Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Evaluation of Different Clustering Algorithms on Different Datasets

Shivanjli Jain
Research Scholar
Department of Computer Science and Engineering
Baba Banda Singh Bahadur Engineering College,
Fatehgarh Sahib, Punjab, India

Amanjot Kaur
Assistant Professor
Department of Computer Science and Engineering
Baba Banda Singh Bahadur Engineering College,
Fatehgarh Sahib, Punjab, India

Abstract: ***Data mining is the method of evaluating data from various prospective and summarizing into valuable information and it is used to gain revenue, decrease price, or both. It is the method of finding interrelation or patterns through different ranges in huge databases. Data mining is well-known in the science and mathematical fields but also making used progressively by marketers trying to extract useful user data from websites. In clustering, data is being organized in classes. However, class labels are unidentified in clustering and the clustering algorithm to explore acceptable classes. Clustering is also called unsupervised classification. It is not a particular algorithm but a common task is being solved. A comparative review of clustering algorithms is performed here on different data items. The various clustering algorithms are compared on the basis of performance with the different parameters to form the estimated clusters. The various clustering algorithms which are executed are depicted as a graph.***

Keywords: ***Data mining, Clustering, Cobweb, Clope, Filtered Cluster, Farthest First, EM, DBSCAN***

I. INTRODUCTION

As we know that there will be a huge amount of data especially from terabytes to petabytes is used in day to day life but the major problem is that the availability of data. There is expanding the data but due to this knowledge is diminished. So to overcome this, data mining is introduced. Companies spending money to construct data warehouses which contain millions of records and attributes but they are not taking the ROI [22]. Due to scarcity of knowledge, staffs and appropriate tools, companies cannot generate sufficient output. Data mining [4] is the method of unsupervised classification of items based on data patterns obtained from a dataset. Several algorithms are developed and extract the information to implement and explore new knowledge patterns which is useful for decision support. KDD [11] is also referred in data mining.

Data mining consists into six classes of tasks:

- Preprocessing: The recognition of uncommon data records that might be elegant or data inaccuracy which want further review.
- Association rule learning [14]: Finds new relationships among variables. To simplify this, we explain a paradigm that a supermarket collected data of customer in daily purchasing. Using this class of data mining, the supermarket resolves the products which are regularly bought and gathered information which is used in marketing and it is referred as market basket analysis.
- Clustering: is the method to explore the data in groups and structures that are in some manner or another similar way which do not use data having recognized structures.
- Classification: is the method of establishing well known structure to implement to innovative data. To understand this we will take an example, an e-mail program may be attempted which is to classify an e-mail as genuine or as spam.
- Regression: attempts to locate a function which duplicates the data with the smallest amount of error.
- Summarization: is the method to provide the data set in highly demonstrated way which includes visualization and report generation.

II. CLUSTERING

Clustering [8], [13] is a technique used in data mining which is used to set data elements into their interrelated groups with no advancement of knowledge regarding grouping of definitions. It is not a particular algorithm but a common task is being solved. The various algorithms which are used in clustering can accomplish that vary extensively in notion of what cluster comprises and how it quickly locates them. Popular concepts of clusters which contains groups through short intervals along with the cluster components, intense fields of the data spaces in different intervals. Thus clustering can be specified as an accession problem. The relevant algorithm of clustering and the parameters which are used rely on the particular data set and expected results are used. It is not a regular task but a repetitive method of knowledge discovery which is used to share multiple objects involving test and inadequacy. It will be essential for data

preprocessing to modify and different parameters which are used to cluster the objects until it satisfies the desired results. Several algorithms of clustering which are used quickly are EM, Cobweb, Filtered Clustering, Clope, DBSCAN, OPTICS, Farthest First etc. These clustering algorithms [20] that are used to execute 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. The amount of clusters which are produced by the data depends upon the uniqueness and dissimilarity present in the data and the variables considered as metrics for clustering. Clustering is a complicated activity for a learner because it has much to understand. In this research, I have used different clustering algorithms with different parameters and applied over different datasets.

III. METHODOLOGY

My methodology [19] is extremely simple. The past information of the data has been collected from the repositories and implement in clustering. I am applying different clustering algorithms because it is the essential method to manage the large databases and predict a helpful outcome which is useful for the creative users and innovative researchers.

IV. DATASET

To perform the comparison of algorithms of clustering we want the datasets which are used in past. In this research the data which I used are being taken from two data repositories. ISBSG [13] and UCI [17] data repositories gives us the data information which are used in past. They should have different types of nature. These repositories are too useful for the researchers for doing the research. We can directly use this information in clustering algorithms of data mining and conclude the result.

V. CLUSTERING ALGORITHMS

In this paper, I have to choose the six clustering algorithms [18]: EM, Clope, Cobweb, DBSCAN, Farthest first, Filtered cluster.

A. EM:

An expectation-maximization (EM) algorithm [1], [12] is a repetitive method to find maximum possibilities or maximum a posteriori (MAP) parameter estimation of analytical models in which the model refers to unobserved underlying variables. This algorithm performs two steps, first is expectation (E) step and second is maximization (M) step. E step evaluates the expectation of log- likelihood with the use of current estimation of parameters. M step computes the parameters which maximize the expected log-likelihood finds in the E step. Now these parameter estimation are used in determining to distribute of the variables in next E step. The end result of the analysis of cluster is written into band naming class indices. The value refers the class indices, where value '0' indicates to the first cluster, value '1' indicates to the second cluster etc.

Algorithm:

generalized-cluster($X \in R^{n\times d}$, k, m, w, $C \in R^{k\times d}$)
while the set of centers C has not converged do
// expectation step
for i ∈ {1 . . . n} do
for j ∈ {1 . . . k} do
compute membership $m(c_j|x_i)$
compute weight $w(x_i)$
end for
end for
// maximization step
for j ∈ {1 . . . k} do
$c_j = \sum^{n}_{i=1} m(c_j|x_i)w(x_i)x_i / \sum^{n}_{i=1} m(c_j|x_i)w(x_i)$
end for
end while

B. Cobweb:

COBWEB [11] is an incremental system for hierarchical clustering. The observations in cobweb are incrementally arranged into a hierarchical tree. Each nodule in a hierarchical tree personifies a class with a probabilistic approach so that it summarizes the classification of objects distribute the attribute and value in each nodule. This hierarchical tree can be helpful to predict the misplaced attributes or the values of a new object. Cobweb uses an inquisitive estimation measure called class efficiency to convoy the construction of the tree. It incrementally combines objects into a hierarchal tree to acquire the class efficiency. Cobweb handles four essential operations to build the hierarchical tree. The operations are:

- Intermixing two nodules.
- Splitting a nodule.
- Inserting a new nodule.
- Passing an object along the hierarchy.


```
Algorithm:
COBWEB (core, log):
 Input: A COBWEB node core, an instance to insert log
 if core has no children then
   children := {copy(core)} \\ changes root to core and record to log
   newcategory(log) \\ adds child with log's feature values.
   insert(log, core) \\ update core's statistics
 else
   insert(log, core)
   for child in core's children do
     calculate Category Utility for insert(log, child),
     set best2, best3 children w. best Category Utility
   end for
   if newcategory(log) yields best Category Utility then
     newcategory(log)
   else if merge(best2, best3) yields best Category Utility then
     merge(best2, best3)
     COBWEB(core, log)
   else if split(best2) yields best Category Utility then
     split(best2)
     COBWEB(core, log)
   else
     COBWEB(best2, log)
   end if
 end
```

C. DBSCAN:

DBSCAN (Density-based spatial coing into groups of applications with noise) [23], [24] is a measure of space between parts based coing into groups algorithm which is suggested by Martin Ester [11], Hans-Peter Kriegel, Jorg Sander and Xiaowei Xu in 1996. A set of points is specified in some space and it combines together which are closely packed. In this algorithm, density reach and density connect capability approach is used. This approach is referred by two input parameters, first is the size of epsilon which controls the cluster size and neighbourhood size and second is the minpts refers the minimum points of nearest neighbours which is distributed locally. All points which are discovered within the surroundings are added, as it is in their own surrounding when the added points are also dense. This process carries on until the density-connected mass group is completely discovered.

```
Algorithm:
DBSCAN(A, eps, MinPts)
  B = 0
  for each unfrequented point Pt in dataset A
    mark Pt as frequented
    NeighbourPts = regionQuery(Pt, eps)
    if sizeof(NeighbourPts) < MinPts
      mark Pt as NOISE
    else
      B = next cluster
      expandCluster(Pt, NeighbourPts, B, eps, MinPts)
expandCluster(Pt, NeighbourPts, B, eps, MinPts)
  add Pt to cluster B
  for each point Pt' in NeighbourPts
    if Pt' is not frequented
      mark Pt' as frequented
      NeighbourPts' = regionQuery(Pt', eps)
      if sizeof(NeighbourPts') >= MinPts
        NeighbourPts = NeighbourPts joined with NeighbourPts'
    if Pt' is not yet member of any cluster
      add Pt' to cluster B
regionQuery(Pt, eps)
  return all points within Pt's eps-neighbourhood (including Pt)
```

D. Clope:

CLOPE (Clustering with sLOPE) [18], [25] is very rapid and scalable, when clustering is applied over greatly sized business recording knowledge-bases with high measures such as web server computer records. Clope is a clustering algorithm that was used over greatly sized knowledge-bases. This algorithm takes low feedback time or results in lesser burst time when made a comparison to simple k-means algorithm over greatly sized knowledge-bases.

Algorithm:

```
    /* Phrase 1 - Initialization */
while not end of the record file  // choose record instead of database
   read the next transaction ⟨r, unknown⟩;
   put r in an existing cluster or a new cluster Ci that maximize profit;
   write ⟨r, i⟩ back to record;
    /* Phrase 2 – Iteration
 repeat
   rewind the record file;
   moved = false;
   while not end of the record file
      read ⟨r, i⟩;
      move r to an existing cluster or new cluster Cj that maximize profit;
      if Ci ≠ Cj then
          write ⟨r, j⟩;
          moved = true;
until not moved;
```

E. Filtered cluster:

The Filtered clustering algorithm [11] 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. For every new information that is given, 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. This algorithm begins by storing the data points in a kd-tree. In each stage the nearest center to each data point is computed and each center is moved to the centroid of the associated neighbors. The data for each node are filtered as they are propagated to the node's children. Since the kd-tree is computed for the data points rather than for the centers, there is no need to update this structure in each stage.

Algorithm:

Input: Training questions, list of informants, Threshold

1. Do:
2. Classify all training questions to the most related informant in the list
3. For each informant:
4. NCorrect ← The amount of questions of the correct type classified by this informant
5. NIncorrect ← The amount of questions of other types misclassified by this informant
6. If NIncorrect*Threshold > NCorrect, remove informant from the list
7. While informants are removed in step 6

F. Farthest First:

Farthest First algorithm [3], [14] suggested by Hochbaum and Shmoys in 1985. It is a made an adjustment form of k-means algorithm which places each cluster in the middle at the point farther most from the existing cluster center. This point should lies within the data area. So it highly increases the clustering speed in most of the situations when less reassignment and modification is needed. A best possible heuristic for the k-center problem works as a fast simple approximate clustered modeled after simple k-means which might be a useful for it. The valid options for this algorithm are:

N -Specify the amount of clusters to generate.

S -Specify random number seed.

Algorithm:

FARTHESTFIRST(A, k, W)

Input: A: input affinity matrix, k: number of clusters, W: constraint penalty matrix

Output: $\{\pi_c^{(0)}\}_{c=1}^{k}$: initial partitioning of the points

1. M = transitive closure of must-link constraints in W.
2. C_M = connected components in M.
3. CC = components from C_M disconnected across cannot-link boundaries.
4. Set $\pi_1^{(0)}$ = largest connected component from CC; set i = 1.
5. If k>i, go to Step 6. Else, go to Step 8.
6. Find component CC_j that is farthest from the current set of selected components $\{\pi_c^{(0)}\}_{c=1}^{i}$.
7. Set $\pi_{i+1}^{(0)} = CC_j$; set i = i + 1. Go to Step 5.
8. Return $\{\pi_c^{(0)}\}_{c=1}^{k}$

VI. EXPERIMENTAL RESULTS

These algorithms are performed in consideration to measure the performance of different parameters applied over five datasets and their individual parameters for the separate algorithms are made clear by the tables shown below. These results are summarized by the graphs.

EM

Table 1: EM clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model	Number of clusters selected by cross validation	Log likelihood
Audiology	70	226	4	27.73 seconds	4	-15.07045
Breast Cancer	10	286	3	5.4 seconds	3	-9.36546
Vote	17	435	5	24.32 seconds	5	-7.24322
Mushroom	23	8124	14	8613.74 seconds	19	-9.7878
Nursery	9	12960	6	3120.93 seconds	10	-9.71469

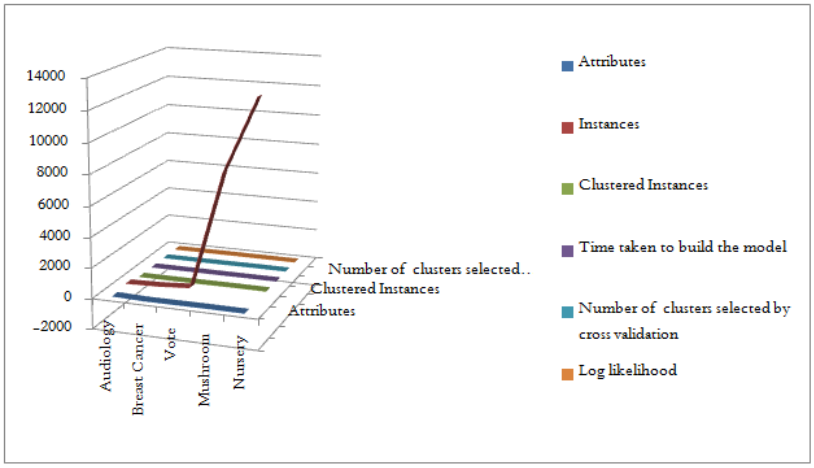

Chart 1: Comparison between parameters of EM clustering algorithm

CLOPE

Table 2: Clope clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model
Audiology	70	226	4	0.14 seconds
Breast Cancer	10	286	57	0.28 seconds
Vote	17	435	14	0.23 seconds
Mushroom	23	8124	23	6.27 seconds
Nursery	9	12960	1440	334.68 seconds

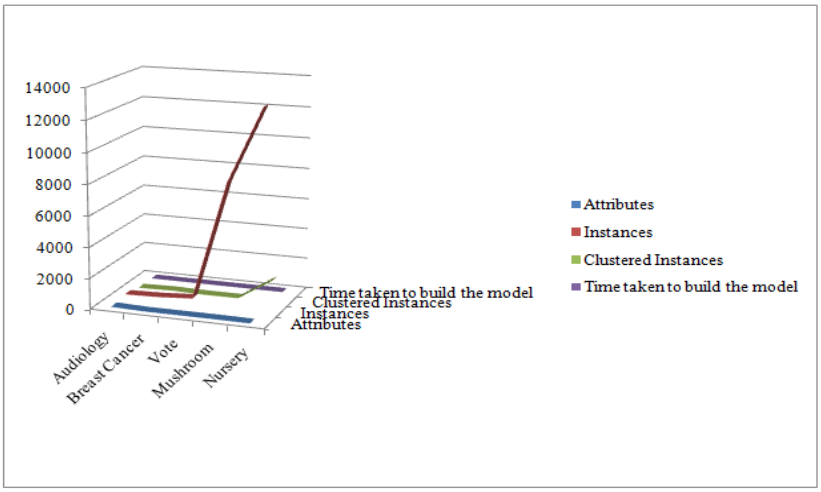

Chart 2: Comparison between parameters of Clope clustering algorithm

COBWEB

Table 3: Cobweb clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model	Number of merges	Number of splits
Audiology	70	226	172	0.92 seconds	90	87

Breast Cancer	10	286	256	0.33 seconds	96	75
Vote	17	435	208	0.3 seconds	104	97
Mushroom	23	8124	7213	76.26 seconds	2608	1927
Nursery	9	12960	11458	61.42 seconds	5039	3946

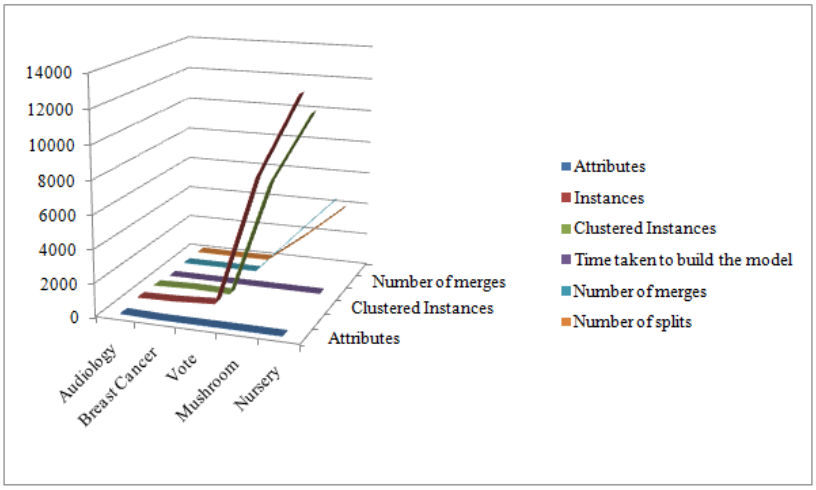

Chart 3: Comparison between parameters of Cobweb clustering algorithm

DBSCAN

Table 4: Dbscan clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model	Epsilon	Minpts
Audiology	70	226	226	0.25 seconds	0.9	6
Breast Cancer	10	286	286	0.08 seconds	0.9	6
Vote	17	435	14	0.18 seconds	0.9	6
Mushroom	23	8124	8124	112.24 seconds	0.9	6
Nursery	9	12960	12960	347.39 seconds	0.9	6

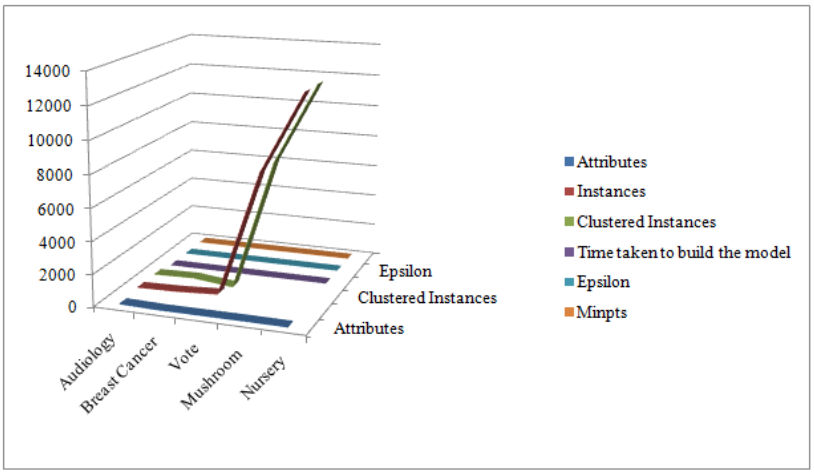

Chart 4: Comparison between parameters of Dbscan clustering algorithm

FARTHEST FIRST

Table 5: Farthest first clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model
Audiology	70	226	2	0.02 seconds
Breast Cancer	10	286	2	0 seconds
Vote	17	435	2	0.01 seconds
Mushroom	23	8124	2	0.06 seconds
Nursery	9	12960	2	0.05 seconds

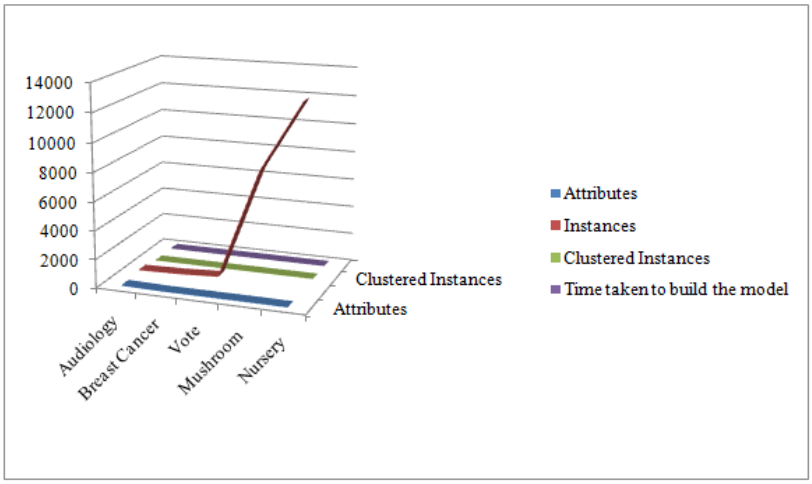

Chart 5: Comparison between parameters of Farthest first clustering algorithm

FILTERED CLUSTER

Table 6: Filtered clustering algorithm

Dataset Name	Attributes	Instances	Clustered Instances	Time taken to build the model	Within cluster sum of squared errors	No. of iterations
Audiology	70	226	2	0.03 seconds	1204.0	3
Breast Cancer	10	286	2	0 seconds	1177.0	3
Vote	17	435	2	0.01 seconds	1510.0	3
Mushroom	23	8124	2	1.14 seconds	61292.0	11
Nursery	9	12960	2	0.28 seconds	65892.0	3

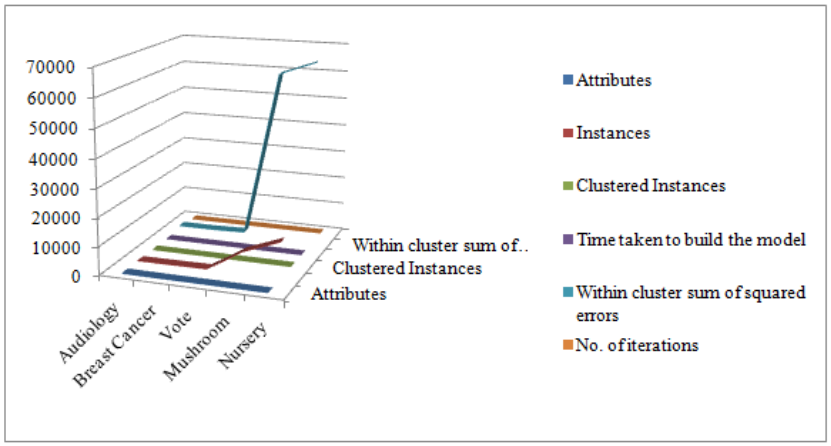

Chart 6: Comparison between parameters of Filtered clustering algorithm

VII. COMPARISON

These results are compared the performance of clustering algorithms according to time taken and clustered instances which is shown by a graph.

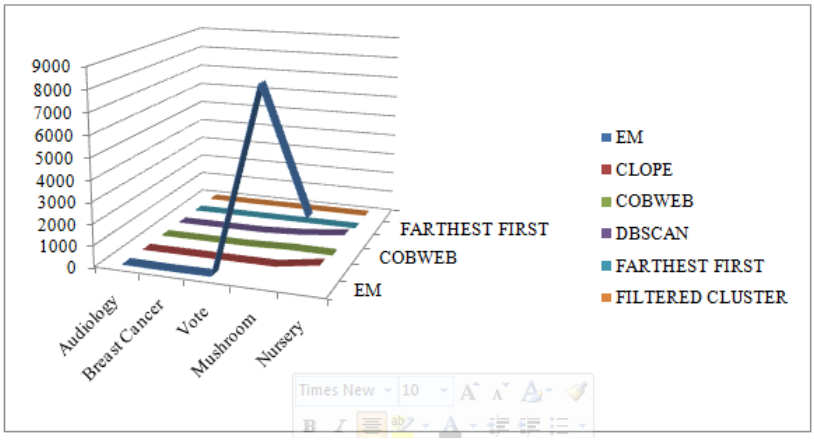

Chart 7: Comparison according to time taken

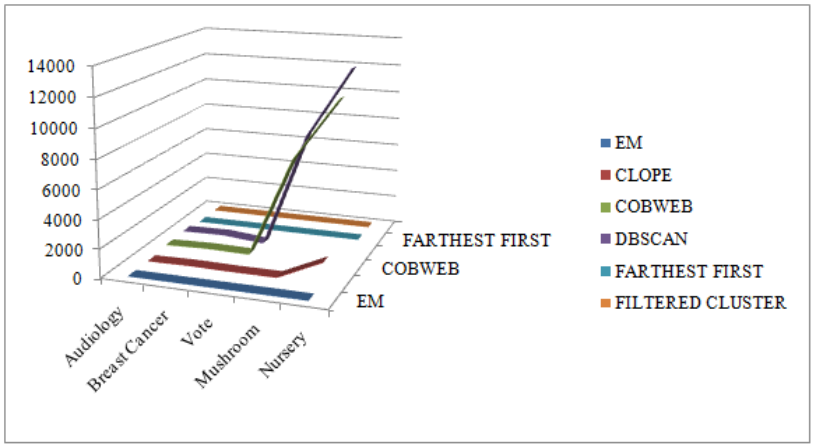

Chart 8: Comparison according to clustered instances

VIII. CONCLUSION & FUTURE WORK

Data mining is wrapping each and every area of our life. Basically, it is useful in many areas that are banking, business, medical, education and so on. The main aim of this paper is to make a comparison among various clustering algorithms on the basis of time taken and number of clusters formed over five datasets. For given datasets, EM algorithm took more time to perform clustering whereas farthest first algorithm took very less time. In case of clustered instances, DBSCAN algorithm formed larger amount of clusters whereas farthest first algorithm and filtered cluster algorithm formed less amount of clusters. So according to time taken, farthest first algorithm is preferred more than other algorithms and according to clustered instances, DBSCAN algorithm is preferred more than other algorithms. According to datasets, breast cancer has less amount of data so it takes less time when it is implemented with farthest first algorithm but according to cluster formation, nursery dataset has large amount of data so it is better when it is implemented with DBSCAN algorithm. In future, other proposed clustering algorithms will be implement and compare to present the results of these algorithms with practical examples.

REFERENCES

[1] Sonam Narwal, Kamaldeep Mintwal," Comparison the Various Clustering and Classification Algorithms of WEKA Tools", ISSN: 2277 128X, Volume 3, Issue 12, December 2013.

[2] Namita Bhan, Deepti Mehrotra, " Comparative Study of EM and K-Means Clustering Techniques in Weka Interface", ISSN No: 2250-3536 Volume 3, Issue 4, July 2013.

[3] Brian Kulis, Sugato Basu, Inderjit Dhillon, Raymond Mooney, (2008) Semi-supervised graph clustering: a kernel approach.

[4] http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques

[5] Mrutyunjaya Panda , Manas Ranjan Patra," A novel classification via clustering method for anomaly based network intrusion detection system", Vol. 2, No. 1, Nov 2009.

[6] Murlidher Mourya, Phani Prasad," An Effective Execution of Diabetes Dataset Using Waikato Environment for Knowledge Analysis or WEKA", Vol. 4 (5), 2013, 681-682.

[7] Minky Jindal, Nisha Kharb, " K-means Clustering Technique on Search Engine Dataset using Data Mining Tool", ISSN 0974-2239, Volume 3, Number 6 (2013), pp. 505-510.

[8] https://en.wikipedia.org/wiki/Cluster_analysis

[9] Mahesh Singh, Anita Rani, Ritu Sharma, (2014) An Optimised Approach for Student's Academic Performance by K-Means Clustering Algorithm Using Weka Interface.

[10] Xiu-yu Zhong, (2011) The research and application of web log mining based on the platform weka.

[11] Richard Khoury, (2012) Sentence Clustering Using Parts-of-Speech.

[12] Pankaj Saxena, Sushma Lehri," Analysis of various clustering algorithms of data mining on health informatics", Volume-4, Issue-2, 2013.

[13] Xindong Wu, 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 (2007) Top 10 algorithms in data mining.

[14] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", ISSN 2250-2459, Volume 2, Issue 5, May 2012.

[15] https://en.wikipedia.org/wiki/Data_mining

[16] Samir K. Sarangi, Vivek Jaglan, (2013) Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques.

[17] Pallavi, Sunila Godara," A Comparative Performance Analysis of Clustering Algorithms", Vol. 1, Issue 3, pp.441-445.

[18] S. Revathi, Dr.T.Nalini," Performance Comparison of Various Clustering Algorithm", Volume 3, Issue 2, February 2013 ISSN: 2277 128X.

[19] D.Ramya, D.T.V.Dharmajee Rao, (2014) Performance Evaluation of Learning by Example Techniques over Different Datasets.

[20] Suman, Pooja Mittal, "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weka tool", Volume 3, Issue 2, March – April 2014.

[21] Garima Sehgal, Dr. Kanwal Garg, "Comparison of Various Clustering Algorithms", Vol. 5 (3) , 2014, 3074-3076.

[22] Vishnu Kumar Goyal, An Experimental Analysis of Clustering Algorithms in Data Mining using Weka Tool.

[23] Swasti Singhal, Monika Jena, (2013) A Study on WEKA Tool for Data Preprocessing, Classification and Clustering.

[24] Aastha Joshi, Rajneet Kaur," A Review: Comparative Study of Various Clustering Techniques in Data Mining", ISSN: 2277 128X , Volume 3, Issue 3, March 2013.

[25] https://en.wikipedia.org/wiki/DBSCAN

[26] Yiling Yang, Xudong Guan, Jinyuan You, (2002) CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data.

