# Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining

Pritam H. Patil
M.E.(Comp)
Pune University
India

Suvarna Thube
M.E. (Comp).
Pune University
India

Bhakti Ratnaparkhi
M.E.(Comp)
Pune University
India

K.Rajeswari
Ph.D Research Scholar
SASTRA University
Tanjore, Tamilnadu
India.

## ABSTRACT

Now days in all fields to extract useful knowledge from data, data mining techniques like classification, clustering, association rule mining are useful. In data mining classification is categorization of different objects and Clustering is methodology using which we will be able to club objects of similar type. Another methodology like association rule mining (ARM) [1] is useful to find out association relationship among different objects. This paper compares performance of different data mining tools [2] like WEKA [3], XLMiner [4] and KNIME [5] for these data mining techniques. We have used Statlog heart disease dataset [6] for analyzing performance of tools.

## General Terms

Data mining, Classification, Clustering.

## Keywords

Classification, Clustering, Association rule mining, WEKA, KNIME, XLMiner.

## 1. INTRODUCTION

Data can be in any form like facts, text or numeric which can be computed by computer. This data when converted to meaningful data becomes information. Knowledge is nothing but useful information. Data mining is process of finding knowledge from huge collection of data. It has various methods to extract hidden knowledge from large data set. Techniques like association rule mining, classification and clustering can be used to analyze data. Classification is the process of finding a set of models that describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown [7]. In clustering each object is similar to the other objects in the cluster and different from objects in all the other clusters. In association rule mining hidden association relationships are discovered from available data set. This data is used for further analysis to generate patterns. Using techniques of data mining on heart disease dataset with the help of tools we can predict whether person will have heart attack in future or not [8]. In India availability of expert doctors is less. To utilize their valuable time only critical patients can be determined and only those can be treated by expert doctors. In this way death rate can be reduced and also expert's time can be utilized efficiently.

In this paper we are describing analysis of three different tools on the basis of various performance parameters.

## 2. TOOLS DESCRIPTION

WEKA (Waikato Environment for Knowledge Analysis) is Collection of machine learning algorithm for data mining task written in a JAVA. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. WEKA is open source software issued under the GNU General public License. KNIME, the Konstanz Information Miner [9], is an open source data analytics reporting and integration platform. It is developed by University of Konstanz and Silicon Valley Software Company. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), modeling and data analysis and visualization. It is written in a java based on Eclipse. Since 2006, KNIME has been used in pharmaceutical research. XLMiner is the only comprehensive data mining add-in for Excel, with neural nets, classification and regression trees, logistic regression, linear regression, Bayes classifier, K-nearest neighbors, discriminate analysis, association rules, clustering, principal components, and more. For analysis purpose we using WEKA 3.6.10, KNIME 2.9.2 and XLMiner Education Edition on windows 32 bit OS.

## 3. DATASET DESCRIPTION

We tested the Statlog heart disease dataset on different data mining tools such as WEKA, KNIME and XLMiner. This dataset is taken from the University of California, Irvine (UCI) [6] Machine Learning Database. It contains total 270 instances of healthy persons and patients with heart problem. It includes class information and total 13 attributes as listed below:

1) Age

2) Sex

3) Chest pain type (four values)

4) Resting blood pressure

5) Serum cholesterol (in milligrams per deciliter)

6) Fasting blood sugar > 120 mgldL

7) Resting electrocardiographic results (values 0, 1, 2)

8) Maximal heart rate achieved

9) Exercise-induced angina

10) Old peak (ST depression induced by exercise relative to rest)

11) The slope of the peak exercise ST segment

12) Number of major vessels (0-3) colored by fluoroscopy

13) Thai: 3 - normal; 6 – fixed defect; 7 - reversible defect.

The class information is included in the dataset as absent and present regarding the absence and presence of heart disease, respectively.

## 3.1 Dataset Format & Preprocessing

Each data mining tool supports different extension formats of dataset. We used .arff format for WEKA tool. If data format is not suitable / supported then it needs to be converting into require format. For these purpose WEKA Converter has different loaders such as ArffLoader, CSVLoader, DatabaseLoader, etc. For XLMiner .xlsv or .csv format is supported. This tool takes sample from Excel Worksheet or Database. In case of KNIME tool we have to take IO node for reading data before processing it. Under IO node various subnodes comes such as Read, Write, Cache and Other.

# 4. EXPERIMENT DETAILS

## 4.1 Classification [10]

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attributeWe have used NaiveBayes classifier for all tools. A NaiveBayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions.

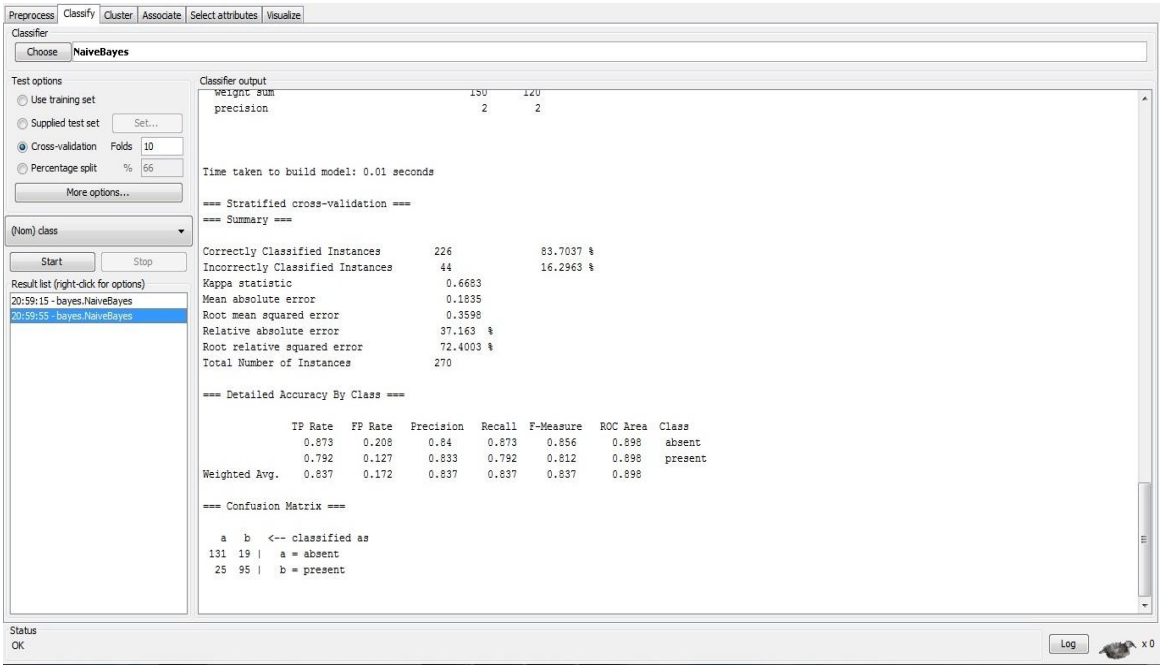



**Fig 1: Classification Using WEKA**

In above Fig.1, we have seen that correctly classified and incorrectly classified instances for WEKA. In WEKA if we used classification then statistical analysis like mean, variance is shown. For KNIME as shown in Fig.2 below, we need to add prediction node after the classification then only the accuracy can be found. In XLMiner we can directly apply a classifier and in that we can get accuracy with confusion matrix.

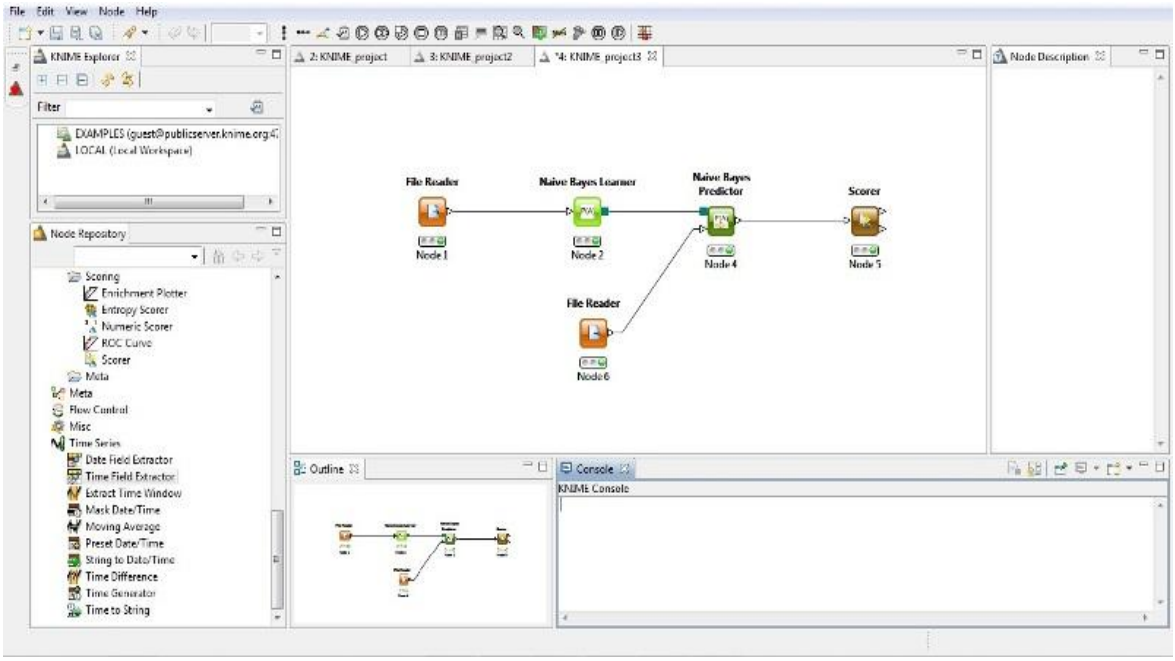


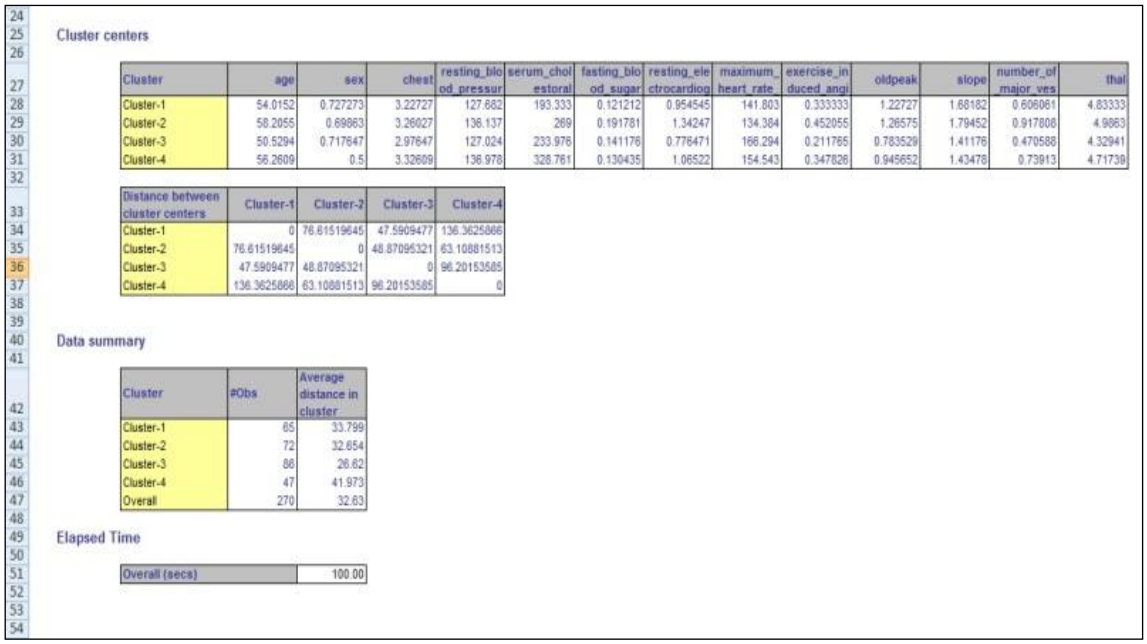

**Fig 2: Classification Using KNIM**

## 4.2 Clustering [11]

Clustering techniques are applied when class labels are not known in advance. It comes under unsupervised learning. For analysis we are using K-Means algorithms in all tools. This algorithm aims at minimizing an objective function, in this case a squared error function (Refer Equation 1) [12]. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| X_i^{(j)} - C_j \right\|^2 \quad \ldots (1)$$

In above equation $\left\| X_i^{(j)} - C_j \right\|^2$ is a chosen distance measure between a data point $X_i^{(j)}$ and the cluster centre $Cj$, is an indicator of the distance of the $n$ data points from their respective cluster centers.



Cluster centers

| Cluster | age | sex | chest | resting_blood_pressur | serum_cholestoral | fasting_blood_sugar | resting_electrocardiog | maximum_heart_rate | exercise_induced_angi | oldpeak | slope | number_of_major_ves | thal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster-1 | 54.0152 | 0.727273 | 3.22727 | 127.682 | 193.333 | 0.121212 | 0.954545 | 141.803 | 0.333333 | 1.22727 | 1.68182 | 0.606061 | 4.83333 |
| Cluster-2 | 58.2055 | 0.69863 | 3.26027 | 136.137 | 269 | 0.191781 | 1.34247 | 134.384 | 0.452055 | 1.26575 | 1.79452 | 0.917808 | 4.9863 |
| Cluster-3 | 50.5294 | 0.717647 | 2.97647 | 127.024 | 233.976 | 0.141176 | 0.776471 | 166.294 | 0.211765 | 0.783529 | 1.41176 | 0.470588 | 4.32941 |
| Cluster-4 | 56.2609 | 0.5 | 3.32609 | 136.978 | 328.761 | 0.130435 | 1.06522 | 154.543 | 0.347826 | 0.945652 | 1.43478 | 0.73913 | 4.71739 |

| Distance between cluster centers | Cluster-1 | Cluster-2 | Cluster-3 | Cluster-4 |
|---|---|---|---|---|
| Cluster-1 | 0 | 76.61519645 | 47.5909477 | 136.3625866 |
| Cluster-2 | 76.61519645 | 0 | 48.87095321 | 63.10881513 |
| Cluster-3 | 47.5909477 | 48.87095321 | 0 | 96.20153585 |
| Cluster-4 | 136.3625866 | 63.10881513 | 96.20153585 | 0 |

Data summary

| Cluster | #Obs | Average distance in cluster |
|---|---|---|
| Cluster-1 | 65 | 33.799 |
| Cluster-2 | 72 | 32.654 |
| Cluster-3 | 86 | 26.62 |
| Cluster-4 | 47 | 41.973 |
| Overall | 270 | 32.63 |

Elapsed Time

| | |
|---|---|
| Overall (secs) | 100.00 |

**Fig 3: Clustering Using XLMiner**

Fig. 3 shows clustering results using XLMiner. Four clusters are formed based on each attribute. In 2nd table distance between cluster centers is given. In 3rd table average distance in cluster is mentioned. KNIME provide better understanding and different symbol are used. In XLMiner we need to select attribute each time and then it give better analysis. Analysis of tools is given in Table 1. 'Y' indicates the left hand column feature is supported by respective tool.

**Table 1. Features Supported by Tools**

| Features | WEKA | KNIME | XLMiner |
|---|---|---|---|
| Data Range | | | Y |
| Distance Function | Y | | |
| Input Variable | | Y | Y |
| No. of Clusters | Y | Y | Y |
| No. of Iterations | Y | Y | Y |

## 4.2 Association Rule Mining

Association rules are if/then statement that helps uncover relationship between seemingly unrelated data in a relationship database or other information repository. An example of an association rule would be "If person age is greater or equal to 18years then only he can vote". Generating rule we have using Apriori algorithm and also calculate support and confidence.

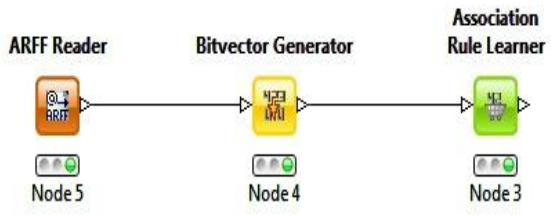



**Fig. 4 Association Rule Mining in KNIME**

Before applying association rule mining on a dataset we need to preprocess it. As shown in Fig. 4 Node 5 is for ARFF file reader, Node 4 is Bivector Generator used for preprocessing of data and Node 3 as Association Rule Learner. Below Table 2 gives basic requirements while performing association rule mining using different tools. These basic requirements must be satisfied before rule generation.

**Table 2. Rule Mining Features**

| Features | WEKA | KNIME | XLMiner |
|---|---|---|---|
| Pre-Processing | Y | Y | |
| Rule generation Count | Y | | |
| Support Count | Y | Y | Y |
| Confidence | Y | Y | Y |
| Item Set | Y | Y | Y |

# 5. RESULT ANALYSIS & DISCUSSION

Experiment is performed using tools on data mining techniques as explained above. To analyse these tools following criteria are chosen:

## 5.1 Accuracy

From the below Table 3, we can see that XLMiner gives higher accuracy for NaiveBayes Classification. WEKA is comparatively giving less accuracy than other two tools. KNIME is somewhere near to WEKA but better than it Accuracy is calculated in a percentage (%).

**Table 3 Classification Accuracy for Different Tools**

| Tools | Accuracy (%) |
|---|---|
| WEKA | 83.70 |
| KNIME | 85.18 |
| XLMINER | 91.11 |

## 5.2 GUI

If we compare tools in the GUI perspective, then KNIME and XLMiner are quite good. For beginners it is easy to work with these tools. WEKA is also user friendly but it takes time to understand how to work with it. In case of XLMiner, we have to just manage Add-Ins for XLMiner in our Excel Sheet. If someone knows how to work with Excel then it is trivial to operate XLMiner. We need to understand how data mining take place, flow of operations. As per flow of operation we need to add nodes and manage links among them. We can configure nodes as per requirement.

## 5.3 Algorithms

WEKA supports different algorithms for data mining. Each strategy has various algorithms in it. In XLMiner Education edition, there are few algorithms supported. For ARM both XLMiner and KNIME does not provide special algorithm. But WEKA has different algorithms for ARM such as, Apriori, Tertius, Predictive Apriori, etc. Except ARM, KNIME supports lot of algorithms for other approaches.

## 5.4 Elapsed Time

Each tool has different processing time as shown in Table 4 below

**Table 4 Processing time for WEKA and XLMiner**

| Features | WEKA | KNIME |
|---|---|---|
| Pre-Processing | Y | Y |
| Rule generation Count | Y | |
| Support Count | Y | Y |
| Confidence | Y | Y |
| Item Set | Y | Y |

## 5.5 Operating System [OS] Support

Table 5 represents the mapping of different operating system verses different tools

**Table 5 Os Support for Tools**

| Features | WEKA | KNIME | XLMiner |
|---|---|---|---|
| Windows Server 2000 | | Y | |
| Windows 2000 | | Y | |
| Windows Me, 98/95 | | Y | |
| Windows x86, x64 | Y | Y | Y |
| Linux | Y | Y | Y |
| Mac OS x (With Java Support) | Y | | Y |

## 5.6 Other

We also consider some other factor for analysis as follow

1) KNIME supports image mining. WEKA supports only Image classification. XLMiner does not support this feature.
2) In KNIME various report items available such as Chart, label, text, data, image, grid, list, and table. But WEKA and XLMiner has only limited report items.
3) Other than java XLMiner supports other languages such as C (Procedural), C++, C#, Visual Basic, VB.NET and MATLAB

# 6. CONCLUSION AND FUTURE WORK

In order to compare different data mining tools we have carried out comparative analysis of those tools for classification, clustering and association rule mining using Statlog heart dataset. According to our results WEKA provides different algorithms for data mining techniques also processing time required is less. KNIME has better GUI so understandability of flow is good for beginners. It is a user friendly tool. Results may vary with different datasets. Thus as discussed in this paper, using efficient tool critical patients can be found and by treating them death rate can be controlled.

# 7. ACKNOWLEDGMENTS

The authors thank to Pimpri Chinchwad College of Engineering for giving an opportunity to publish the paper. The authors also thank University of Waikato for WEKA KNIME and XLMiner tool availability as an open source.

# 8. REFERENCES

[1] A. M. Khattak, A. M. Khan, Sungyoung Lee, Young-Koo Lee, "Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka", International Journal of Database Theory and Application Vol. 3, No. 1, March, 2010

[2] M.Vijayakamal, Mulugu Narendhar "A Novel Approach for WEKA & Study On Data mining Tools", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 2, August 2012

[3] http://www.cs.waikato.ac.nz/ml/weka

[4] http://www.solver.com/xlminer-data-mining

[5] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Giuseppe Di Fatta, Thomas R. Gabriel, Florian Georg, Thorsten Meinl, Peter Ohl, Christoph Sieb, and Bernd Wiswedel. "Knime: The Konstanz Information Miner", White paper

[6] http://repository.seasr.org/Datasets/UCI/arff

[7] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques, Third Edition"

[8] P.Yasodha, M. Kannan "Analysis of a Population of Diabetic Patients Databases in Weka Tool", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011 ISSN 2229-5518

[9] http://www.knime.org

[10] C.V.Subbulakshmil, S.N.Deepa, N.Malathi, "Comparative Analysis of XLMiner and WEKA for Pattern Classification", 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies

[11] Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[12] P.Isakki alias Devi, S.P.Rajagopalan "Analysis of Customer Behavior using Clustering and Association Rules", International Journal of Computer Applications (0975 – 8887) Volume 43– No.23, April 2012