DEVELOPMENT OF DATA MINING SYSTEM TO ASSESS THE PERFORMANCE OF EAHC WITH FARTHEST FIRST CLUSTERING ALGORITHM USING SENSOR DISCRIMINATION DATASET

[1]K.Thulasiram, [2]Dr.S.Ramakrishna, [3]Dr.M.Jayakameswaraiah
[1]Research Scholar Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India,
[2]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India,
[3]Assistant Professor, School of Computer Science and Applications, Reva University, Bangalore, Karnataka, India,

***Abstract*:** Data mining has evolved into a vital and active area of research because of theoretical challenges and practical applications associated with the problem of discovering (or extracting) interesting and previously unknown knowledge from very large real-world databases. Many aspects of data mining have been investigated in several related fields. The last decade had experienced a revolution in the information availability and exchange via the internet. The main aim of the research is decision making approach to take right decision for design and development of classification and clustering techniques of data mining system. Regarding the classification and clustering algorithm, we rely on Upgraded Random Forest classification algorithm and EAHC with Farthest First Clustering as a greedy classification method. Using this method, several approaches are proposed to work on binary and multiclass scenarios. The proposed system gives better classification and clustering performance when compared to other popular algorithms.

***IndexTerms* - Data Mining, Classification, Clustering, Farthest First Algorithm, EAHC Clustering Algorithm.**

I. INTRODUCTION

The database technologists have been seeking efficient means of storing, retrieving and manipulating data; the machine learning communities have focused on developing techniques for learning and acquiring knowledge from the data. At times the data can be considered to be mined for strategic planning on research and development in this area which is often referred to as Data Mining (DM) and Knowledge Discovery in Databases (KDD)[5]. The role of scientist is to develop from data, new information. To discover the underlying model that governs the functioning of the physical world and encapsulate the same in theories that can be used for predicting the future. As the background of all scientific innovations particularly theories are the same. In Data mining system, it is important to understand the difference between a model and a pattern[6,18]. Model is a global summary of the dataset and makes statements about any point in the full measurement space while pattern describes a structure and the relationship to a relatively small part of the data or the space in which the data would occur.

II. LITERATURE REVIEW

Data Mining is widely used in diverse areas. There are a number of profitable data mining systems offered today yet there are many challenges in this field. Data mining tools make use of data to construct a representation of reality in the form of a model. The resulting model describes pattern and relationship present in the data[3,11]. From a process orientation view, data mining activities fall into three general categories

Discovery: The process of sorting a database to find hidden patterns without a predetermined idea or hypothesis.
Predictive Modeling: The process of referring the patterns discovered from the database and using them to predict the future.
Forensic Analysis: The process of applying the extracted patterns to find anomalous or unusual data elements.

Data Mining is defined as extracting the information from a large pool of data. In other words, we can say that data mining is mining the knowledge from data[10,16]. This information can be used for any of the following applications:

- Fraud Detection
- Science Exploration
- Production Control
- Market Analysis and Management
- Biological Data Analysis

- Intrusion Detection
- Customer Retention
- Financial Data Analysis and Other Applications
- Corporate Analysis & Risk Management

2.1 Classification and Prediction

Classification is the process of finding a model that elaborates the data classes or concepts. The purpose is the ability to use this model to predict the class of objects whose class label is unknown. This resultant model is based on the analysis of a set of training data[2,17]. The derived model can be accessible in the following forms

- Classification Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

2.1.1 Classification: It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.

2.1.2 Prediction: It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for the prediction. Prediction can also be used for the identification of distribution trends based on available data[7,20].

2.1.3 Outlier analysis: Outliers are data elements that cannot be grouped in a given class or cluster. The outliers can be considered as noise and discarded in some applications. Outliers are defined as the data objects that do not comply with general behavior or model of the data available.

2.1.4 Evolution Analysis: Evolution Analysis refer to description and model regularities or trend for objects whose behavior changes over time[15,21].

2.2 Clustering

Cluster is a group of objects that belong to the same class. In other words, the similar objects are grouped in one cluster and dissimilar are grouped in another. Sometimes clustering is useful even when labels are available by illuminating groupings of cases or suggesting data features to be calculated to help other methods[13]. Like nearest neighbor method, clustering depends on a distance metric such as Euclidean distance (root sum squared of distances along each feature) or Manhattan distance (sum of absolute feature differences). For both algorithms, it is a serious challenge to incorporate symbolic features alongside numeric ones[1,19]. As with nearest neighbors and neural networks, clustering also uses all dimensions provided, making feature reduction essential. The main advantage of Clustering over classification is that, it is adaptable to changes and help single out useful features that distinguishes different groups.

2.3 Clustering Techniques

clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters[9.12]. The clusters can be formed based upon various parameters depending upon the clustering method. Different Clustering methods have a database of n objects or data tuples.

- Partitioning methods
- Hierarchical based methods
- Density based method
- Grid-based methods
- Model-based methods
- Clustering high-dimensional data
- Constraint-based clustering

III. EAHC CLUSTERING ALGORITHM

3.1 Cluster Divergence

In order to select which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of clarifications are required. In furthermost methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of clarifications), and a linkage criterion which identifies the deviation of sets as a function of the pairwise distances of interpretations in the sets. Here we applied a metric function to integrate the

performance of the proposed algorithm using squared Euclidean distance formula[4,8,14]. The Enhanced Agglomerative Hierarchical Clustering comprises of the following three parameters need to be considered.

1. Single Linkage
2. Complete Linkage
3. Group Average

Algorithm:

Step-1: Scan the Entire Database

Step-2: collects the reduced data set by using agglomerative technique.

Step-3: Partition the reduced dataset.

Step-4: Eliminate Outliers.

Step-5: Cluster Labeled data as Partial Cluster using Squared Euclidean distance equation.

$$||a - b||_2^2 = \sum_i (a_i - b_i)^2$$

Setp-6: Initially each item $x_1, \ldots, x_n$ is in its own cluster $C_1, \ldots, C_n$.

Step-7: Repeat until there is only one cluster left:

Step-8: Merge the nearest clusters, say C_i and C_j. The result is a cluster. One can cut the tree at any level to produce different clustering. A little thought reveals that the nearest clusters are not well-defined, since we only have a distance measure $d(x, x')$ between items. This is where the variations come in:

$$d(C_i, C_j) = \min x \epsilon C_i, x' \epsilon C_j \, d(x, x')$$

This is known as *single-linkage*. It is identical to the minimum spanning tree algorithm. Single can set a threshold and stop clustering once the distance between clusters is above the threshold. Single-linkage tends to produce long and skinny clusters.

$$d(C_i, C_j) = \max x \epsilon C_i, x' \epsilon C_j \, d(x, x').$$

This is known as *complete-linkage*. Clusters tend to be compact and roughly equal in diameter.

$$d(C_i, C_j) = \frac{\sum x \epsilon C_i, x' \in C_j d(x, x')}{|C_i| \times |C_j|}$$

Step-9: The result is the *average* Euclidean distance between items. Somewhere in between single-linkage and complete-linkage and a million other ways you can think of.

Step-10: Cluster formation followed after the incorporation of Squared Euclidean distance

IV. EHAC WITH FARTHEST FIRST CLUSTERING ALGORITHM

EHAC with Farthest First algorithm has identical procedure as k-means, this also chooses centroids and assign the objects in cluster but with max distance and initial seeds are value which is at largest distance to the mean of values, here cluster assignment is different, at initial cluster we get link with high Session Count, like at cluster-0 more than in cluster-1, and so on. EHAC with Farthest First algorithm need less changes and simple working as defined here, it also defines initial seeds and then on basis of "k" number of cluster which we need to know prior. In EHAC with Farthest First algorithm it takes point P_i then chooses next point P_i which is at maximum distance.

P_i is centroid and p_1, p_2, ...,p_n are points or objects of dataset belongs to cluster from equation given below.

$$min\{max\ dist(p1, p2), max\ dist(p1, p2) \ldots.\}$$

EHAC with Farthest First algorithm actually solves problem of k-centre and it is very efficient for large set of data. In EHAC with Farthest First algorithm we are not finding mean for calculating centroid, it takes centrod arbitrary and distance of one centroid from other is maximum Figure-1 shows cluster assignment using EHAC with Farthest First. When we performed outlier detection for our dataset we get which objects is outlier as shown in figure 1.

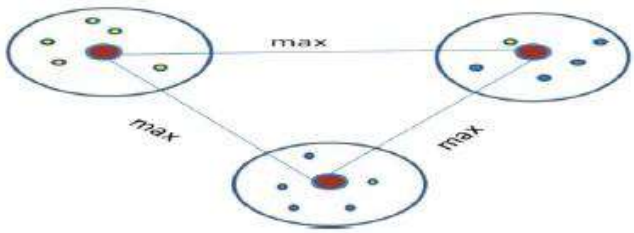

Fig.1: Object assignment in cluster

Initially the algorithm loads the training set to classify the instances that are in the form of .arff file. Here we use Manhattan distance metric formula to calculate the distance of the subsets. The Manhattan distance function computes the distance for travel one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components.

The formula for this distance between a point $X=(X_1, X_2$, etc.) and a point $Y= (Y_1, Y_2$, etc.) is:

$$D(x,y) = \sum_{i=1}^{n} | X_i - Y_i|$$

Where n → is the number of variables, and *Xi* and *Yi* → are the values of the i^{th} variable, at points *X* and *Y* respectively.

Through the dimensionality reduction process it divides training set into sub sets. If the distance is greater than 55 then the concerned instance is belonging to the same group otherwise these belong to another. This entire process continues up to completion of available attributes in training set. On each subset we can apply Enhanced Agglomerative Hierarchical Clustering with Farthest First Algorithm for better performance in effective clustering of data set. If the given instance examples are positive, this algorithm returns single node tree with positive value of root element. If all the examples are negative, then this algorithm returns a single node, i.e. tree with negative value labeled root node. If the given data set instances majority are empty, then this algorithm returns single node tree with common value of majority nodes. All of these cases are not satisfied then we have to calculate entropy for a given set of instances. The entropy is calculated through following equation.

$$Entropy\ (S) = \sum_{i=1}^{c} P_i\ log_2 P_i$$

Where P_i → is the proportion of S belonging to class i. Log2 → is log base 2. Note that S is not an attribute but the entire sample set and c represents total number of samples.

If the entropy is zero it returns root node subset containing records having the same value of all available categorical attributes. If the entropy is non-zero, then we have to calculate information gain for each attribute left to find the maximum gain and create child nodes for this node. This entire process repeats for available individual nodes until the entropy value becomes zero or if we successfully reach the leaf node.

V. IMPLEMENTATION AND RESULTS

In this experiment we use Weka 3.8.1 and Window 7 ultimate to evaluate the Enhanced Agglomerative Hierarchical Clustering Algorithm with Farthest First Clustering for generating decision making approach using respective number of clusters. Weka is machine learning/data mining software written in Java language and it is an open source. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for developing new machine learning schemes. It can be used for Pre-processing, Classification, Clustering, Association and Visualization. Data Set is taken for this algorithm; the input data set is an integral part of data mining application. The data used in our experiment is either real world data obtained from UCI machine learning repository or widely accepted data set available in Weka toolkit. Sensor Discrimination data set comprises 2212 instances and 13 attributes in the area wireless communication while some of them contain missing value.

Table 5.1: Performance of Clusterer using Training Set as Cluster Mode

Clustering Techniques	Total Number of instances in dataset	Number of Clustered Instances	Number of Clusters formed	Percentage of Clustering
Enhanced Agglomerative Hierarchical Clustering	2212	2151 61	Cluster 0 Cluster 1	97% 3%
EAHC with Farthest First Clustering	2212	2212	Cluster 0	100%

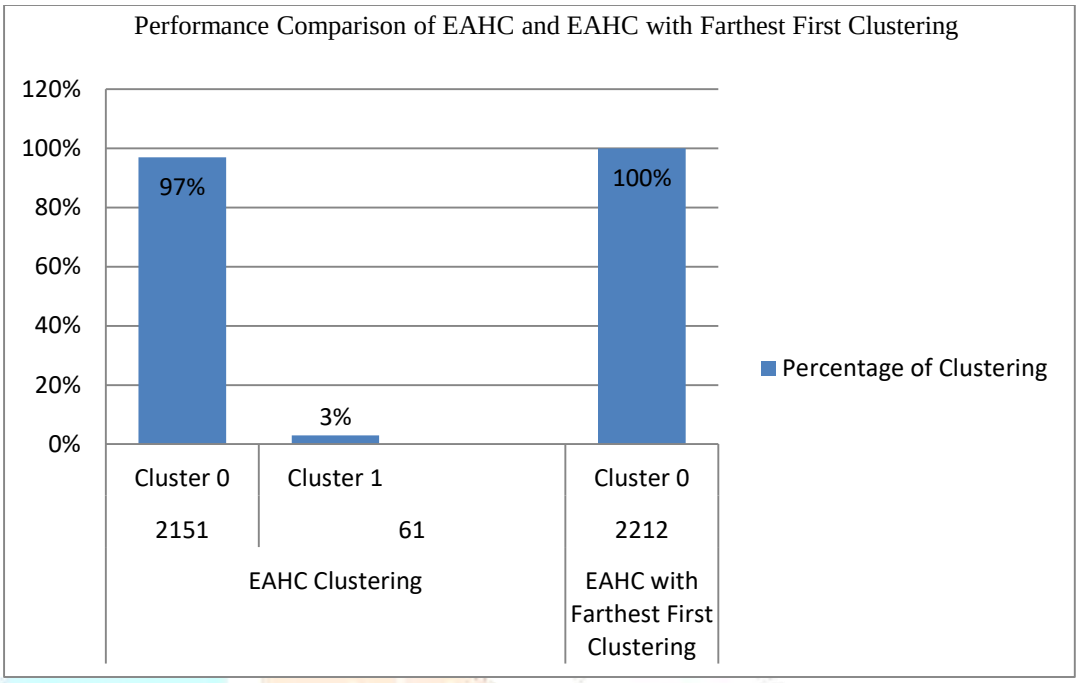

Fig.2: Performance Comparison of EAHC and EAHC with Farthest First Clustering

Table 5.2: Performance of Clusterer using Percentage Split as Cluster Mode

Clustering Techniques	Percentage Split 33%	Percentage Split 66%	Percentage Split 99%
Enhanced Agglomerative Hierarchical Clustering(EAHC)	97% 3%	97% 3%	100%
EAHC with Farthest First Clustering	100%	100%	100%

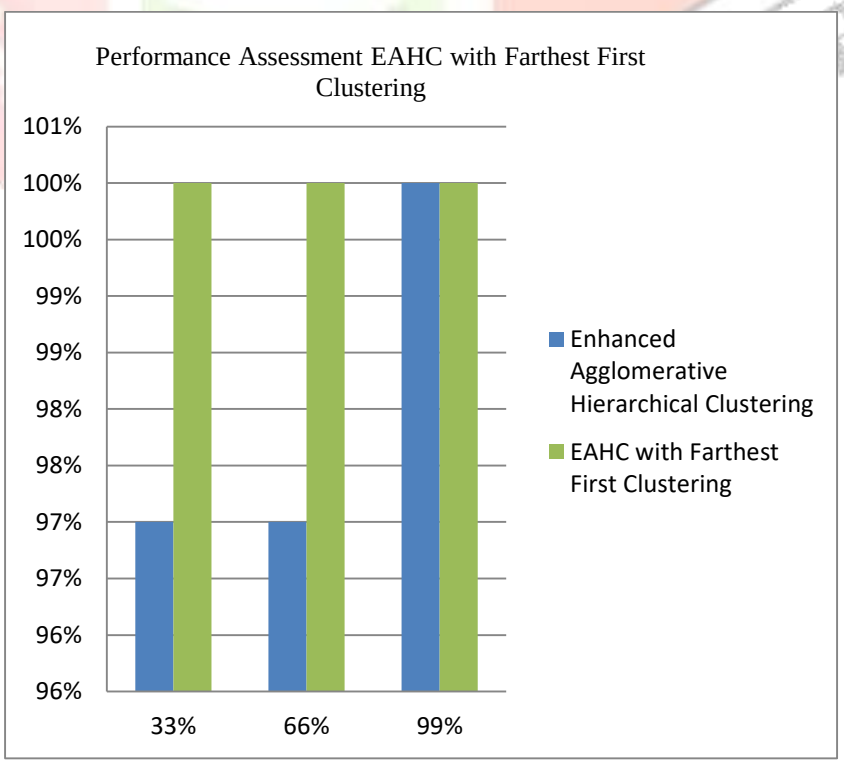

Fig.3: Performance Assessment of EAHC with Farthest First Clustering

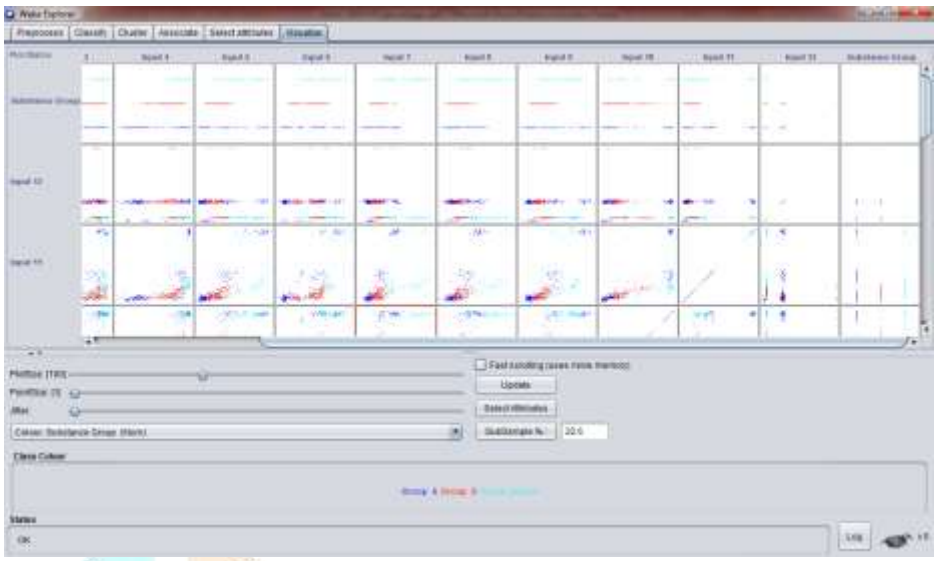

Fig.4: Visualization of attributes with Cluster Assignments of Sensor Discrimination dataset

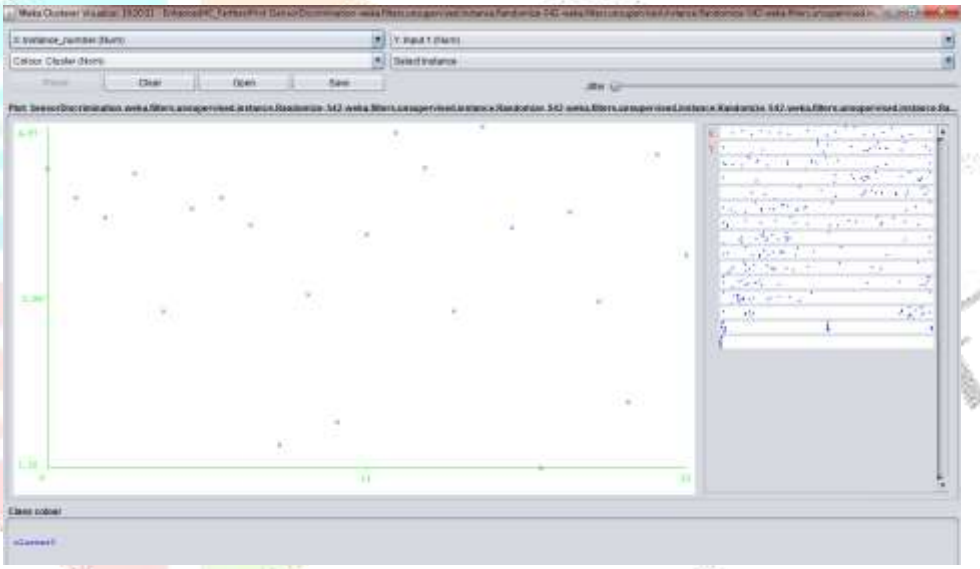

Fig.5: Visualize Cluster Assignments using EAHC with Farthest First Clustering

VI. CONCLUSION

Data mining is an application oriented technology and is having wide applications in many fields. It also estimates, integrates and motives to guide the solution of practical problems and discover the connections between events. In my research work we proposed and developed an innovative algorithm called Enhanced Agglomerative Hierarchical Clustering (EAHC) with Farthest First Clustering algorithm is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The process of Enhanced Agglomerative Hierarchical Clustering starts with these single observation clusters and gradually combines pairs of clusters, forming smaller numbers of clusters that contain more observations. Then clusters successively merged until the desired cluster structure is obtained. In this assessment we expected the performance of Enhanced Agglomerative Hierarchical Clustering and our proposed EAHC with Farthest First Clustering algorithm using Sensor Discrimination dataset from the UCI Machine Learning Repository. The proposed method gives admirable performance when compared the results with other algorithms. In future work we will develop, integrate and embed the EAHC with Farthest First Clustering algorithm with another utmost classification or clustering algorithms in data mining system to apply on real world datasets from the UCI-Machine Learning Repository. It gives tremendous performance when compared to other clustering algorithms in the data mining system.

REFERENCES

[1]. Chim H and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229 Sept. 2008.

[2]. Chee Keong Chan, Duc Thang Nguyen, Lihui Chen and Senior Member, IEEE, "Clustering with Multi viewpoint-Based Similarity Measure", IEEE transactions on E. Hernandez and J. Recasens, "A general framework for induction of decision trees under uncertainty", Modelling with Words, LNAI 2873, pp.26–43, Springer-Verlag, 2003.

[3]. Gaurav L. Agrawal, & Prof. Hitesh Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013.

[4]. IndiraPriya P, Ghosh DK. A survey on different clustering algorithms in data mining technique. International Journal of Modern Engineering Research, 3(1):267– 74, 2013.

[5]. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", the Morgan Kaufmann/Elsevier India, 2006.

[6]. Joun Mack.,"An Efficient k-Means Clustering Algorithm, Analysis and Implementation", IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002.

[7]. Kantardzic, M.," Data Mining: Concepts, Models, Methods and Algorithms", Wiley-IEEE Press, 2011.

[8]. K.Thulasiram, Dr. S. Ramakrishna, Dr. M. Jayakameswaraiah, "To Assess the Performance of EAHC Algorithm Using Sensor Discrimination Dataset for the Improvement of Data Mining System", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3, Issue 1, pp.1141-1146, January-February.2018.

[9]. Kilian Q.Weinberger, Lawrence K. Saul, "Distance Metric Learning for Large Margin nearest Neighbor Classification", Journal of Machine Learning Research, 207-244, 2009.

[10]. L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu, "A fast divisive clustering algorithm using an improved discrete particle swarm optimizer", Pattern Recognition Letters, 31, pp. 1216-1225, 2010.

[11]. Mehmed Kantardzic, Jozef Zurada," Next Generation of Data-Mining Applications", New York：Wiley-IEEE Press, 3, 2005.

[12]. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "A survey of multi objective evolutionary algorithms for data mining: part I," IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, pp. 4–19, 2014.

[13]. M.Jayakameswaraiah and S.Ramakrishna, "A Study on Prediction Performance of Some Data Mining Algorithms", International Journal of Advanced Research in Computer Science and Management Studies, Volume 2, Issue 10, ISSN: 2321-7782, October 2014.

[14]. Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo, "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science,5,3, pp.229- 240, 2012.

[15]. N.Elgendy and A.Elragal, "Big data analytics: a literature review paper," in Advances in Data Mining. Applications and Theoretical Aspects, vol. 8557 of Lecture Notes in Computer Science, pp. 214– 227, Springer, Cham, Switzerland, 2014.

[16]. Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517– 530, 2005.

[17]. UCI Machine Learning Repository – [http://mlearn.ics.uci.edu/databases]

[18]. Verma M, Srivastava M, Chack N, Diswar AK, Gupta N. A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications (IJERA); 2(3):1379–84, 2012.

[19]. Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K.C. Wong, Fellow, IEEE, and Yang Wang, Member, IEEE ,"Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", Sep. 15, 2004.

[20]. WEKA Software, the University of Waikato. [http://www.cs.waikato.ac.nz/ml/weka]

[21]. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.