ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

Performance Evaluation of Learning by Example Techniques over Different Datasets

D.Ramya [1], D.T.V.Dharmajee Rao [2]

Final year M.Tech Student, Department of Computer Science and Engineering, Aditya Institute of Technology & Management(AITAM),Tekkali,Srikakulam,Andhra Pradesh,India[1]

Professor, Department of Computer Science and Engineering, Aditya Institute of Technology & Management(AITAM),Tekkali,Srikakulam,Andhra Pradesh, India[2]

ABSTRACT: The clustering activity is an unsupervised learning observation which coalesce the data into segments. Grouping of data is done by identifying common characteristics that are labeled as similarities among data based on their characteristics. Scheming the Performance of selective clustering algorithms over different chosen data sets are evaluated here. Burst time is a performance parameter chosen in evaluating the performance of various selective clustering based machine learning algorithms. Here the investigational results are represented in a table. In our investigation we also suggest a clustering algorithm that performs quicker over a selected data set with reference to the parameter Burst time.

KEYWORDS: Clustering, Weka, Clustering algorithms, simple K-means, EM, Farthest First, CLOPE, COBWEB, Filtered Clustering, Hierarchical Clustering, Machine learning.

I. INTRODUCTION

Data Mining [1] is a day to day activity which is used to determine hidden relations among the facts. A fact is a real time value that exists in real time. The facts can be a measurable value which is also used as a metric for an activity. In facts the relational values represents the measure such as age, gender, salary, cost amount etc which are real time values. Over the facts data mining is applied. Data mining applications are broadly classified into 2 types- Descriptive and Predictive. The descriptive based applications involve classification, times series analysis etc. The Predictive based applications involve clustering, prediction etc. This paper deals with various clustering based algorithms which are used to cluster the facts into various clusters. Clustering is a data mining application which is used for classifying the tuples into different group of densed clusters. A densed cluster is a group of facts with a common relation set of facts which are available in the original data source. The clustering activity is an unsupervised learning activity which makes different facts based on dynamic measures. In the clustering activity the cluster are formed dynamically. The clusters are not predefined classes; the numbers of clusters formed are also depended upon the relations of the facts. The clustering algorithms are used in many business applications especially in production and marketing areas. In designing a catalog for a company if the company decides to develop different catalogs for different groups of people based on some measures such as age. Gender, occupation, favourite items, location etc., which does not have a set of finite values instead of classification clustering yields better results for these types of problems. This paper is organized as various sections; each section describes about the phenomenal activities of clustering. The sections are discussed below.

II. LITERATURE SURVEY

Data Mining is a continuous process of extracting information from the large volumes of data. In the current trends of information technology the amount of information gathered is rapidly growing to large volumes of data. A very little work is done in this area, thus new tools and systems are required for supporting all kinds of human activities. The systems must have the goals of accuracy, robustness and Security. A brief detail of the existing work is described below: Here the usage of data by its availability involves over a series of different types of applications. We are now availing the methods of data mining such as classification, clustering, Association rules etc. Application in which data

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

mining is applied is vast these are used in criminal, forensic investigations and also in many other Matthew Sparks in his papers proposed a series of successful data mining applications. In all these applications we use the techniques like clustering classification and association rules for providing solutions. In clustering there are various algorithms indira priya and Dr.D.K.Gosha have proposed a series various clustering algorithms which exhibit different nature. The algorithm performance is measured over the time complexity which is called execution time and space complexity followed by the simplicity of the algorithm. We cannot assure all the algorithms yield same results when applied to clustering. Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih have proposed a method in which they have selected a series of classification algorithms and applied them over selected datasets to determine the complexity of classification algorithms. A novel approach was proposed there in which they have used exploratory analysis for analysing the classification algorithms. Revathi and Dr. T .Nalini had also propsed a novel method for evaluating the clustering algorithms [2]. They have process some 2D graphs and a table for visualizing the results. Yaminee S. Patil, M.B.Vaidya both have proposed a technical survey on the clustering algorithms. Tiwari M, and Singh R had also contributed by comparing the k-means and k-medoid algorithms over the iris dataset. Athman Bouguettaya[3]research article for clustering the data online had been a note making research article in identifying the dependencies of various clustering algorithms. R. Davé and R. Krishnapuram, have proposed clustering methods that are tough and used in various real world applications they also specified the requirement of having the robustness of various clustering algorithms [4].

III. CLUSTERING

Clustering is an unsupervised learning activity of division of data into different sets. It can also be considered as a special type of classification [5,6]. It is used mainly for identifying similarities among given data items. For performing clustering activity over various types of data different types of clustering algorithms were proposed. Clustering algorithms perform cluster among the given data. A cluster is a group of similar items defined by their similarity among the data elements belong to same cluster. In clustering there are different methods of measures or metrics for clustering the data. Some of the clustering algorithms which are used rapidly are simple k-means, EM, Clope, Farthest First, Make density cluster, Filter clustering, Hierarchical clustering[5] etc. These are the clustering algorithms that are used to perform clustering over the data. In clustering each and every algorithm has its own measures of similarity so the cluster formed by the different clustering algorithms need not be same. One algorithm may make 5 clusters the other may define only 3 cluster groups. The number of clusters formed by the data depends upon the uniqueness and dissimilarity present in the data and the variables considered as metrics for clustering. Clustering is a tricky activity for a beginner because it has much to interpret. In this paper we use various selective clustering algorithms being selected and applied over different dataset which is discussed in future section.

IV. WEKA

Weka is an open source tool developed by Waikato University New Zealand. It is a collection of various types of machine learning algorithms which are used for data mining tasks. Weka is a GUI tool which is used for developing different new machine learning algorithms and schemes. Weka also has predefined datasets and algorithms in their libraries for performing different data mining based operations. Weka can also be used in Pre-processing the data before performing any operation, Classification, Clustering, Association and Visualization of data [2, 5, 6]. In this paper we use this tool for visualizing the results and applying selective clustering algorithms like simple k-means, EM, farthest first clope, cobweb etc. We also choose different data sets such as Qualitative Bankruptcy [7], iris 2D [8], etc., for performing clustering.

V. CLUSTERING ALGORITHMS

In this paper we have chosen the following clustering algorithms: simple K-means, EM, Farthest First, CLOPE, COBWEB, Make Density Cluster, Filtered Clustering, Hierarchical Clustering

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

A. Simple K-Means

K-means is an iterative clustering algorithm [2, 5], here the items are transferred between the set of clusters until the respective related set is assigned to the data elements of the dataset. It is a partition based clustering algorithm. K-means can also be assumed as a variety of squared error algorithm [2]. Cluster mean of $k_i=\{t_{i1}, t_{i2}, \ldots..,$ time$\}$ can be defined as, $\boldsymbol{m_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij}}$

Algorithm: K-Means Clustering

Input: D= $\{t_1, t2\ldots tm\}$ //set of elements
k //number of desired clusters
Output: K //set of clusters

Procedure:

Assign initial values for means a_1, a2…, ak;
Repeat
Assign each item a_i to the cluster which has the closest mean;
Calculate new mean for each cluster;
Until convergence criteria is met;

In many of the cases, the simple k means clustering algorithm takes more time to form clusters. It is advised not to be considered for large datasets.

B. Farthest First

Farthest first is a sibling of k means clustering algorithm. The FF places the cluster center at the point farther from the present cluster [5]. This point has to be there within the data area. Points that are farther are clustered together first. Farthest first clustering algorithm performs faster the clustering process because of this modification from K-means. In various situations like less reassignment and adjustment are needed for this algorithm.

C. Clope

CLOPE is a clustering algorithm that was applied over a large datasets. This algorithm is very quick and it is also scalable, and yields low response time or results in lesser Burst time when compared to simple k-means over large datasets. For clustering larger datasets CLOPE is quite effective. The algorithm for clope is shown below [2].

Algorithm: Clope

/* Phrase 1 – Initialization */

- **While** not end of the database file
- Read the next transaction <*t*, unknown>;
- Put *t* in an existing cluster or a new cluster c_i that maximize profit;
- Write <*t*,*i*> back to database;

/* Phrase 2 – Iteration */

- **Repeat**
- Rewind the database file;
- Moved=**false**;
- **While** not end of the database file
- Read <*t*, *i*>;
- Move *t* to an existing cluster or new cluster c_j that maximize profit;
- If $c_i \neq c_j$ then
- write <*t*, *j*>;
- Moved=true;
- **Until** not moved;

DOI: 10.15680/IJIRSET.2014.0309025

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

In a clustering C = {C_1, ..., C_k}, the following is assumed as a straightforward definition of the criterion function Profit[2].

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) \times |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^2} \times |C_i|}{\sum_{i=1}^{k} |C_i|}$$

The criterion function profit(c) can be generalized by applying a parametric exponential value *Power r* as[2]

$$Profit_r(C) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|}$$

Here, *r* is considered as a non-negative real number called as *repulsion*, which is used to control the level of intra-cluster similarity [6]. If *r* is larger value, transactions representing the same cluster have to share a larger portion of common Items when compared to other clusters.

D. Cobweb

COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. It incrementally incorporates objects into a classification tree in order to get the highest category utility. And a new class can be created on the fly, which is one of big difference between COBWEB and K-means methods. COBWEB provides merging and splitting of classes based on category utility, this allows COBWEB to be able to do bidirectional search. For example, a merge can undo a previous split. While for K-means, the clustering is usually unidirectional, which means the cluster of a point is determined by the distance to the cluster centre. It might be very sensitive to the outliers in the data [9].

E. Expectation Maximization

EM algorithm is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. an expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an Expectation (E) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and Maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc[9].

F. Make Density Clustering

The cluster in a dense region is a set of points that are separated by low density by which the regions form dense regions which are tight. The Make Density Cluster clustering algorithm is very useful when clusters are not regular. This make density based cluster algorithm can also be used if the data has noise and when there are outliers in the data. The points of same density and present within the respective same areas will be connected while forming clusters [2].

Algorithm: Make Density Clustering

- Compute the ε-neighborhood for all objects in the data space.
- Select a core object CO.
- For all objects co ε CO, add those objects y to CO which are density connected with co. Proceed until no further y are encountered.
- Repeat steps 2 and 3 until all core objects have been processed.

G. Filtered Clustering

The Filtered clustering algorithm is used for filtering the information, data or pattern in the given dataset. Here user supplies keywords or a set of samples that contain relevant information [6]. For every new information that is given,

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

they are now compared against the available filtering profile and the information that is matched to the keywords is presented to the user. Filtering profile can be corrected by the user by providing relevant feedback on the retrieved information. The filtering algorithm as follows:

Algorithm: Filtered Clustering

- Find pre-filtering threshold θ.
- Cluster the pre-filtered set.
- Select clustering threshold σ, on the basis of the keyword and initial relevant document set.
- For each new information or pattern α within the distance θ from the filtering profile[2,6].
-

H. Hierarchical Clustering

The Hierarchical clustering is also called as Connectivity based clustering, which is mainly based on the idea of objects that are being more relative to the nearby objects than to the objects far away. Hierarchical methods are generally classified as Agglomerative and Divisive methods these are depended upon how the hierarchies are formed. The Hierarchical algorithms connect the "objects" and form "clusters" by measuring their distance. A cluster can also be considered as large with the maximum distance required to connect the parts of the cluster. At various distances, many clusters are formed. These algorithms cannot provide a single partitioning in the data set, but they provide an extensive hierarchy of clusters that are merged with each other at particular distance [2].

VI. IMPLEMENTATION

In this paper we use the weka API [10] and selective datasets and clustering algorithms for performing comparative analysis. The measures considered here is burst time i.e. the time taken by an algorithm to cluster a dataset. Then the respective times are compared and the quickest algorithm is represented. The same method is applied over the various datasets is applied and we derive various faster clustering algorithms over the datasets. The experimentation results and implementation results are shown in the next section.

VII. RESULT

By using weka API and real time datasets we are evaluating the performance of various clustering algorithms. These algorithms are applied on five datasets and their individual time complexities for the respective algorithms are shown in the table below:

TABLE I
TIME COMPLEXITY COMPARISON

ALGORITHM	QB	VENDOR	IRIS	VOTE	LABOUR
Simple K-means	254.751	19.734	91.719	101.668	75.192
Expectation Maximization	1762.529	10006.534	420.106	12924.743	627.787
CLOPE	41.523	480.399	89.892	135.393	60.518
COBWEB	49.301	480.399	12.949	203.078	20.391
Farthest First	**6.388**	**3.65**	**5.368**	**14.786**	**6.827**
Filtered cluster	8.809	3.7	7.419	20.537	7.958
Hierarchical cluster	64.924	3.798	32.265	611.611	17.558
Make Density Cluster	8.36	8.556	28.251	28.398	8.047

The above table represent the respective burst times of eight clustering algorithms over five data sets. From all the above burst times we can analyse that the fast computed clustering algorithm is Farthest First in each set. And also the EM algorithm is performing its clustering with high burst time. So we are displaying a chart below with respective burst times of seven clustering algorithms excluding EM over five data sets.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

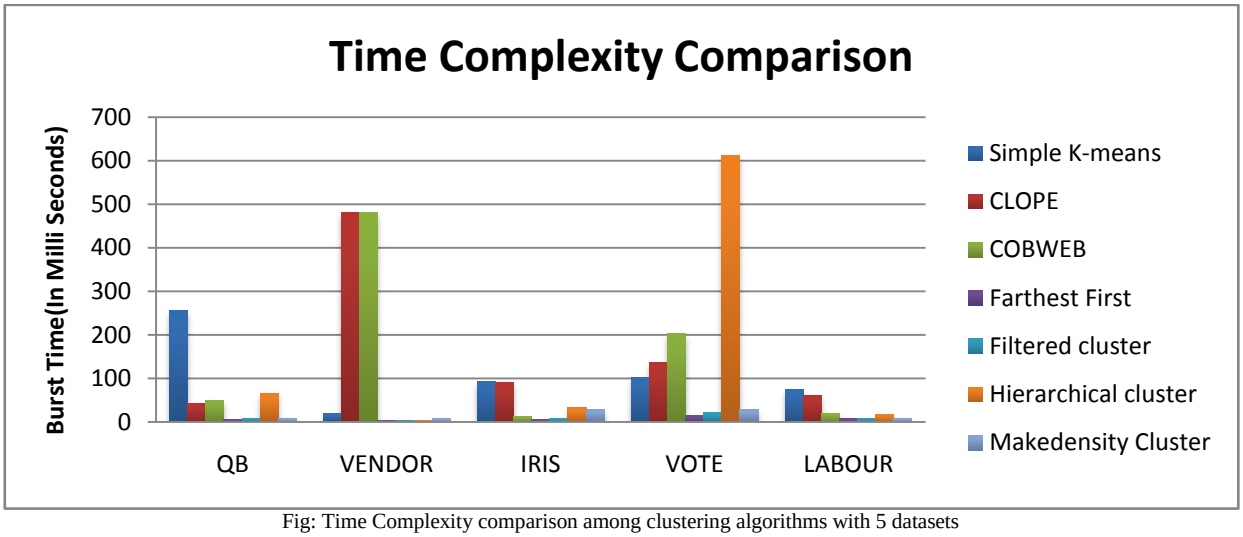

Fig: Time Complexity comparison among clustering algorithms with 5 datasets

VIII. CONCLUSION

In this paper, we observed that all the considered clustering algorithms had performed their clustering over different datasets yield variable results. For example, Cobweb algorithm performed its clustering with minimum burst time over Iris dataset; whereas hierarchical clustering algorithm performed its clustering with minimum burst time over Vendor dataset. The only algorithm which showed consistent result is Farthest First over variety of datasets.

IX. FUTURE WORK

In today's world business intelligence is taking new shape where many new solutions are being provided through clustering. In coming era many new clustering algorithms will be proposed, the behavior of these algorithms can be observed by performing this type of evaluation.

REFERENCES

1. Datamining: concepts and techniques byjiawei han, micheline chamber, Elsevier publication.
2. Yiling Yang, Xud ong Guan, Jinyuan You, CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data. 2002 ACM 1-58113-567-X/02/0007.
3. AthmanBouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996.
4. R. Davé and R. Krishnapuram, "Robust clustering methods: A unified view," IEEE Trans. Fuzzy Syst., vol. 5, no. 2, pp. 270–293, May 1997.
5. Bhoj Raj Sharmaa, Aman Paula, Clustering Algorithms: Study and Performance Evaluation Using Weka Tool, International Journal of Current Engineering and Technology, 2013, ISSN 2277 – 4106.
6. Yiling Yang, Xud ong Guan, Jinyuan You, CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data. 2002 ACM 1-58113-567-X/02/0007.
7. http://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy
8. http://www.cs.waikato.ac.nz/ml/weka/datasets.html
9. Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya,Comparison the various clustering algorithms of weka tools, International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459, Volume 2, Issue 5, May 2012)
10. http://www.cs.waikato.ac.nz/ml/weka
11. Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, Andr´e C. Ponce Leon F. de Carvalho, A Survey of Evolutionary Algorithms for Clustering, ieee transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 39, no. 2, march 2009, pg: 133-151.
12. Tiwari M and Singh R (2012) ,Comparative Investigation of K-Means and K-Medoid Algorithm of IRIS Data, International Journal of Engineering Research and Development, 4: 69-72.