(IJIRSE) International Journal of Innovative Research in Science & Engineering
ISSN (Online) 2347-3207

An Experimental Analysis of Clustering Algorithms in Data Mining using Weka Tool

Vishnu Kumar Goyal
Dept. of Computer Engineering
Govt. R.C. Khaitan Polytechnic College, Jaipur, India
vishnugoyal_jaipur@yahoo.co.in

***Abstract*—Cluster analysis divides data into meaningful or useful groups (clusters). It is a process for discovering groups and identifying interesting patterns. There are different types of clusters: Well-separated clusters, Center-based clusters, Contiguous clusters, Density-based clusters, Shared Property or Conceptual Clusters. Predictive and the descriptive are the two main tasks of the data mining. Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. This paper analyze the five major clustering algorithms: COBWEB, DBSCAN, EM, FARTHEST FIRST and K-MEANS clustering algorithm and compare the performance of these major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. The results are tested on three datasets namely Iris, Haberman diabetes and glass dataset using WEKA interface and compute the correctly cluster building instances in proportion with incorrectly formed cluster.**

***Index Terms*— Cluster analysis, Clustering, Data Mining.**

I. INTRODUCTION

Clustering is the process of grouping a collection of objects (usually represented as points in a multidimensional space) into classes of similar objects. Cluster analysis is a very important tool in data analysis. It is a set of methodologies for automatic classification of a collection of patterns into clusters based on similarity. Intuitively, patterns within the same cluster are more similar to each other than patterns belonging to a different cluster. It is important to understand the difference between clustering (unsupervised classification) and supervised classification. Cluster center is the heart of the cluster. The process of making data clusters is defined in fig. 1.

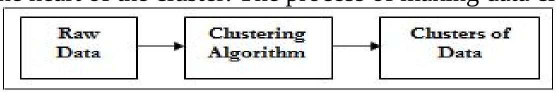

Fig. 1. Clustering Process

Firstly, we take raw data, then apply clustering algorithm on the raw data and after that we will get the clusters of data. This is the process of making data clusters with the help of Clustering algorithm.

II. THE DATA MINING PROCESS

Data mining is an iterative process that typically involves the number of phases. Figure 2 shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model

1. Problem definition

A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

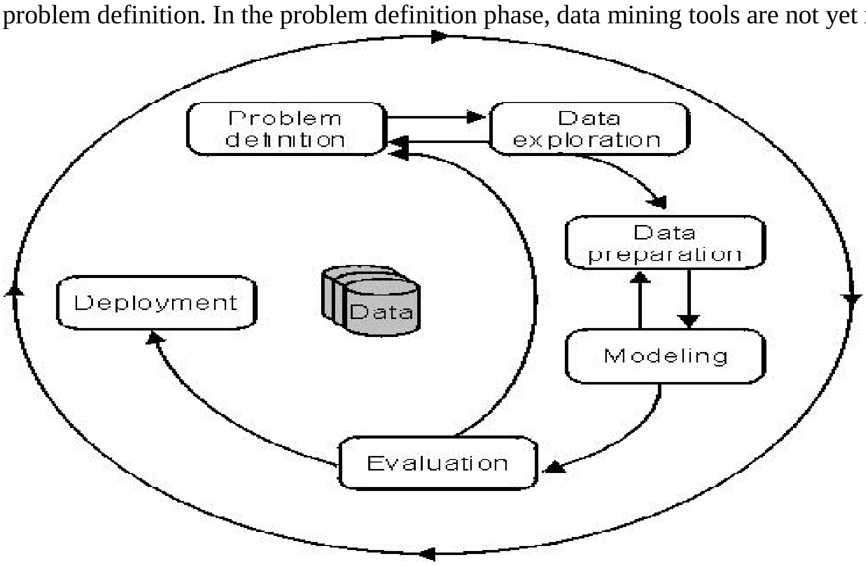

Fig. 2. CRISP DM process model

2. Data exploration
Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3. Data preparation
Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. Modeling
Data mining experts select and apply various mining functions because we can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

5. Evaluation
Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: Does the model achieve the business objective? Have all business issues been considered? At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

6. Deployment
Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets. The Intelligent Miner™ products assist to follow this process.

III. THE WEKA TOOL

For a successful clustering implementation, Weka 3.6.8 was used to aid the investigation. Data mining [1] isn't solely the domain of big companies and expensive software. In fact, there's a piece of software that does almost all the same things as these expensive pieces of software the software is called WEKA. WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997[14]. It uses the GNU General Public License (GPL). The figure of weka is shown in the figure 2.The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API, so we can embed WEKA, like any other library, in our own applications to such things as automated server-side data mining tasks.

Fig. 3. Weka GUI

For working of weka we not need the deep knowledge of data mining that's reason it is very popular data mining tool. Weka also provides the graphical user interface of the user and provides many facilities [4, 7].

IV. DATASET

For performing the comparison analysis we need the past project datasets. In this research I am taking data from two data repositories. ISBSG and PROMISE data repositories provide the past project data. This should have been taken the different- different nature. These repositories are very helpful for the researchers. We can directly apply this data in the data mining tools and predict the result. We have taken four datasets containing nominal attributes type that is all these datasets contains the continuous attributes. Each dataset's instance has contained an assigned class with it. On the basis of this class the cluster are generating by applying the above mentioned algorithms using the Weka interface. Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of Weka was released 11 years ago). These datasets have been taken from UCI machine learning repository system.

1 Iris plants dataset contains 3 classes of 50 instances each where each class refers to a type of iris plant. One class is linearly separable from the other 2, the latter are NOT linearly separable from each other. No. of instances are 150(50 in each of the 3 classes). No of attributes are 5 including the class attributes.

2 Haberman's Survival Dataset contains cases from a study that was conducted on the survival of patients who had undergone surgery of breast cancer. No. of instances 306 and no. of attributes are 4 including the class attribute.

3 Diabetes Dataset The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

4 Glass Dataset In determining whether the glass was a type of "float" glass or not. The study of classification of types of glass was motivated by criminological investigation.At the scene of the crime, the glass left can be used as evidenc, if it is correctly identified. Number of Instances: 214

V. METHODOLOGY

My methodology is very simple. I am taking the past project data from the repositories and apply it on the weka. In the weka I am applying different-different clustering algorithms and predict a useful result that will be very helpful for the new users and new researchers. Five well known and important algorithms COBWEB, DBSCAN, EM, FARTHEST FIRST and K-MEANS were applied on the Iris, Haberman, diabetes, and glass datasets and the outputs were tabulated and plotted in a 2 dimensional graph. Then one by one these datasets are evaluated and their clustering performance is evaluated. Amount of correctly clustered instances and incorrectly clustered instances have been recorded. Each algorithm is run over five predefined datasets and their performance is evaluated.

VI. EXPERIMENTAL SIMULATION AND RESULTS

The above discussed five algorithms have their implemented source code in the Weka 3.6.4 version upon which simulations have carried out in order to measure the performance parameters of the algorithms over the datasets. The results are summarized in the following tables and graphs.

TABLE I. PERFORMANCE OF COBWEB ALGORITHM

Dataset	**Instances**	**Incorrect cluster instances**	**Percentage % Incorrect cluster instances**
Iris	150	50	33.33
Haberman	306	299	97.71
diabetes	768	757	98.56
glass	214	115	53.73

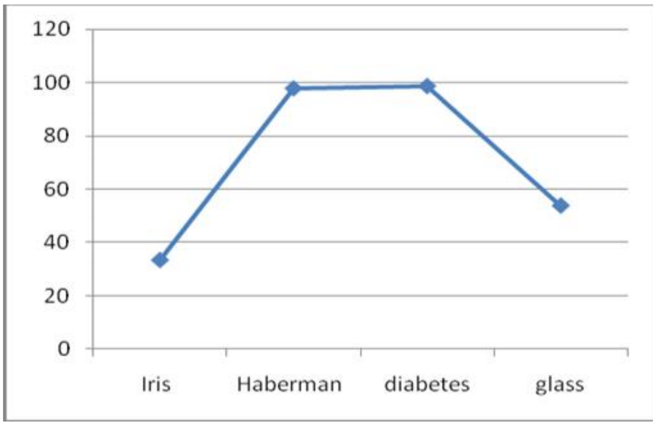

Fig. 4. COBWEB Algorithm: Percentage % Incorrect cluster instances

The COBWEB algorithm performs well for Iris dataset. It is also good for glass datasets, but for the Haberman and diabetes datasets, it not performing well.

TABLE II. PERFORMANCE OF DBSCAN ALGORITHM

Dataset	**Instances**	**Incorrect cluster instances**	**Percentage % Incorrect cluster instances**
Iris	150	100	66.66
Haberman	306	269	87.9
diabetes	768	268	34.89
glass	214	136	63.55

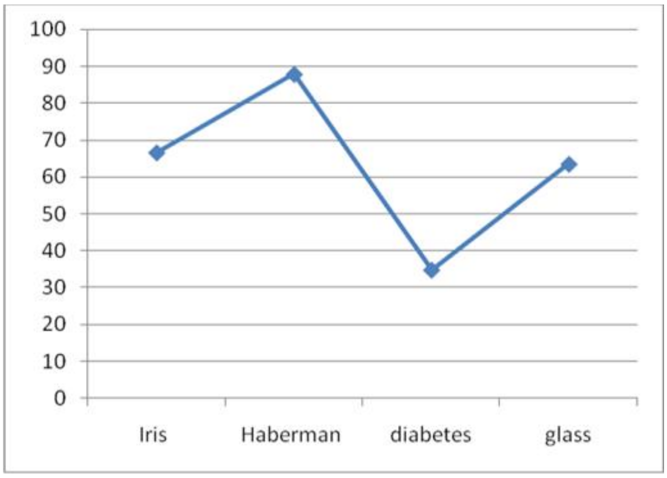

Fig. 5. DBSCAN Algorithm: Percentage % Incorrect cluster instances

As shown in the Fig. 5 the DBSCAN algorithm perform well for diabetes dataset. It is also very poor for Haberman dataset, but for the Iris and glass datasets, it performing average.

TABLE III. PERFORMANCE OF EM ALGORITHM

Dataset	**Instances**	**Incorrect cluster instances**	**Percentage % Incorrect cluster instances**
Iris	150	60	40
Haberman	306	102	33.33
diabetes	768	469	61.06
glass	214	116	54.2

As shown in the Fig. 6 the EM algorithm perform well for Haberman dataset. It is also improved for Iris dataset, but for the diabetes and glass datasets, it performing very poor.

As shown in the Fig. 7 the FARTHEST FIRST algorithm perform well for Iris, Haberman and diabetes dataset. But for the glass datasets, it performing very poor.

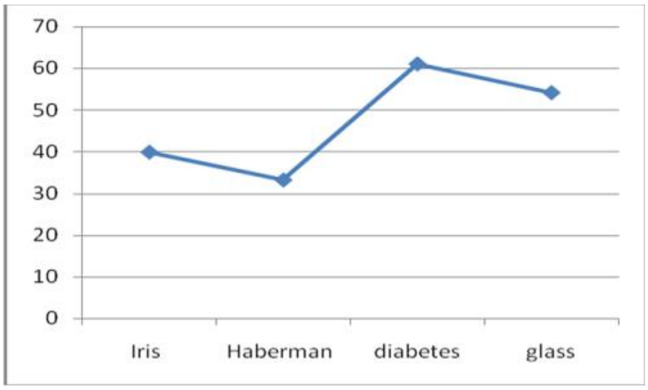

Fig. 6. EM Algorithm: Percentage % Incorrect cluster instances

TABLE IV. PERFORMANCE OF FARTHEST FIRST ALGORITHM

Dataset	**Instances**	**Incorrect cluster instances**	**Percentage % Incorrect cluster instances**
Iris	150	50	33.33
Haberman	306	93	30.39
diabetes	768	263	34.24
glass	214	139	64.95

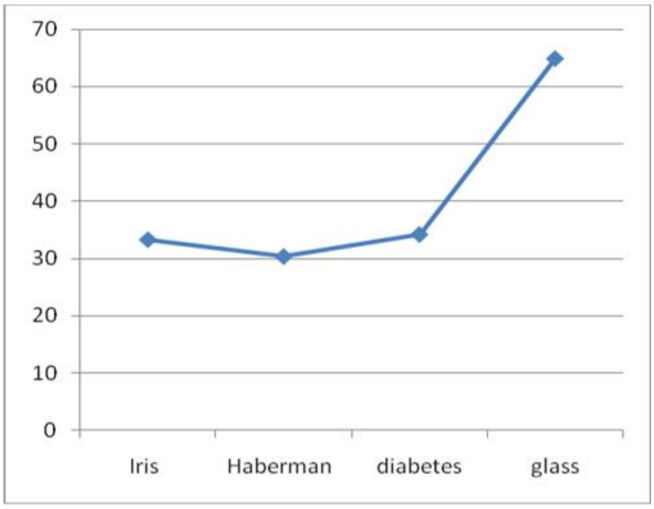

Fig. 7. FARTHEST FIRST Algorithm: Percentage % Incorrect cluster instances

TABLE V. PERFORMANCE OF K-MEANS ALGORITHM

Dataset	Instances	Incorrect cluster instances	Percentage % Incorrect cluster instances
Iris	150	50	33.33
Haberman	306	148	48.36
diabetes	768	255	33.2
glass	214	119	55.6

As shown in the Fig. 8 the K-MEANS algorithm perform well for Iris and diabetes dataset. It is also improved for Haberman and glass dataset.

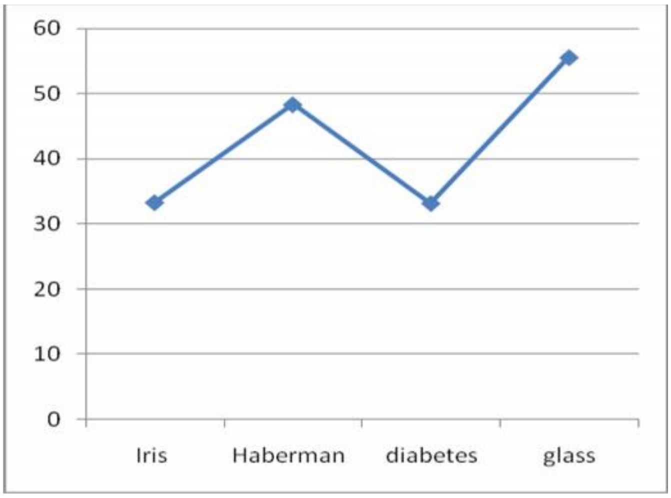

Fig. 8. K-MEANS Algorithm: Percentage % Incorrect cluster instances

VII. COMPARISON

The COBWEB, DBSCAN, EM, FARTHEST FIRST, and K-MEANS algorithm clustering techniques were used on the Iris, Haberman, diabetes, and glass datasets and the Consolidated outputs were tabulated and plotted in a 2 dimensional graph as shown below.

TABLE VI. PERFORMANCE COMPARISON

Dataset	COBWEB	DBSCAN	EM	FARTHEST FIRST	K-MEANS
Iris	33.33	66.66	40	33.33	33.33
Haberman	97.71	87.9	33.33	30.39	48.36
diabetes	98.56	34.89	61.06	34.24	33.2
glass	53.73	63.55	54.2	64.95	55.6

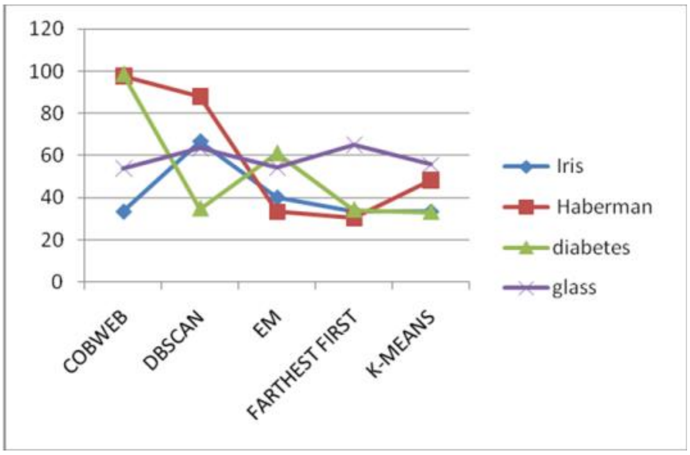

Fig. 9. Clustering Algorithm: Percentage % Incorrect cluster instances

VIII. CONCLUSION

For Iris dataset, the COBWEB, EM, and K-MEANS perform well, but the DBSCAN and EM are not good. For Haberman dataset, the EM, FARTHEST FIRST and K-MEANS perform well, but the DBSCAN and COBWEB are worst. For diabetes dataset, the DBSCAN, FARTHEST FIRST and K-MEANS perform well, but the COBWEB and EM are not good. For glass dataset, the COBWEB, EM, and K-MEANS perform average, but the DBSCAN and FARTHEST FIRST are not good. For the given datasets, COBWEB algorithm is performing worst and DBSCAN is performing average. The FARTHEST FIRST shows some improvement over these two. The EM and K-MEANS are performing well among the all algorithms, but the K-MEANS can be considered as the best among these algorithms.

REFERENCES

[1] Xiaozhe Wang, Kate Smith and Rob Hyndman: "Characteristic-Based Clustering for Time Series Data", Data Mining and Knowledge Discovery, Springer Science Business Media, LLC Manufactured in the United States, 335–364, 2006.

[2] Ehsan Hajizadeh, Hamed Davari Ardakani and Jamal Shahrabi:"Appilication of data mining techniques in stock market", Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010.

[3] Jiangjiao Duan, WeiWang , Bing Liu and Baile Shi:" Incorporating with recursive model training in time series clustering", Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05), IEEE2005.

[4] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.

[5] Sapna Jain, M Afshar Aalam and M N Doja, "K-means clustering using weka interface", Proceedings of the 4th National Conference; INDIACom-2010.

[6] MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.University of California Press. 1967, pp. 281–297.

[7] Lloyd, S. P. "Least square quantization in PCM". IEEE Transactions on Information Theory 28, 1982,pp. 129–137.

[8] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers,second Edition, (2006).

[9] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384

[10] E.B Fawlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78:553–584, 1983

[11] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and intialization methods. Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA.

[12] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, 14:315–332

[13] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.

[14] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[15] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery, 2:283–304, 1998.