlan , Joydeep Ghosh · Qiang Yang , Hiroshi Motoda ,Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu ,Zhi-Hua Zhou · Michael Steinbach · David J. Hand , Dan Steinberg, Top 10 algorithms in data mining, Springer-Verlag London Limited ,2007.
8. C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, C-20:68–86, 1971.
9. Y. Liu, Kefe and X. Liz," A Hybrid Genetic Based Clustering Algorithm", Proceedings of the Third International Conference on Machine Leaming and Cybernetics, Shanghai, 26-29 August 2004.
10. Ayhan D., K. P. Bennett, M. J. Embrechts," Semi-Supervised Clustering Using Genetic Algorithms",2002.
11. Yin Xiang, Alex Tay Leng Phuan," Genetic Algorithm Based K-Means Fast Learning Artificial Neural Network" ,Nanyang Technological University,2004.
12. R.C Chakraborty,Fundamentals of Neural Networks,June,2010.
13. S. Chiang, S. C. Chu, Y. C. Hsin and M. H. Wang,"Genetic Distance Measure for K-Modes Algorithm", International Journal of Innovative Computing, Information and Control ICIC,ISSN 1349-4198, Volume 2, Number 1, pp 33-40 February 2006.
14. Jawai Han and M. Kamber," Data Mining Concepts andTechniques", second edition, Elsevior.
15. Rui Xu ,and Donald Wunsch, "Survey of Clustering Algorithms", IEEE Transactions On Neural Networks, Vol. 16, No. 3, MAY 2005.
16. Phimphaka Taninpong, Sudsanguan Ngamsuriyaroj," Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification" , 2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing. Materials Research, Vols. 171-172, pp. 543-546, 2011.
17. Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.
18. Tanagra-Data Mining tutorials http://data-miningtutorials.blogspot.com/.

19. http://www.c4.5-Wikipedia, the free encyclopedia.htm accessed on 16/12/2010.