*May - June 2020*
*ISSN: 0193-4120 Page No. 3572-3576*

# A Comparative Study of Data Mining Techniques

**[1]Limbada Ahmad Ilyas, [2]S. Senthil**

[1]MCA Student, [2] Professor and Director,
[1,2]School of Computer Science and Applications, Reva University, Bangalore, Karnataka, India
[1]ahmadlimbada@gmail.com, [2]dir.csa@reva.edu.in

***Article Info***
***Volume 83***
***Page Number: 3572-3576***
***Publication Issue:***
***May - June 2020***

***Article History***
***Article Received:*** *19 August 2019*
***Revised:*** *27 November 2019*
***Accepted:*** *29 January 2020*
***Publication:*** *12 May 2020*

**Abstract**
Data mining is defined by different people with different manners but generally, it is nothing but a method of extracting knowledge from the data. For decades, Humans have been generating and storing large amounts of data and dumping them into the storage. Data mining can be used to extract the knowledge from this vast size of data. Different techniques are available for analysis and comparison of different available algorithms are analyzed in this paper. Among the Data mining techniques, there are different task are performed on the data based on the data and knowledge required from the acquired data. There are a variety of algorithms like Naïve Bayes, Decision Tree-based Algorithms, Rule-based classifiers, etc. Data mining techniques are divided into several methods like Classification, Regression, Clustering, Association Rules, etc. In this comparative study, different algorithms available in some of the mentioned categories will be compared based on the accuracy and time complexity of the algorithms.

***Keywords:*** *Data mining, Algorithm, Classification, Clustering, Regression*

## 1. Introduction

In this rapidly developing era of Information Technology, Humans are generating around 2.5 quintillion bytes of the data every day [1]. This data are majorly generated by Social Media Platforms, Web Streaming Platforms, and the Gaming Industry, etc. This much vast amount of data storage is possible because of low data storage costs. Major companies like Facebook, Google, YouTube, and Netflix are storing and analyzing this data for finding useful hidden patterns, relations in the data. Most of the data which is being generated is about user details and behavior. One of the examples is Facebook uses Data Mining and Digital Image Processing for tagging people into the post, Facebook by default identifies the people in the photo and suggest user to tag that person.

All these things are possible due to Data Mining. Data Mining is an interdisciplinary field that includes many fields like Machine Learning, Artificial Intelligence, Big Data, Statistics and Business Intelligence. All these fields use data mining for different purposes and the type of techniques is decided based on the purpose. While a system like face recognition, it is a combination of more than one technique.

Data mining is defined as "It is a process of extracting accurate and useful knowledge in the data"[2]. People might use different names for Data Mining like Deductive Learning, Data Dredging, and Knowledge Extraction, etc. Data Mining consists of a variety of techniques like Classification, Regression, Clustering, Association Rules, Summarization, etc. but all these techniques can be classified into two major categories, Predictive Techniques, Descriptive Techniques.

Before applying any of the technique, Preprocessing is needed to increase the correctness of the data. Correct data will lead to more accurate results. After preprocessing the data, segregation of the data into training data and the test data will happen. Training data is used to train the model and later test data is used to check the accuracy and error ratio in the model. A model with the best accuracy can be said to be the best model for a particular technique.

Predictive technique's objective is to predict the value of an attribute based on the values of available

*May - June 2020*
*ISSN: 0193-4120 Page No. 3572-3576*

values of other attributes. An attribute which is to be predicted is Target or dependent variable while Attributes which will be used for making prediction are to be known as Explanatory or Independent variables. Predictive Techniques are majorly four which are Classification, Regression, Time Series Analysis and Prediction. Each of these techniques has some algorithms in it, out of all algorithms some of the algorithms will be compared to check the accuracy of the results generated by the algorithm in further discussion in this paper.

A Descriptive technique's objective is to find the patterns and/or relationship in data. It does not predict the values as Predictive Techniques but It identifies the hidden meaning in the existing data. Descriptive Techniques are majorly four which are Clustering, Summarization, Association Rules, and Sequence discovery. A comparison of various clustering techniques results will be done.

This paper is organized in such a manner that Section 1 contains a brief Introduction about Data mining and its types, Section 2 has summary of research done by other researchers, Section 3 contains a variety of prediction techniques available while section 4 has different descriptive techniques available. Section 5 contains information about the tool used for comparison and Datasets used for the comparison and comparison between various algorithms in each technique.

## 2. Literature Review

This section contains various reviews about technical papers on a comparison between data mining techniques carried out by many researchers. This section summarizes the brief research and survey done by other researchers.

In Sumit Garg & Arvind K.[3], discussed the Comparative Analysis of Data Mining Techniques on Educational Dataset in which they have compared various merits and demerits in each of the algorithms and concluded that this algorithm can be used to improve the student's performance.

In Keshav Singh Rawat[4], explained the Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics. Paper has an introduction of various data mining techniques types and mentioned various algorithms in each of the types. Paper also contains information about various tools,their features, and comparison between these tools is mentioned.

In Mr. Nilesh Kumar Dokania and Ms. Navneet Kaur[5], Paper was written on Comparative Study of Various Techniques in Data Mining. The paper contains information about preprocessing of data, classification of various data mining techniques and various types of tasks available in each classified technique. Paper also has brief details description about a comparison of K-means and Y means clustering as well as the application of some of the techniques.

## 3. Prediction Techniques

### 3.1. Classification

Classification is a type of supervised learning which assigns the data to a collection of predefined target classes. Classification models majorly used to predict to which predefined class a particular member belongs. Classification uses majorly four types of the algorithm: Statistical based classifiers, Decision Tree-based Classifiers, Distance-based classifiers, Rule-based classifiers, and Neural network Classifiers. Classification generally analyzes the importance of attribute for classification from the training data set and later based on that it classifies the test data.

#### 3.1.1. Statistical based Classifier

Statistical based classifier uses statistical measures to classify the tuple to a particular class. Some of the algorithms for statistical-based classifiers are Naïve Bayes Classifier, Regression.

#### 3.1.2. Decision Tree-based Classifier

Decision Tree-based classifier uses a tree data structure to make decisions and classify the tuple to a particular class. Some of the algorithms for Decision Tree-based classifiers are ID3[6], C4.5[7], J48[8] and CART[9].

#### 3.1.3. Distance-based Classifier

The distance-based classifier uses Euclidian Distance and/or Manhattan distance to find the distance of the particular tuple from the class. The tuple will be assigned to the nearest class to that tuple. An example of this type of classifier is K Nearest Neighbor [10].

#### 3.1.4. Neural Network Classifier

Neural Network classifier uses the graphs and linking of the graphs based on the attribute values to classify the tuples. Major algorithms used in this category are Propagation [11] and Back Propagation [12].

#### 3.1.5. Rule-Based Classifier

The rule-based classifier uses the if-else rules to classify the tuple. This type of classifier generates the model with conditions based on data items in the dataset which condition can be applied to classify. 1R [13] and RIPPER [14] are an example of Rule-Based Classifier.

### 3.2. Regression

Regression is used to predict future values based on continuous values. The classification works on discrete values while Regression works on continuous values. In Regression, Identification and Analysis of the continuous behavior of the data is performed to predict future behavior. Regression is classified into several types which are Simple Linear Regression, Linear Regression, etc.

*May - June 2020*
*ISSN: 0193-4120 Page No. 3572-3576*

### 3.2.1. Simple Linear Regression

Simple Linear Regression establishes a relationship between the target variable (Y) and single predictor variables (X) by using a best fit straight line (also referred as a regression line).It is represented by an equation Y=a+b*X , where a is the intercept, b is the slope of the straight line. This equation is often used to predict the value of the dependent variable based on the given independent variable.

### 3.2.2. Multiple Linear Regression

Multiple Linear Regression establishes a relationship between target variable (Y) and two or more predictor variables (X) by using a best fit straight line (also referred as a regression line).It is represented by an equation $Y=a+b_1X_1+b_2X_2+b_nX_n$, where a is the intercept, $b_n$ is the slope of the straight line. This equation is often used to predict the value of the dependent variable based on the given independent variable(s).Multiple Linear Regression will be chosen over Simple Linear Regression if the predictor variable is more than one.

### 3.3. Time Series Analysis

Time Series Analysis can be defined as analyzing the behavior of the data over a period of time. It is used to predict future behavior at a particular moment of the time.

## 4. Description Techniques

### 4.1. Clustering

Clustering is a type of unsupervised learning in which it tries to group the data into different groups so that the data with similar patterns are assigned to each group. This whole process of grouping the clusters on runtime is called clustering. A variety of measurements like similarity or dissimilarity are used to form the clusters. Unlike Classification, groups are not predefined in Clustering but they are decided based on the type of data provided to train the model. A variety of clustering algorithms are available, some of the examples are Hierarchical algorithms, Partition-Based Algorithms, Density-based Algorithms, etc. Clustering generally suffers from outliers and deciding the number of clusters to be generated.

### 4.1.1. Hierarchy-based Algorithms

Hierarchy based algorithms use levels to create the cluster at each level. It uses dendogram (a kind of tree data structure) to illustrates the hierarchy of the clusters. There two major types of algorithms in this type which are Agglomerative and Divisive. Agglomerative uses the bottom-up approach to form the clusters while Divisive uses the top-down approach to form the clusters. Agglomerative generally starts with n cluster where n is the number of data items in dataset and stops when k number of cluster is achieved while Divisive starts with 1 single big clusters and stops when k number of clusters are achieved. Some of the algorithm is Simple Agglomerative which is divided into Single Link[15], MST Single Link[16], and Complete Link[17], etc.

### 4.1.2. Partition-based Algorithms

Partition based algorithms do not use hierarchy to form the cluster instead they create the cluster in one step as opposed to several steps. This type of algorithm generally deals with static data sets. Some algorithms available in this category are K-Means[18], BEA[19], and PAM[20], etc. This type of algorithm has limitations like it cannot handle categorical data, a number of clusters must be predefined, etc.

### 4.1.3. Density-based Algorithms

Density-based clustering can be defined as identifying the clusters based on the data space, in a continuous range of high point density and low point density. In which the data on the low point density data space are typically outliers. Some algorithms available in this category are DBSCAN[21],P-DBSCAN[22], etc.

### 4.2. Association

Association or also known as Association Rule Techniques is used to generate the rules from frequent data items from datasets. It is based on two factors called support and confidence. Support is nothing but no times particular data item appears in Datasets while Confidence is the appearance of data item x with data item y. Confidence can be defined as the possibility of the appearance of data item x when data item y appears in the dataset. There are varieties of algorithms available in this technique and the quality of rules generated by these algorithms is determined via Support, Confidence, Interest, Conviction and Chi-Squared Test. Algorithms available in this category are Apriori[23], Sampling Algorithm[24], Partitioning algorithm[25], etc.

## 5. Experimental Results

### 5.1. Tool

A tool that is going to be used for experiments is WEKA v3.8.4. WEKA (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) is an open-source machine learning software with interactive GUI and easy to access different Data Mining Techniques. All of the techniques discussed in sections 2 and 3 are one click away in Weka. Weka Explorer is selected and inside that various tabs like Preprocess, Classify, Cluster, Associate, and Visualize are available for data mining and machine learning. As discussed in section 2 before applying any technique preprocessing of the data is required and these facilities are also given by Weka and it is just a click away. More than one algorithm can be applied to the same dataset and results can be compared.

To setup the WEKA just visit this link[26] and download the WEKA which is compatible with particular

TEST Engineering & Management

*May - June 2020*
*ISSN: 0193-4120 Page No. 3572-3576*

OS. WEKA is available for Windows, Mac OS as well as Linux. Weka 3.8.4 requires Java 8 because it is the minimum Java version required to run WEKA. Windows OS also requires a high pixel density display for appropriate scaling of WEKA's GUI. WEKA support varieties of data files like .arff, .names, .csv, .dat, .json, etc. Dataset can be directly imported from URL, too. After adding the dataset pre-process the data in the pre-process tab and move towards the desired technique.

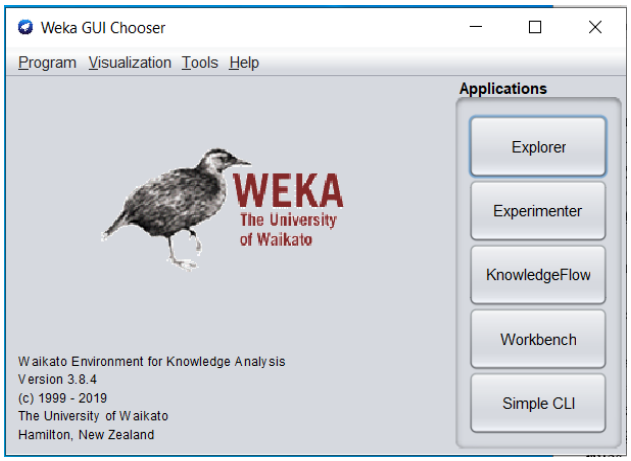


Figure 1: WEKA Home Page

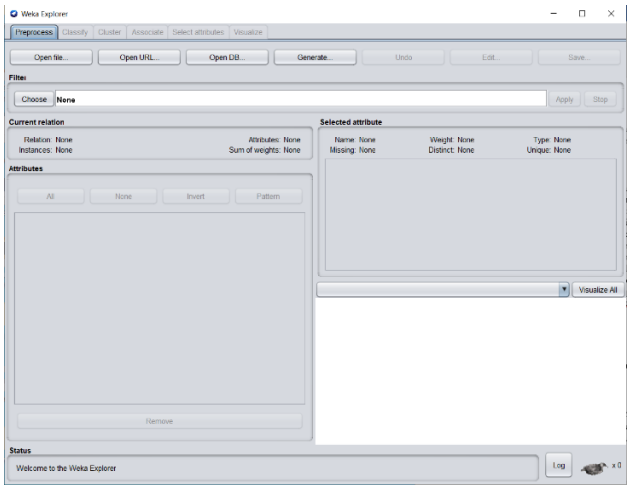

Figure 2: WEKA Explorer

### 5.2. Datasets

Datasets that are going to be used can be found on this link [27]. WEKA has specific data sets for specific problems so this data sets can be easily downloaded for free and used for the comparison.

### 5.3. Experimental Results Evaluation

Various mentioned algorithms above in classification technique are compared in WEKA. Naïve Bayes from Statistical based classifier, C4.5 from Decision Tree-based Classifier, KNN from Distance-based classifier and 1R, RIPPER from Rule-based classifier. While comparing this algorithm A % is accuracy in percentage which is derived via a number of correctly classified instance and T(s) is the time taken by an algorithm to generate results in seconds. 80-20 rule is used for training and testing the model during the comparison. 80% instance from data set is used to train the model while the remaining 20% instance is used for testing the model. Amongst all algorithms, J48 came out most accurately with aggregate 84.491% accuracy and 1R was least accurate with aggregate 73.29 % accuracy for chosen datasets. NN backpropagation is the most time consuming with aggregate 47.401 s and Naïve Bayes is the least time consuming with only 0.036 s.

| SL NO | Name of Dataset | Naïve Bayes | | J48 | | KNN | | NN Back Propagation | | 1R | | RIPPER | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A % | T(s) | A % | T(s) | A % | T(s) | A % | T(s) | A % | T(s) | A % | T(s) |
| 1 | breast-cancer | 70.17 | 0 | 71.92 | 0 | 71.92 | 0.01 | 64.91 | 2.71 | 61.4 | 0 | 68.42 | 0 |
| 2 | diabetes | 77.27 | 0.01 | 75.97 | 0.01 | 74.67 | 0.01 | 74.02 | 0.41 | 75.32 | 0.01 | 81.81 | 0.06 |
| 3 | glass | 51.16 | 0 | 60.46 | 0.01 | 62.79 | 0 | 55.81 | 0.26 | 60.46 | 0 | 65.116 | 0.04 |
| 4 | heart-c | 86.88 | 0 | 85.24 | 0.01 | 73.77 | 0.01 | 85.24 | 0 | 78.68 | 0 | 78.68 | 0.03 |
| 5 | iris | 96.66 | 0 | 100 | 0 | 96.66 | 0 | 96.66 | 0.06 | 96.66 | 0 | 96.66 | 0 |
| 6 | letter | 64.75 | 0.35 | 87.4 | 1.19 | 95.6 | 3.59 | 81.225 | 126.43 | 16.75 | 0.08 | 85.07 | 45.48 |
| 7 | mushroom | 95.38 | 0 | 100 | 0.01 | 100 | 0.58 | 100 | 329.04 | 98.4 | 0.01 | 100 | 0.15 |
| 8 | sick | 91.77 | 0 | 98.67 | 0.05 | 96.28 | 0.28 | 97.08 | 13.11 | 96.551 | 0.02 | 98.54 | 0.12 |
| 9 | vehicle | 44.37 | 0 | 71 | 0.01 | 69.82 | 0.01 | 85.79 | 1.49 | 54.43 | 0 | 70.41 | 0.08 |
| 10 | vote | 87.35 | 0 | 94.25 | 0 | 88.5 | 0 | 95.4 | 0.5 | 94.25 | 0 | 95.4 | 0 |
| | | 76.576 | 0.036 | 84.491 | 0.129 | 83.001 | 0.449 | 83.613 | 47.401 | 73.29 | 0.012 | 84.010 | 4.596 |

Figure 3: Comparison between Naïve Bayes, J48, KNN, NN Back Propagation, 1R, and RIPPER]

#### 5.3.2. Practical Comparison of Clustering Algorithms

Some of the algorithms for clustering like Simple EM(Expectation-Maximization), Agglomerative and Simple K Means are compared in WEKA and their results are mentioned in the below table. Out of these three algorithms, Simple EM is most accurate compared to Agglomerative and Simple K Means.

| SL No | Name of Dataset | No Of Clusters | Simple EM | Agglomerative | SimpleKMeans |
|---|---|---|---|---|---|
| 1 | car | 4 | 69.44 | 69.85 | 43.35 |
| 2 | glass | 7 | 43.93 | 36.45 | 44.86 |
| 3 | sonar | 2 | 53.85 | 52.89 | 54.33 |
| 4 | wine | 3 | 97.2 | 38.7 | 94.38 |
| 5 | zoo | 7 | 67.33 | 87.13 | 72.28 |
| | | | 66.35 | 57.004 | 61.84 |

Figure 4: Comparison between Simple EM, Agglomerative and Simple K Means

### 6. Conclusion

The result of this paper indicates that each and every data mining technique comprises of more than one algorithms. Each of the algorithms in each category has its own advantages and disadvantages. To achieve the best accuracy for any problem, The selection of algorithms

*May - June 2020*
*ISSN: 0193-4120 Page No. 3572-3576*

totally dependent on the dataset in the problem. While choosing the algorithm, One can apply multiple algorithms to the same dataset and choose the algorithm with the best result. A particular algorithm is best for all the problems cannot be decided directly. In future work,a specific type of dataset can be chosen and multiple algorithms can be applied to get more clear results about the algorithms.

**References**

[1] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#a7b36a560ba9

[2] https://economictimes.indiatimes.com/definition/data-mining

[3] Garg, Sumit, and Arvind K. Sharma. "Comparative analysis of various data mining techniques on educational datasets." *International Journal of Computer Applications* 74, no. 5 (2013).

[4] Rawat, K. S. "Comparative Analysis of Data Mining Techniques, Tools and Maching Learning Algorithms for Efficient Data Analytics." *JOSR J. Comput. Eng.* 19, no. 4 (2017): 56-60.

[5] Mr. Nilesh Kumar Dokania and Ms. Navneet Kaur, "COMPARATIVE STUDY OF VARIOUS TECHNIQUES IN DATA MINING." *IJESRT, ISSN: 2277-9655*, pp. 202-209, May 2018.

[6] Quinlan, J. R. "Semi-autonomous acquisition of pattern-based knowledge." *Introductory readings in expert systems* 12 (1982).

[7] Quinlan, J. Ross. *C4. 5: programs for machine learning*. Elsevier, 2014.

[8] Quinlan, J. Ross. "Improved use of continuous attributes in C4. 5." *Journal of artificial intelligence research* 4 (1996): 77-90.

[9] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. "Classification and regression trees. Belmont, CA: Wadsworth. " *International Group* 432 (1984): 151-166.

[10] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46, no. 3 (1992): 175-185.

[11] https://towardsdatascience.com/coding-neural-network-forward-propagation-and-backpropagtion-ccf8cf369f76

[12] Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016). "6.5 Back-Propagation and Other Differentiation Algorithms". Deep Learning. MIT Press. pp. 200–220. ISBN 9780262035613.

[13] Holte, Robert C. "Very simple classification rules perform well on most commonly used datasets." *Machine learning* 11, no. 1 (1993): 63-90.

[14] Cohen, William W. "Fast effective rule induction." In *Machine learning proceedings 1995*, pp. 115-123. Morgan Kaufmann, 1995.

[15] Sibson, Robin. "SLINK: an optimally efficient algorithm for the single-link cluster method." *The computer journal* 16, no. 1 (1973): 30-34.

[16] Gower, John C., and Gavin JS Ross. "Minimum spanning trees and single linkage cluster analysis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 18, no. 1 (1969): 54-64.

[17] Defays, Daniel. "An efficient algorithm for a complete link method." *The Computer Journal* 20, no. 4 (1977): 364-366.

[18] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281-297. 1967.

[19] Arabie, Phipps, and Lawrence J. Hubert. "The bond energy algorithm revisited." *IEEE transactions on systems, man, and cybernetics* 20, no. 1 (1990): 268-274.

[20] Kaufman, Leonard, and Peter J. Rousseeuw. "Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm." *Y. Dodge, Ed* (1987): 405-416.

[21] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231. 1996.

[22] Kisilevich, Slava, Florian Mansmann, and Daniel Keim. "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos." In *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*, pp. 1-4. 2010.

[23] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." In *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487-499. 1994.

[24] Toivonen, Hannu. "Sampling large databases for association rules." In *Vldb*, vol. 96, pp. 134-145. 1996.

[25] Swarndeep Saket, J., and D. S. Pandya. "An overview of partitioning algorithms in clustering techniques." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 5, no. 6 (2016): 1943-1946.

[26] https://waikato.github.io/weka-wiki/downloading_weka/

[27] https://waikato.github.io/weka-wiki/datasets/