Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 10, October 2015, pg.46 – 60

RESEARCH ARTICLE

Machine Learning Technique for Comparison and Evaluation of Clustering

Anurag Verma
anuragverma434@gmail.com
Department of Computer Science
HP University, Shimla, India

Dr. A.J.Singh
aj.hpucs@gmail.com
Department of Computer Science
HP University, Shimla, India

ABSTRACT: Data explosion in every field whether it is business, pharmaceutical or medical field. Where enormous data storage occurs, name comes Data Warehouse and for piling its mountain down to its knowledgeable form is Data Mining. Association, Classification and Clustering represents basics Data Mining techniques. The recent paper works on the three univariant datasets to finalise the evaluation of clusters on three clustering algorithms and finally draw the accuracy behaviour of these three algorithms.
Keywords: - Data mining, clustering, Machine Learning, K-Means, EM, DBSCAN, Accuracy, Clustering Error

1. INTRODUCTION

Data residing in small chips and hard disks are so much huge that they cannot be easily analysed manually. So to access useful and knowledgeable information , one has to go beyond human's reach and rely on a technique of the computer science, that is ,Data Mining which has been in a research area since a last decade.

2. Data Mining

Data mining is an important field of computer science which involves computational process of large data sets patterns discovery. It has been possible for organizations, with the invention of new data collection and storage technology, to accumulate huge amounts of data at a very low cost [2].

Data mining helps end users extract useful business information from large databases. Data warehousing allows you to build Data Mountain. Data mining allows you to shift that mountain down to the essential information that is useful to your business [1].

Data mining is the process of exploration and analysis, by automatic or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules [5].

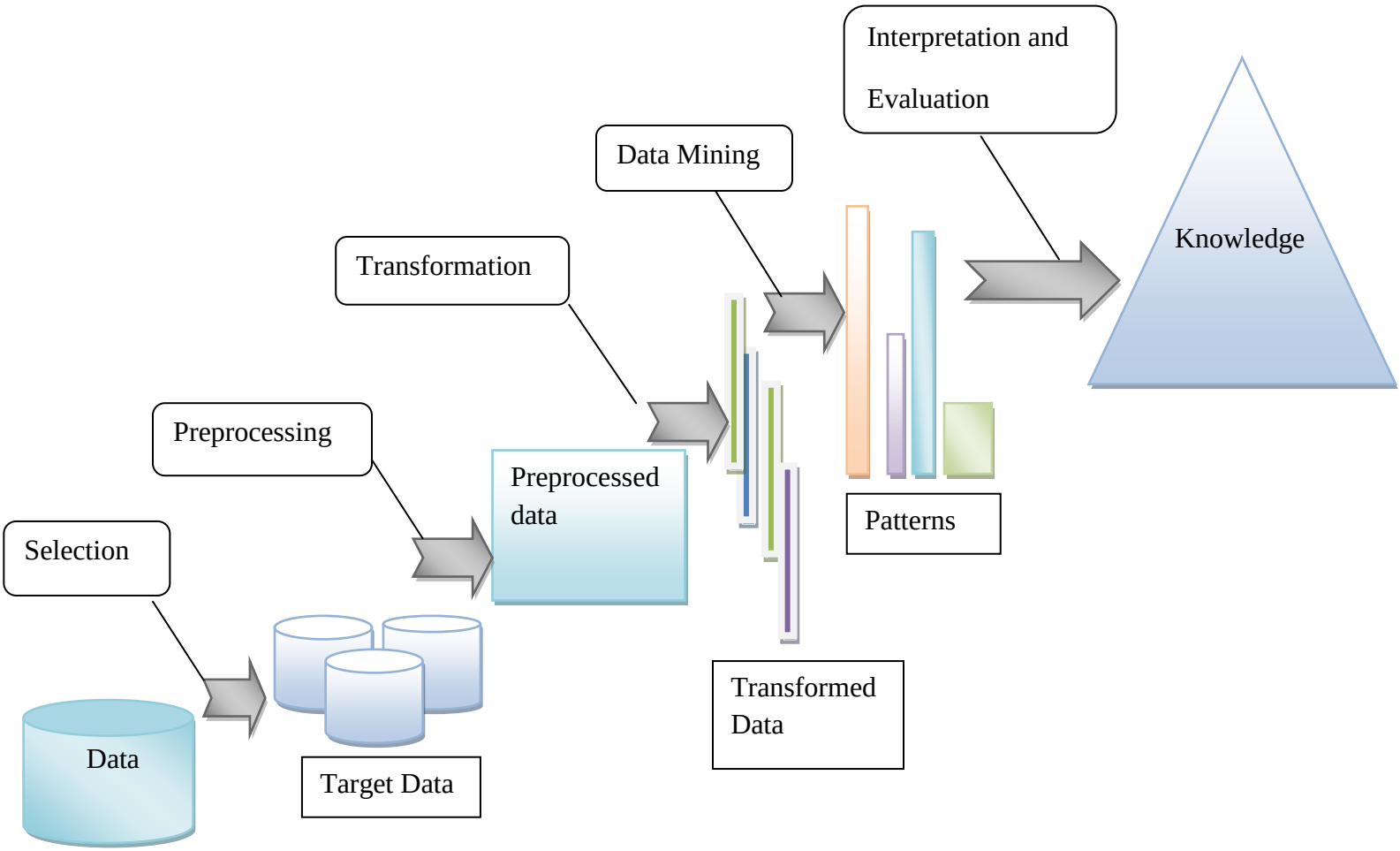

Fig 1.Data Mining Process

3. Clustering

When one make groups on such basis that similar data objects inside one group and these data objects must be dissimilar to other groups, then that process is called Clustering. Therefore, clustering is an unsupervised learning of a hidden data concept.

As such clustering does not use previously assigned class labels, except perhaps for verification of how well the clustering worked.

Clustering technique represents by two steps as such follows:

- Each cluster should be containing data items of homogeneous nature.
- Each cluster should be heterogeneous in nature from other clusters.

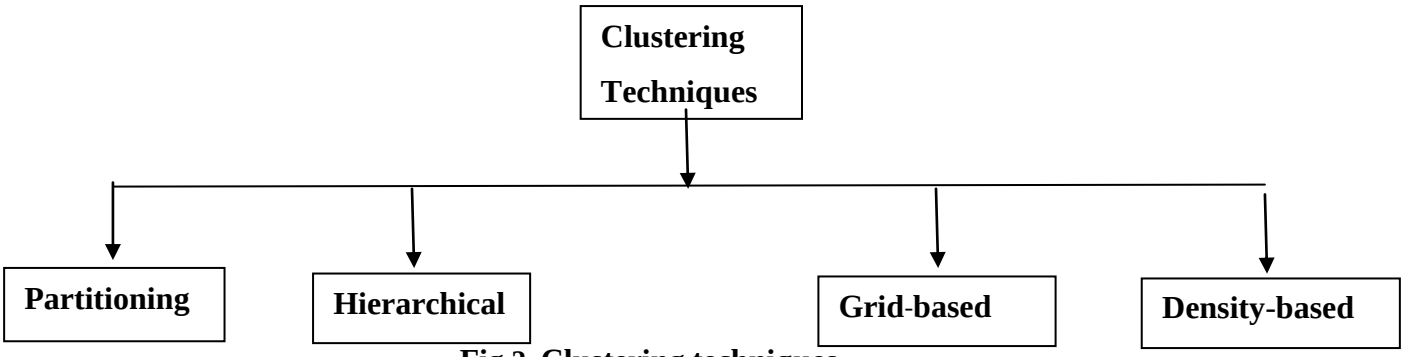

Fig 2. Clustering techniques

4. Machine Learning

Machine Learning is a technique which provide many tools to automatically analyse large dataset which are not possible to be analysed manually. It looks for informative patterns and relationships among attributes to improve decision making and prediction in certain domain. Attributes can be numeric or nominal

5. K-Means Clustering (Partitioning Clustering)

K-means clustering, also known as Partitioning Clustering, where objects are classified into one of K-groups. First of all randomly an initial centroids are chosen for each randomly partitioning clusters. Next, we compute the clusters means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step

until cluster means don't change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.[4]

Algorithmic steps for k-means clustering

1. Select K points as the initial centroids.
2. Choose k points at random as cluster centers.
3. Assign all instances to their closest cluster center.
4. Calculate the centroid (i.e., mean) of instances in each cluster.
5. These centroids are the new cluster centers.
6. Continue until the cluster center don't change.

Minimizes the total squared distance from instances to their cluster centers Local, not global, minimum!

- Sum of Squared Error (SSE) is the most crucial measuring parameter.
  - For each point, the error is the distance to the nearest cluster

6. Expectation Maximization Clustering (Probabilistic Clustering)

EM algorithm also known as, Probabilistic Clustering, is extremely complex in nature but gives very useful result for the real world dataset. Using this algorithm is useful when one wants to perform cluster analysis of a small scene or region of interest and when one is not satisfied with the results so obtained from K-Means. EM is an iterative method to search out the maximum likelihood. The EM iteration alternates between performing an expectation (E) step, which is computing the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which is computing 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.

General EM Algorithm:

Alternate steps until model parameters don't change much:

E step: To be given a certain fixed model, it will estimate distribution over labels

M step: In order to maximize expected log-likelihood of observed data and hidden variables, it will choose new parameters for model.

7. DBSCAN (Density based Clustering)

DBSCAN means Density-based Spatial Clustering of Applications with noise. DBSCAN is used in the formation of clusters irrespective of shape, size and location of the clusters. It works on two main concepts, that is, Density Reachability and Density Connectability. Both these concepts is depending on two parameters of the DBSCAN clustering; the size of epsilon neighbourhood e and the minimum points in the cluster m[3] .

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

8. WEKA

WEKA has its full form- Waikato Environment for Knowledge Learning, [7] which is an open source tool meaning available at public use. Developed at the University of Waikato in New Zealand, WEKA is a computer program which supports various standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA code is written in Java and has GUI interface.

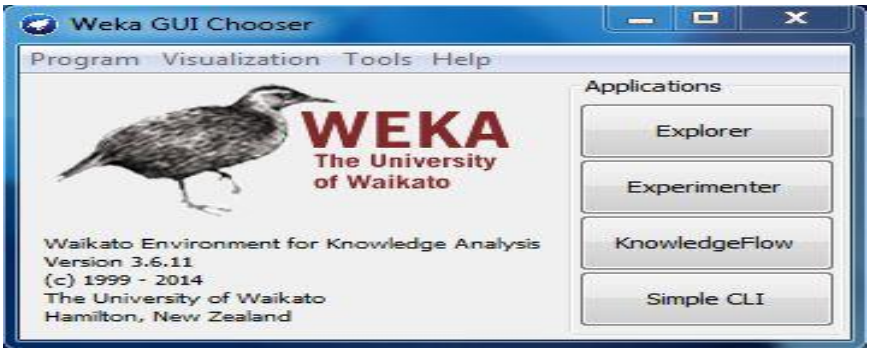

Figure 3: Weka GUI Interface

The GUI Chooser consists of four buttons, that is, Explorer, Experimenter, Knowledge Flow and Simple CLI.

9. Implementation

A) System requirement

System requirements are the basic requirements of the computer on which machine learning tool is to be run. It has been run on Weka 3.6.9, which is configured on Intel Core 2 Duo, 1 GB RAM, 32 bit Operating system having Windows 7 Home Premium.

B) Selection of Dataset

There are three datasets which have been downloaded from UCI repository [6] with varying instances. However, all three datasets are univariant and contains numeric attributes except one nominal class. First Dataset is of Wine Dataset having 178 instances and 14 attributes, second is of Glass Dataset having 214 instances and 10 attributes and finally third one is Diabetic patients having 768 instances and 9 attributes.

C) Selection of Tool

Along with the selection of appropriate dataset, choice of tool is also very important. The current study is implemented on machine learning tool WEKA which is Graphical user interface and it is an open source tool.it supports arff, csv file formats.

D) Selection of Technique

In Clustering, we normally make clusters based on the values given in the attributes and finally get clusters but here we use a technique of evaluation of clusters. In this technique, clusters are formed based on the attributes values, after then we evaluate the clusters so formed with already declared classes and finally find out which algorithm gives much accuracy.

Evaluation of clusters can be done by three ways:-

(i) ADD CLUSTER-In this method, we use an unsupervised filter Add Cluster and apply. By clicking edit button we can easily visualize cluster assignment.

(ii) Classes-to-Cluster Evaluation –In this method, there is a cluster mode similar to Use training set, supplied test set etc. and it finally shows incorrectly clustered instances.

(iii) Meta ClassificationViaClustering – In this method, one has to go Classification technique and under it one goes to Meta and then choose ClassificationViaClustering and finally choosing appropriate Clustering Algorithm simply runs it and get the desired results

We strict to Classes-to-Cluster Evaluation, however, above all three evaluation techniques lead to the same result.

10.Results and Analysis

Table 1 Classes to Cluster Evaluation of Wine Dataset

Name	Cluster Instances	Number of Iterations	Within clusters sum of squared errors	Time taken to build model (ms)	Log likelihood	Speed (instances/ms)	Incorrectly Clustered Instances
K-Means Algorithm	0:60(34%) 1:55(31%) 2:63(35%)	**8**	**48.970291155139165 99101**	**30**		**5.9333**	**10(5.618%)**
EM Algorithm	0:51(29%) 1:57(32%) 2:70(39%)			**160**	**-18.50709**	**1.1125**	**5(2.809%)**
DBSCAN	0:178 (100%)			**50**		**3.56**	**107(60.1124%)**

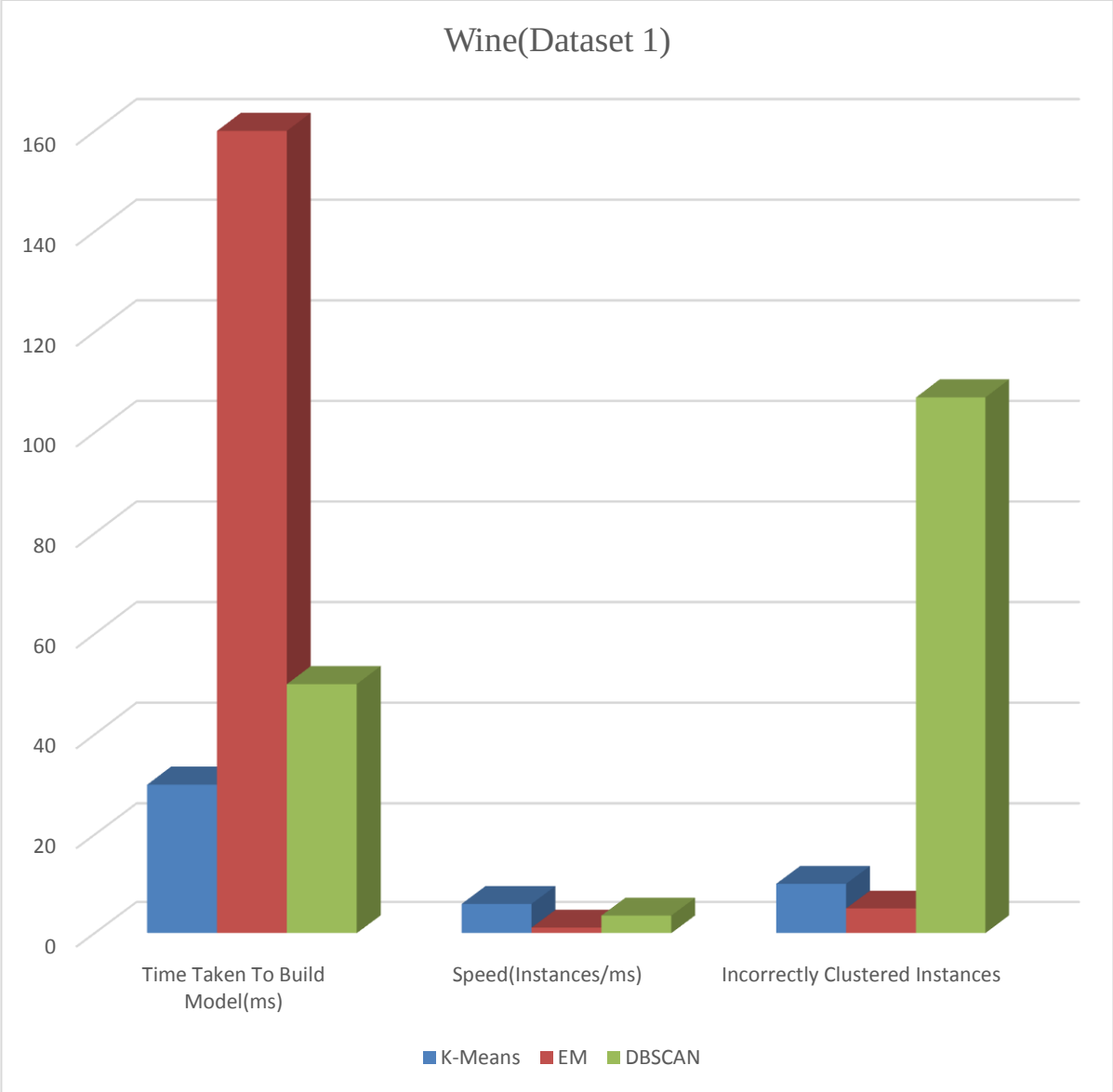

Fig 4

Table 2 Classes to Cluster Evaluation of Glasses Dataset

Name	Cluster Instances	Number of Iterations	Within clusters sum of squared errors	Time taken to build model (ms)	Log likelihood	Speed (instances/ms)	Incorrectly Clustered Instances
K-Means Algorithm	0:38(18%) 1:24(11%) 2:2(1%) 3:22(10%) 4:94(44%) 5:16(7%) 6:18(8%)	**12**	**17.199155810403063**	**50**		**4.28**	**118(55.1402%)**
EM Algorithm	0:27(13%) 1:44(21%) 2:21(10%) 3:6(3%) 4:28(13%) 5:77(36%) 6:11(5%)			**230**	**4.39188**	**0.91304348**	**116(54.2056%)**
DBSCAN	0:211(100%)			**60**		**3.5**	**136(63.5514%)**

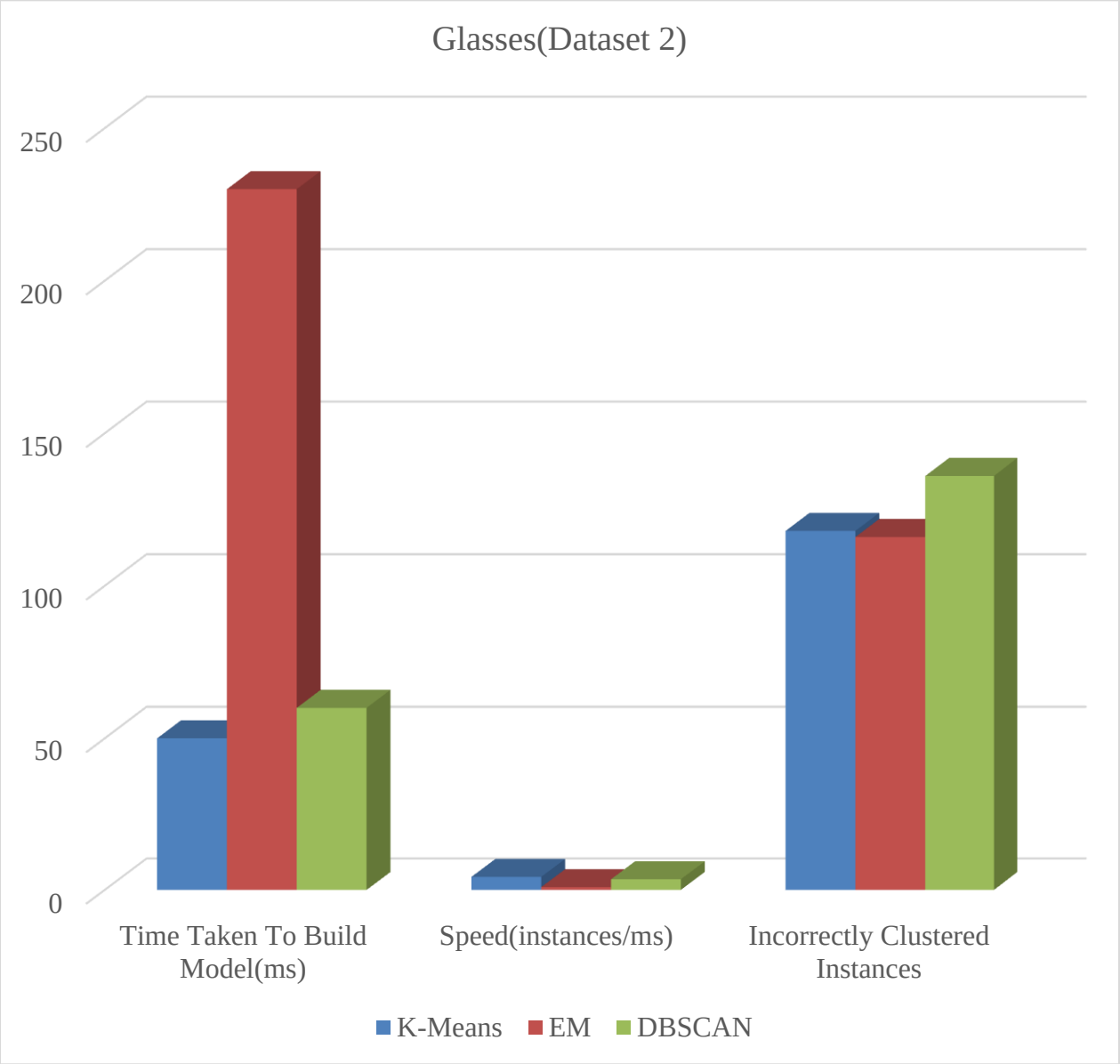

Fig 5

Table 3 Classes to Cluster Evaluation of Diabetic Dataset

Name	Cluster Instances	Number of Iterations	Within clusters sum of squared errors	Time taken to build model (ms)	Log likelihood	Speed (instances/ms)	Incorrectly Clustered Instances
K-Means Algorithm	0:515(67%) 1:253(33%)	**7**	**121.25790179 99101**	**30**		**25.6**	**255(33.2031%)**
EM Algorithm	0:353 (46%) 1:415(54%)			**560**	**-29.12836**	**1.37142857**	**261(33.9844%)**
DBSCAN	0:768 (100%)			**590**		**1.30169492**	**268(34.8958%)**

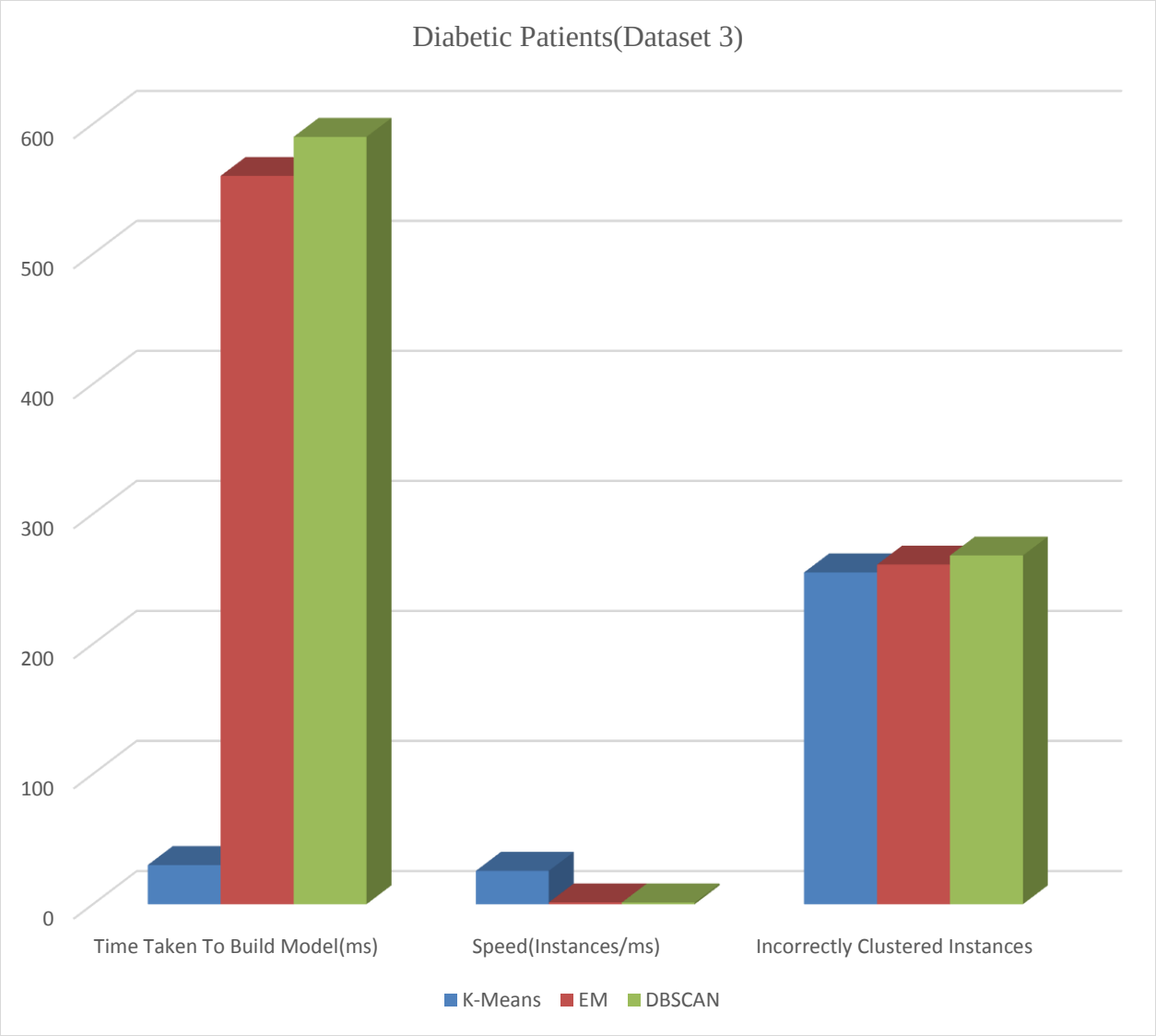

Fig 6

Accuracy

Classes to Cluster evaluation mode make clusters on the basis of numeric values of the attribute of the dataset and then nominal class value is evaluated with cluster so formed. Accuracy refers to the correctly classified instances in the dataset.

Clustering Error

Clustering error defines incorrectly classified instances, that is, instances which are wrongly clustered despite their original class is different

Table 4 .Comparative Analysis for different datasets of performance parameters for three selected algorithms

	Dataset 1		Dataset 2		Dataset 3	
Algorithm	**Accuracy**	**Clustering Error**	**Accuracy**	**Clustering Error**	**Accuracy**	**Clustering Error**
K-Means	**94.38%**	**5.62%**	**44.86%**	**54.21%**	**66.80%**	**33.20%**
EM	**97.19%**	**2.81%**	**44.40%**	**52.80%**	**66.02%**	**33.98%**
DBSCAN	**39.89%**	**60.11%**	**36.45%**	**63.55%**	**65.10%**	**34.90%**

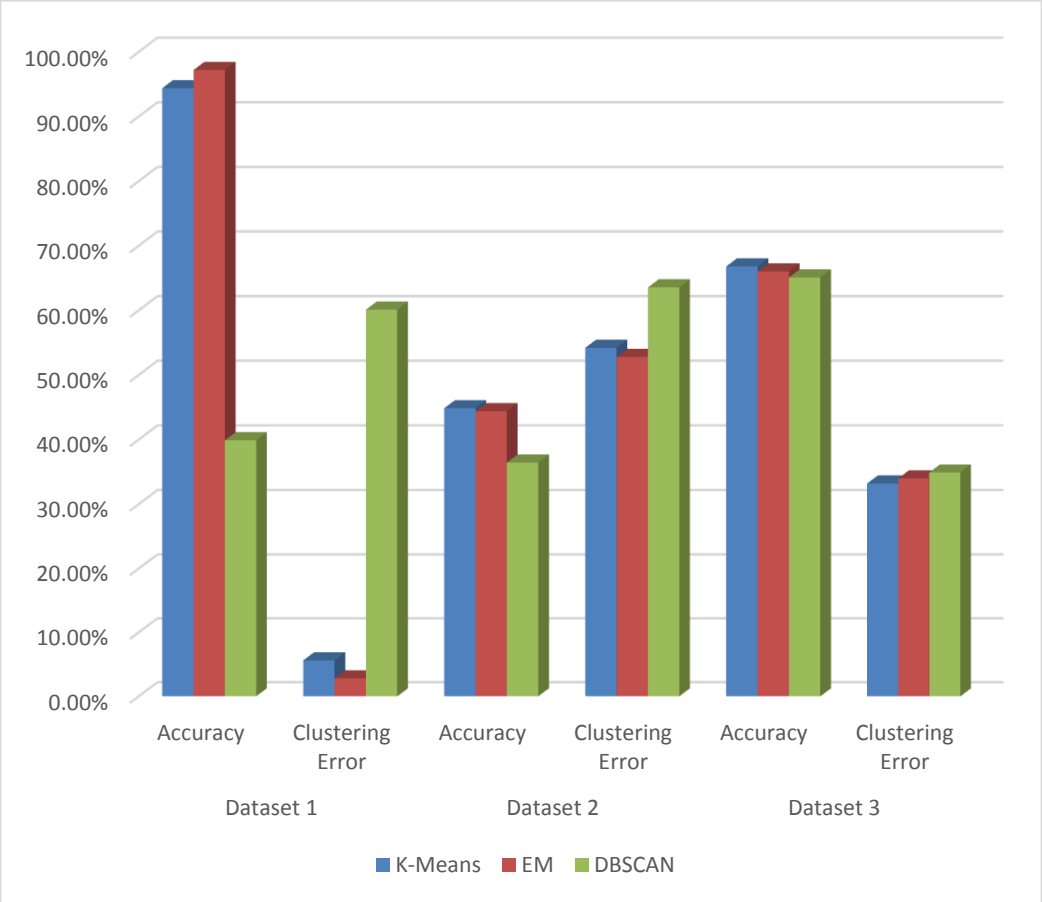

Fig 7

Fig 7 represents the graphical representation of the results obtained from the Table 4 .The graph clearly shows that K-means algorithm and EM algorithm shows almost the same type of accuracy and clustering error as compared to DBSCAN with varying instances of datasets .However, it is also shown from previous **figures 4, 5 and 6**; K-means take less time to build the model and has highest speed among all three algorithms .This shows that for numeric univariant attributes K-Means is the most suitable above EM and DBSCAN.

11.Conclusion

In Data Mining, first it is difficult to know which data mining technique to use and then which algorithm is the most suitable and this decision is the most probably taken by hit and trial method. This research has been done to give comparison of three clustering algorithms which is based by evaluation of clusters on three different datasets. Finally graph has been made on the correctly classified instances and accuracy. Atlast but not at least the conclusion comes that for numeric type of attributes values K-Means is the best one algorithm in clustering whereas EM is the best one when instances are in small numbers. As a future research, other clustering algorithms can be tested for multivariate attributes or a new algorithm having more efficiency can be proposed.

References

[1]Alex Berson, J.Smith Stephen"Data Warehousing, Data Mining, & OLAP". (2004)

[2]Ian H.Witten and Eidbe Frank, "Data Mining-Practical Machine learning tools & Techniques-Second Edition.

[3]Astha Joshi"A Review: Comparative Study of Various Clustering Techniques on Data Mining "Volume 3, Issue 3, March 2013 ISSN 2277-128X.

[4] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta "A Comparative Study of Various Clustering Algorithms in Data Mining" International Journal of Engineering Research & Applications,Volume 2,Issue 3,May-June 2012,pp 1379-1384, ISSN-248-9622

[5] Xingquan Zu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershley, New York, 2007

[6] http://www.ucirepository.com

[7] http:// www.cs.waikato.ac.nz/~ml/WEKA

