Computation Accuracy of Hierarchical and Expectation Maximization Clustering Algorithms for the Improvement of Data Mining System

Dr.M.Jayakameswaraiah[1], Mr.M.Veeresh Babu[2], Dr.S.Ramakrishna[3] ,Mrs.P.Yamuna[4]

[1]Assistant Professor, Department of Computer Applications, Madanapalle Institute of Technology & Science, Madanapalle, Chittoor District, Andhra Pradesh, India, drjayakameswar@gmail.com
[2]Assistant Professor, Department of Computer Science Engineering, Sir Vishveshwaraiah Institute of Science & Technology, Madanapalle, Chittoor District, Andhra Pradesh, India, veeruchandra@gmail.com
[3] Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India, drsramakrishna@yahoo.com
[4]M.Tech Student, Department of EEE, Madanapalle Institute of Technology & Science, Madanapalle, Chittoor District, Andhra Pradesh, India, yamunapadigala@gmail.com

---***---

Abstract: *Several aspects of data mining have been investigated in several related fields. Although the database technologists have been seeking efficient way of storing, retrieving and manipulating data, the machine learning communities have focused on developing techniques for learning and acquiring knowledge from the data. The demand for grouping the important data and mine the optimistic information from data is increased. Clustering is the distribution of data into groups of identical objects which has affinity within the cluster and disparity with the objects in the other groups. Characterizing data into a less number of clusters will definitely lead to a loss in some details but data will be interpreted. It represents data objects by less numbers of clusters and thus, it models, data by using its own clusters. Analysis of clustering is the arrangement of a set of patterns into clusters based on similarity. Patterns within the same cluster are closely related than to the data in the adjacent clusters. In this research we evaluated the Hierarchical clusterer with Expectation Maximization Clusterer using Shannon entropy and compared using GPS Trajectory dataset to get better performance and results.*

Key Words: Data Mining, Clustering, Hierarchical Clustering, EM Algorithm.

1. INTRODUCTION

Data mining involves an integration of techniques from multiple disciplines such as data warehouse, machine learning, neural networks, information recovery, data visualization, pattern recognition, image and signal processing and spatial or temporal data study. That is, importance is placed on the capable and scalable data mining techniques. For an algorithm to designate scalable, its running time should be developed almost linear in proportion to the volume of the data, given the existing system resources such as the main memory and disk space. Data mining can be performed with motivating knowledge regularities, the sophisticated information can be extracted from databases and viewed or browsed from dissimilar angles. Clustering is an important analysis tool in many fields, such as pattern recognition, image classification, biological sciences, marketing, city-planning, document retrievals, etc. Hierarchical clustering is one of the most widely used clustering methods. At present, several existing clustering algorithms focus on combination of the advantages of hierarchical and partitioning clustering algorithms[2,7].

2. LITERATURE REVIEW

The main methods of data mining involve with classification and prediction, clustering, sequence analysis, outlier detection, association rules time series analysis and text mining. Among these methods, clustering is considered as among the widely and intensively studied by many data mining researchers.

2.1 Steps Involved in Knowledge Discovery:

Many people treat data mining as a synonym for one more widely used term, Knowledge Discovery from Data (KDD). On the other hand, data mining can be viewed simply as an essential step in the process of knowledge discovery[7]. The Knowledge Discovery consists of an iterative sequence of the following steps:

1. Data cleaning: Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistency in information. Data cleaning involves transformation to accurate the incorrect data. Data cleaning is performed as data preprocessing step before preparing the data for a data warehouse.

2. Data integration: Data Integration is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store. Data integration may absorb unpredictable data therefore needs data cleaning.

3. Data selection: Data Selection is the process where data relevant to the analysis task is retrieved from the database. Occasionally data transformation and consolidation are performed prior to data selection procedure[4,6].

4. Data transformation: In this step data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

5. Data mining: In this step intelligent methods are applied in order to extract data patterns.

6. Pattern evaluation: In this step, the data patterns are estimated.

7. Knowledge presentation: In this step, the knowledge is represented.

2.2 Data Mining Functionalities:

Classification is the process of finding a model that elaborates the data classes or concepts. The principle is the ability to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of a set of training data[9,11]. The resulting model can be represented in the following forms

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

1. **Classification:** It predicts the class of objects whose class label is unidentified. Its objective is to discover a derived model that describes and distinguishes data module or concepts. The resulting model is based on the analysis of a set of training data i.e. the data object whose class label is identified.
2. **Prediction:** It is used to predict missing or unavailable numerical data values rather than class labels. Regression analysis is commonly used for the prediction. Prediction can also be used for the identification of distribution trends based on available data.
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. The Outliers may be defined as the data objects that do not comply with general behavior or model of the data available.
4. **Evolution Analysis:** Evolution Analysis refer to description and model regularities or trend for objects whose behavior changes over time.

3. IMPLEMENTATION

3.1 Hierarchical Clustering Algorithms

Hierarchical clustering is a technique of cluster analysis to present clusters in hierarchy manner. Most of the distinctive methods are not able to make clusters rearrangement or alteration after merging or splitting process. As a result, if the merging processes of objects have problems, it might create the low quality of clusters. One of the solutions is by integrating the cluster with various clusters using a few different methods.

(a) Clustering Using Representatives Algorithm:

Clustering Using Representatives (CURE) algorithm that utilizes various representative points for every cluster. CURE is a type of class-conscious bunch algorithmic regulation that requires dataset to be partitioned. A combination of sampling and partitioning is applied as a policy to deal with huge information. A random sample from the dataset is partitioned to be component of the clusters. CURE first partitions the random sample and then partially clusters the data points according to the partition. After remove all outliers, the pre clustered data in each partition is then clustered again to create the ultimate clusters. The clustering algorithm can identify randomly produced clusters. The algorithm is strong to the identify the outliers and the algorithm uses space that is linear in the key size n and has a worst-case time complexity of $O(n^2 \log n)$. The clusters created by CURE are also superior than the other algorithms[12,13].

There are two fundamental approaches to generate a hierarchical clustering:

Agglomerative: Begin with the points as entity clusters and, at every step, merge the neighboring pair of clusters. This requires defining the concept of cluster proximity.

Divisive: Begin with one, all-inclusive cluster and, at every step, split a cluster until only singleton clusters of individual points remain. In this case, we want to come to a decision which cluster to split at each step and how to do the splitting.

(b) Agglomerative Hierarchical Clustering Algorithm:

Various agglomerative hierarchical clustering techniques are variations on a particular approach: Starting with individual points as clusters, consecutively merge two clusters until only one cluster remains. This approach is articulated more properly in Algorithm

Step-1: Compute the proximity graph, if necessary.
Step-2: repeat
Step-3: Merge the neighboring two clusters.
Step-4: Update the proximity matrix to reproduce the proximity between the new cluster and the original clusters.
Step-5: until Only one cluster remains

3.2 Improved Expectation Maximization Algorithm

Although EM and its variants have been widely used for learning mixture models, several researchers have approached the problem by identifying innovative techniques that give excellent initialization. Other standard techniques like deterministic annealing, genetic algorithms have been applied to obtain a good set of parameters. Though, these techniques have asymptotic guarantees, they are extremely time consuming and hence cannot be used for most of the practical applications. In many clustering methods, clusters are often determined by estimating the location and dispersion of different sample groups within a given dataset. Under probabilistic mixture, these estimates are calculated based on maximum likelihood (ML), and solved by expectation-maximization (EM) algorithm. In Gaussian model, for example, location is the mean and dispersion is the covariance matrix. However, if outliers exist in the data, they can have two effects, called "masking" and "swamping", which potentially affect the estimates of these parameters.

Each iteration of the EM algorithm consists of two processes:

The E-step, and the M-step. In the probability, or E-step, the missing data are anticipated given the practical data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the options of terminology.

In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the exact missing data. Convergence is guaranteed since the algorithm is guaranteed to boost the likelihood at every iteration[14].

3.2.1 Derivation for Improvement of the EM Algorithm

Let X be random vector which results from a parameterized family. We wish to find θ such that P(X|θ) is a maximum. This is known as the Maximum Likelihood (ML) estimate for θ. In order to estimate θ, it is distinct to introduce the log likelihood function defined as,

$$L(\theta) = \ln P(X|\theta)\text{..}(1)$$

The likelihood function is measured to be a function of the parameter θ given the data X. Since ln(x) is a strictly increasing function, the value of θ which maximizes P(X|θ) also maximizes L(θ).

The EM algorithm is an iterative procedure for maximizing L(θ). Assume that after the n^{th} iteration the current estimate for θ is given by θ_n. While the objective is to maximize L(θ), we wish to calculate an efficient estimate θ such that,

$$L(\theta) > L(\theta_n)\text{..}(2)$$

Equivalently we want to maximize the difference,

$$L(\theta) - L(\theta_n) = \ln P(X|\theta) - \ln P(X|\theta_n)\text{......}(3)$$

So far, we have not considered any unobserved or missing variables. In problems where such data exist, the EM algorithm provides a natural framework for their inclusion. Alternately, hidden variables may be introduced purely as an artifice for making the maximum likelihood estimation of θ tractable. In this case, it is assumed that knowledge of the hidden variables will make the maximization of the likelihood function easier. Either way, denote the hidden random vector by Z and a given realization by z. The total probability P(X|θ) may be written in terms of the hidden variables z as,

$$P(X|\theta) = X\ z\ P(X|z, \theta)P(z|\theta)\text{..................}(4)$$

We may then rewrite Equation (3)

$$L(\theta) - L(\theta_n) = \ln X\ z\ P(X|z, \theta)P(z|\theta) - \ln P(X|\theta_n)\text{.......}(5)$$

Observe that this expression involves the logarithm of a sum. In equation (2) using Jensen's inequality, it was shown that,

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)\text{..}(6)$$

for constants λi ≥ 0 with $\sum_{i=1}^{n} \lambda_i = 1$ This result may be applied to Equation (5) which involves the logarithm of a sum provided that the constants λi can be identified. Consider letting the constants be of the form $P(z|X, \theta_n)$. Since $P(z|X, \theta_n)$ is a probability measure, we have that $P(z|X, \theta_n) \geq 0$ and that $P\ (z\ P(z|X, \theta_n)) = 1$ as required.

In Equation (6) the expectation and maximization steps are apparent. The EM algorithm thus consists of iterating the:

1. E-step: Verify the conditional expectation $EZ|X,\theta_n$ {lnP(X, z|θ)}

2. M-step: Maximize this expression with respect to θ

At this point it is fair to ask what has been gained specified that we have only traded the maximization of L(θ) for the maximization of $l(\theta|\theta_n)$. The answer lies in the fact that $l(\theta|\theta_n)$ takes into account the unobserved or missing data Z. In the case where we wish to guess these variables the EM algorithms provides a framework for doing so. Also, as alluded to earlier, it may be convenient to introduce such hidden variables so that the maximization of $L(\theta|\theta_n)$ is simplified given knowledge of the hidden variables.

4. WEKA

WEKA is a data mining software developed by the University of Waikato in New Zealand that apparatus data mining algorithms using the JAVA language. WEKA is a milestone in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption. The algorithms are directly to a database. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules; It also includes visualization tools. In this research experiment we use WEKA 3.8 and Window 7 to evaluate the Hierarchical Algorithm and Improved Expectation Maximization Algorithm for generating effective clustering approach using respective parameters using GPS Trajectory Dataset from the UCI Machine Learning Repository. 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. GPS Trajectory data set consists of 163 instances and 10 attributes while some of them contain missing values[15].

4.1 Attribute Information of GPS Trajectory Dataset:

go_track_tracks.csv : A list of trajectories
id_android : it represents the device used to capture the instance;
speed : it represents the average speed (Km/H)
distance : it represent the total distance (Km)
rating : it is an evaluation parameter. Evaluation the traffic is a way to verify the volunteers perception about the traffic during the travel, in other words, if volunteers move to some place and face traffic jam, maybe they will evaluate 'bad'. (3- good, 2- normal, 1- bad).
rating_bus : it is other evaluation parameter. (1 - The amount of people inside the bus is little, 2 - The bus is not crowded, 3- The bus is crowded.
rating_weather : it is another evaluation parameter. (2- sunny, 1- raining).
car_or_bus - (1 - car, 2-bus)
linha : information about the bus that does the pathway

5. RESULTS

Table-1: Accuracy Performance of Algorithms

Name of the Clusterer	**Name of the Data Set**	**Cluster Mode**	**Total Number of Instances**	**Clustered Instances with percentage**	**Number of Cluster instances**	**Time Taken to Build Model**
Hierarchical Clusterer	GPS Trajectory	Training Set	163	Cluster 0(53%) Cluster 1(47%)	2	0.04
EM Clusterer	GPS Trajectory	Training Set	163	Cluster 0(18%) Cluster 1(29%) Cluster 2(16%) Cluster 3(37%)	4	2.85

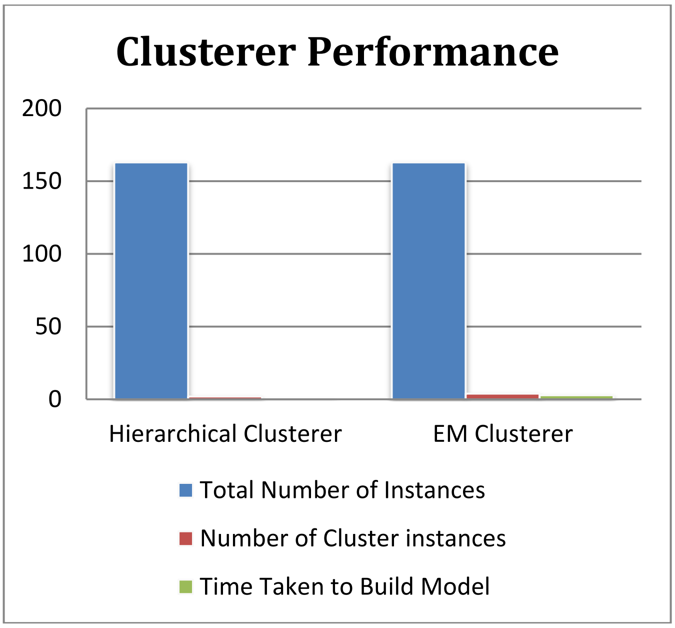
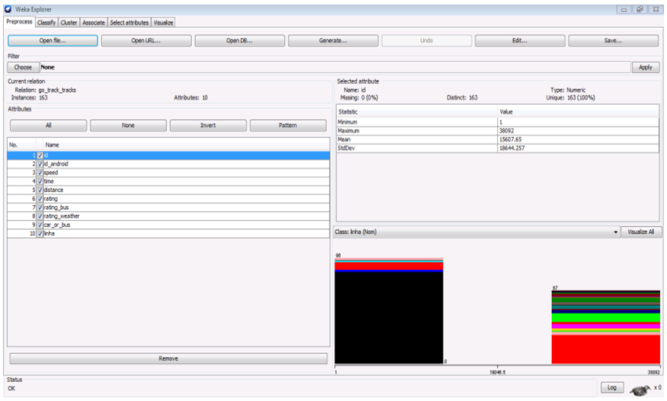


Fig-1:Clusterer Performance of Hierarchical and EM Algorithms



Current relation
Relation: go_track_tracks
Instances: 163 Attributes: 10

Selected attribute
Name: id Type: Numeric
Missing: 0 (0%) Distinct: 163 Unique: 163 (100%)

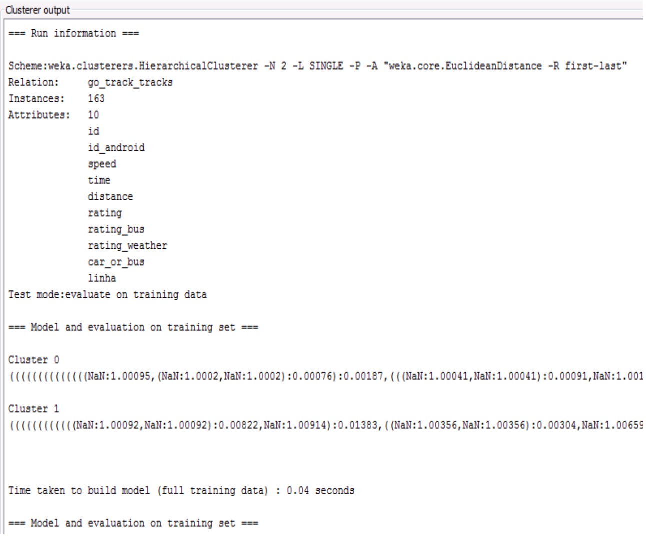
Statistic	Value
Minimum	1
Maximum	38092
Mean	15607.65
StdDev	18644.257

Attributes

No.	Name
1	☑ id
2	☑ id_android
3	☑ speed
4	☑ time
5	☑ distance
6	☑ rating
7	☑ rating_bus
8	☑ rating_weather
9	☑ car_or_bus
10	☑ linha

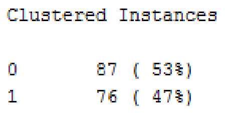
Class: linha (Nom)

Fig-2 : Data preprocessing after deployment of GPS Trajectory Data Set



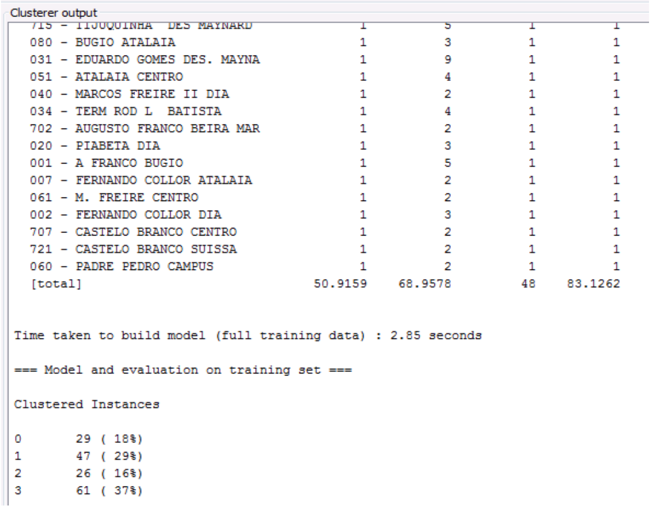
Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     go_track_tracks
Instances:    163
Attributes:   10
              id
              id_android
              speed
              time
              distance
              rating
              rating_bus
              rating_weather
              car_or_bus
              linha
Test mode:evaluate on training data

=== Model and evaluation on training set ===

Cluster 0
((((((((((((((NaN:1.00095,(NaN:1.0002,NaN:1.0002):0.00076):0.00187,(((NaN:1.00041,NaN:1.00041):0.00091,NaN:1.001

Cluster 1
(((((((((((NaN:1.00092,NaN:1.00092):0.00822,NaN:1.00914):0.01383,((NaN:1.00356,NaN:1.00356):0.00304,NaN:1.0065!



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===
```



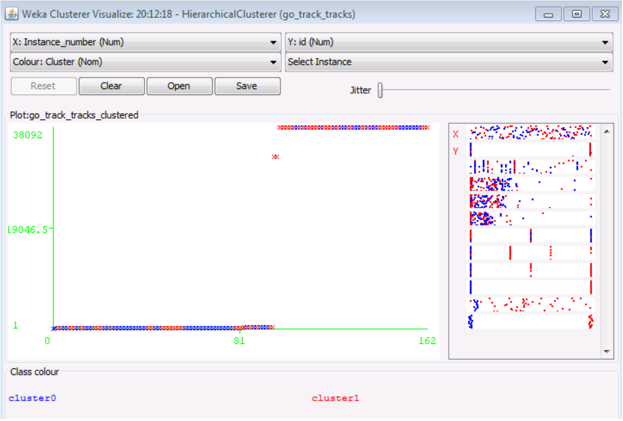
```
Clustered Instances

0       87 ( 53%)
1       76 ( 47%)
```

Fig-3: Cluster Output using Hierarchical Clusterer



```
Clusterer output
  715 - TIJUQUINHA  DES MAYNARD               1         5         1         1
  080 - BUGIO ATALAIA                          1         3         1         1
  031 - EDUARDO GOMES DES. MAYNA               1         9         1         1
  051 - ATALAIA CENTRO                         1         4         1         1
  040 - MARCOS FREIRE II DIA                   1         2         1         1
  034 - TERM ROD L  BATISTA                    1         4         1         1
  702 - AUGUSTO FRANCO BEIRA MAR               1         2         1         1
  020 - PIABETA DIA                            1         3         1         1
  001 - A FRANCO BUGIO                         1         5         1         1
  007 - FERNANDO COLLOR ATALAIA                1         2         1         1
  061 - M. FREIRE CENTRO                       1         2         1         1
  002 - FERNANDO COLLOR DIA                    1         3         1         1
  707 - CASTELO BRANCO CENTRO                  1         2         1         1
  721 - CASTELO BRANCO SUISSA                  1         2         1         1
  060 - PADRE PEDRO CAMPUS                     1         2         1         1
  [total]                                50.9159   68.9578        48   83.1262


Time taken to build model (full training data) : 2.85 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       29 ( 18%)
1       47 ( 29%)
2       26 ( 16%)
3       61 ( 37%)
```

Fig-4: Cluster Output using EM Clusterer

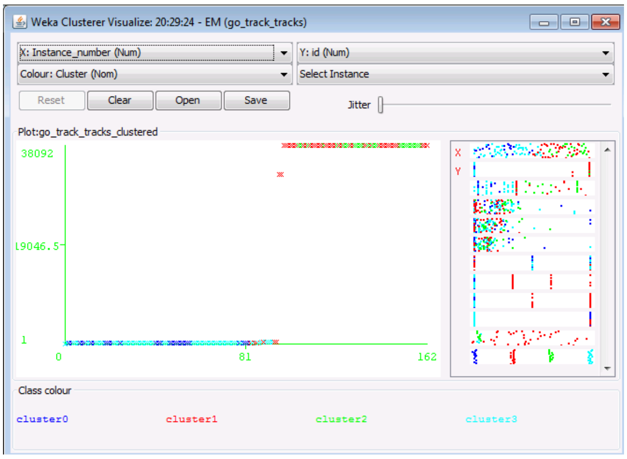


Weka Clusterer Visualize: 20:12:18 - HierarchicalClusterer (go_track_tracks)

X: Instance_number (Num)
Y: id (Num)
Colour: Cluster (Nom)
Select Instance

Plot:go_track_tracks_clustered

Fig-5: Visualization of Hierarchical Cluster Assignments



Weka Clusterer Visualize: 20:29:24 - EM (go_track_tracks)

X: Instance_number (Num)
Y: id (Num)
Colour: Cluster (Nom)
Select Instance

Plot:go_track_tracks_clustered

Fig-6: Visualization of EM Cluster Assignments

6. CONCLUSION

Clustering is the process of grouping objects and data into groups of clusters to ensure that data objects from the same cluster are identical to each other. Data mining is a wide area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms embedded in these fields to execute different data analysis tasks. In this research we compared the clustering results of the Hierarchical algorithm and the our derived and Improved Expectation Maximization algorithm. The results shows that the Improved EM algorithm takes good performance to cluster GPS Trajectory data set and also it gives better accuracy. The specific approaches for clustering are characterized, we developed the WEKA method is based on choosing the file and selecting attributes to convert .csv file to flat file and discussed features of WEKA performance. It is expected that, the state of the art of Improved EM clustering algorithm will help the interested researchers to put forward in proposing more robust and scalable algorithms on different real world datasets in the near future.

REFERENCES

[1]. Baowei Song., Chunxue Wei. “Algorithm of Constructing Decision Tree Based on Rough Set Theory.” International Conference on Computer and Communication Technologies Agriculture Engineering, IEEE (2010).

[2]. Ganti, V., Gehrke, J., and Ramakrishnan, R. “Mining Very Large Databases.” IEEE Computer, Special issue on Data Mining (1999).

[3]. H. Mark, et al. “The WEKA data mining software: an update.” ACM SIGKDD Explorations Newsletter 11.1:10-18 (2009).

[4]. H. Huang, Y. Gao, K. Chiew, K, L. Chen and Q. He, “Towards Effective and Efficient Mining of Arbitrary Shaped Clusters,” IEEE 30th ICDE Conference, pp. 28-39, (2014).

[5]. J. Han., M. Kamber. “Data Mining: Concepts and Techniques (2nd edition).” Morgan Kaufmann Publishers (2006).

[6]. Krakovsky, R., R. Forgac. “Neural network approach to multidimensional data classification via clustering.” Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium, 169–174(2011).

[7]. K. Ding, C. Huo, Y. Xu, Z. Zhong, and C. Pan, “ Sparse hierarchal clustering for VHR image change detection,” Geoscience and Remote Sensing Letters, IEEE, 12 (3), pp. 577 – 581,(2015).

[8]. Mehmed Kantardzic., Jozef Zurada. “Next Generation of Data-Mining Applications.” New York : Wiley-IEEE Press 3 (2005).

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

[10]. Q. Wang., Y. Wu., J. Xiao., and G. Pan. "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics." Automation and Logistics," presented at 2007 IEEE International Conference (2007).

[11]. R.T. Ng, and J. Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining," IEEE Transactions on Knowledge and Data Engineering, 14 (5), pp. 1003-1016, (2005).

[12]. S. Nassar, J. Sander, and C. Cheng. "Incremental and Effective Data Summerization for Dynamic Hierarchical Clustering". In Proc. ACM SIGMOD, (2004).

[13]. Saurav jyoti Sarmah , Dhruba K. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May (2010).

[14]. Tsai, Cheng-Fa, and Chun-Yi Sung. "DBSCALE: An Efficient Density-based clustering algorithm for data Mining in large databases." Circuits, Communications and System (PACCS), 2010Second Pacific-Asia Conference on. Vol. 1. IEEE, (2010).

[15]. WEKA Software. "The University of Waikato". [http://www.cs.waikato.ac.nz/ml/weka].

[16]. Wei Peng., Juhua Chen., and Haiping Zhou. "An Implementation of ID3 - Decision Tree Learning Algorithm" (2012).

[17]. Weiguo Yi., Jing Duan, Mingyu Lu. "Optimization of Decision Tree Based on Variable Precision Rough Set." International Conference on Artificial Intelligence and Computational Intelli gence IEEE (2010).

[18]. Yinghua Lv, TinghuaiMa, MeiliTang, JieCao, YuanTian ,Abdullah Al-Dhelaan , MznahAl-Rodhaan, "An efficient and scalable density-based clustering algorithm for datasets with complex structures", Elsevier, 23 march (2015).