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 of Data Mining Tools: A Case Study Based on Classification Algorithms and Datasets

Nidhi Sharma*, Dr. K L Bansal
Department of Computer Science, HPU,
Shimla, India

Abstract— Data mining is the process of extracting data, analyzing it from every possible dimensions or perspectives, and then producing a summary of the information or knowledge extracted in a useful form that identifies relationships or pattern within the dataset of the application domain. The fundamental algorithms and techniques of data mining and analysis form the strong base for the emerging field of data science that comprises of automated methods to analyze models or patterns of the dataset, with applications ranging from scientific discovery to business intelligence and analytics. There are many techniques of data mining some includes classification, clustering, association and prediction. The data mining technique used in this paper is classification in order to compare the data mining tools. In order to analyse the performance of data mining tools, they run on three different dataset which are having different number of attributes and different number of instances. The evaluation criterion used for the evaluation of tool in this paper is classification accuracy. The analysis process is grouped into three test cases so as to compare the performance of data mining tools in various conditions that is by varying the application domain and classification algorithm.

Keywords— Data Mining, Classification, Knime, Orange, Weka, Rapid Miner, decision tree, Random Forest, KNN

I. INTRODUCTION

A. Data Mining

Nowadays, due to the wide availability of huge amounts of data and the immense need for turning such data into useful information and knowledge, data mining has attracted a great deal of attention in the information industry and business intelligence and analysis. The information and knowledge extracted by using data mining can be used for applications ranging from fraud detection, market analysis, and customer retention, to science exploration and production control [1]. The basis for the emerging field of data science, which includes automated methods to analyze patterns of dataset and models for all kinds of dataset, with applications are deducted from the fundamental algorithms in data mining and analysis. It is helpful in fields ranging from scientific discovery to business intelligence and analytics.

Data mining can be viewed as a by-product of the natural evolution of information technology. Data mining comprises of many algorithms that enable one to gain fundamental insights and knowledge from huge ore of data [2]. It is an interdisciplinary field merging concepts from allied areas such as pattern recognition, statistics, machine learning, and database systems. In fact, data mining itself is part of a larger knowledge discovery process. That knowledge discovery process includes pre-processing tasks such as data extraction, data cleaning, data fusion, data reduction, feature construction and so on, as well as post-processing steps such as pattern and model interpretation, hypothesis confirmation and generation, and so on[3].

B. Classification

Classification is a two-step process. This first step of the classification process comprises of learning a mapping or function, y = f (X), that predict the associated class label y of a given tuple X. for this purpose, first task is to learn a mapping or function that separates the data classes. Generally, this mapping is represented in the form of classification rules, decision trees, neural networks or mathematical formulae [4]. The mapping can be used to categorize unknown tuples into their respective classes and also provide a compressed representation of the data.

In the second step, the model is used for classification. Initially the predictive accuracy of the classifier is estimated. To use training set for estimating the accuracy of the classifier is not a very good idea as the classifier tends to over fit the data. Over fitting is the result of incorporating some particular anomalies of the training data that are not present in the general data set overall. Therefore a dataset known as test data set is used with training set, made up of test tuples and their associated class labels [5]. These tuples are randomly selected from the training data set used for the learning purpose. They are independent of the training tuples which mean that they are not used to construct the classifier. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier after learning from the provided training set. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. There are several methods for estimating classifier accuracy such as cross validation, bootstrapping and percentage split. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify new data tuples for which the class label is not known.

C. Classification algorithm

KNN

The k-nearest-neighbor method was first described in the early 1950s. Although it has been used for many purposes but mostly used for pattern recognition. Nearest-neighbor classifiers are based on learning by analogy. It compares a given test tuple with training tuples that has got similar attributes. The training tuples consists of n attributes which represents a point in an n-dimensional space [6]. Likewise, an n-dimensional pattern space is created by n training set tuples. When an unknown tuple is to be classified, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that has least distance from the unknown tuples. These k training tuples are the k "nearest neighbors" of the unknown tuple. The nearest neighbors are defined in terms of Euclidean distance. The target function could be discrete or continuous valued. It is robust to noisy data by averaging k-nearest neighbors. It can be used both for real valued target function and continuous- valued target function.

Decision Tree

Decision tree induction is the algorithm for constructing decision trees from tuples of class-labeled training data. A decision tree is a flowchart-like tree structure [7]. The topmost node in a tree is the root node of the decision tree. In the decision tree formed each internal node denotes a test on an attribute [8]. The branches of decision tree represent outcome of the test, and each terminal node holds a class label. The tree produced by decision tree algorithms can be only binary trees (where each internal node splits into two branches), whereas others can produce non-binary trees.

Decision trees are best suited for classification. Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple[9]. Decision trees can easily be converted to classification rules. Certain types of decision trees are constructed using different working algorithms such as ID3, C4.5, CART, LMT etc. The main difference lies in the splitting criteria of decision tree such as gini index or information gain.

Random Forest

Random Forests is considered as ensemble algorithm that grows many classification trees. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. To classify a new object from an input vector, start with root to down branches, each of the trees in the forest. Each decision tree gives a classification or in other words the tree "votes" for that class. The forest chooses the classification having the most votes over all the trees in the forest. It has an effective method for estimating the missing data. It can also maintain accuracy when a large proportion of the data are missing. The forest error rate depends on two things. First the correlation between any two trees in the forest and second the strength of each individual tree in the forest. Reducing m reduces both the correlation and the strength. Increasing it increases both. M is the only parameter to which the random forest is sensitive to.

Each tree is grown as follows:

- If the number of cases in the training set is N, N cases are selected at random (with replacement), from the original data. This sample will be considered as the training set for growing the tree.
- If there are M input variables, a number m<<M is specified. m variables are selected at random out of M at each node and the best split on these m is used to split the node. The value of m is held constant during the entire forest growing.
- Each tree is grown to the largest extent possible. There is no pruning.

D. Data Mining Tools

Weka

The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning toolkit introduced by Waikato University, New Zealand. It is open source software written in Java (GNU Public License) and used for research, education and Projects [10]. It can be run on Windows, Linux and Mac. It comprises of many machine learning algorithms for implementing data mining tasks easily. The most daunting task in the comparative study is to collect the huge amount of data to extract information from it. WEKA has made this task easier by data pre-processing. WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers can implement new algorithms without giving concern to supporting infrastructure for data manipulation and evaluation of scheme[11]. WEKA provides various sources for loading data, including files, URLs and databases. It supports file formats include WEKA"s own ARFF format, CSV, Lib SVMs format, and C4.5's format. If applicable, the panel also provides access to graphical representations of models, e.g. decision trees. Moreover, it can visualize prediction errors in scatter plots, and also allows evaluation via AUC curve, ROC curves and other "threshold curves". It is also possible to generate data using an artificial data source and edit data manually using a dataset editor. Many evaluation criteria are also provided such as confusion matrix, recall value, TP and FN values [12]. Models can also be saved for future use and loaded later. Along with supervised algorithms, WEKA also supports application of unsupervised algorithms.

Rapid Miner (RM)

RAPID MINER (RM) provides a good environment for machine learning and data mining processes. It is open-source, free project implemented in Java. The tool has become very popular in several recent years and has a large community support [13]. The drag-and-drop of operators is used to construct the dataflow and the inputs and outputs of

corresponding operators are connected to each other. It represents a new approach to design even very complicated for a huge number of learning problems. Rapid Miner has flexible operators for data input and output in different file formats. It contains more than 100 learning schemes for regression, classification and clustering tasks. The file extensions supported by the data handling is transparent to the operators. It does not give the headache to mess with the actual data format or different data views - the Rapid Miner core takes care of all necessary transformations [14]. Rapid Miner are Aml, Arff, Att, Bib, Clm, Cms, Cri, Csv, Dat, Ioc, Log, Matte, Mode, Obf, A Bar, One Pair, Res, Sim, Thr, Wgt, Wls, Xrff. It supports a large number of the classification and regression, decision trees, Bayesian, logical clusters, association rules, and many algorithms for clustering (k-means, k-medoids, dbscan), and many features for data pre-processing, normalization, features such as filtering, genetic algorithms, neural networks, 3D and data analysis. A huge amount of visualization techniques and the possibility to place breakpoints after each operator give insight into the success of design. It can import data from Oracle, Microsoft SQL Server, PostgreSQL, or MySQL databases.

Knime

KNIME (Konstanz Information Miner) is a general purpose data mining tool implemented on the Eclipse platform, developed and maintained by the Swiss company. It is a user-friendly and interacting open-source data integration, processing, exploration, and analysis platform. From day one, KNIME has been developed using efficient software engineering practices [15]. It is currently being used actively by over 6,000 professionals all over the world, in both industry and academic application. KNIME provides user with a data exploration platform that enables the user to visually create a desired data flows, execute analysis step selectively, and investigate the results later through interactive views on data and models.

The tool adheres to the visual programming paradigm. All the building blocks are placed on a canvas and are connected to obtain a visual program of data mining. In KNIME, these building blocks are called nodes, and according to the official website, more than1000 nodes are available through the core installation and various extensions. A large repository of example workflows is provided to facilitate quicker learning of the tool [16]. Nodes are grouped under the class belonging and can be searched by names too. It has got fully documented algorithms, and the documentation is automatically shown within the interface once the node is selected. One of the alluring features of KNIME is the integration with other data mining tools such as R and WEKA. The integration feature with WEKA gives user to use all the functionality of WEKA as KNIME nodes. It can also run R code as a step in the workflow, opening R views and learning models within R, due to its integration with R. Several other interesting free extensions are also available, e.g., OpenStreetMap and JFreeChart extension that enables working with geographical data and advanced charting respectively [17]. There are also commercial extensions for more specific functionalities.

Orange

ORANGE is a Python-based tool for Data Mining being developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana. It can be used either through Python scripting or through visual programming. Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use ORANGE as a Python library for data manipulation and widget alteration. Through visual programming, widgets can be assembled together into an application by a visual programming tool called Orange Canvas. Its visual programming interface, Orange Canvas, offers a structured view of supported functionalities grouped into nine categories. Functionalities are visually represented by different widgets including a short description. Programming is performed by placing widgets on the canvas and connecting their inputs and outputs. The interface provided by ORANGE is visually appealing, which provides user a pleasant experience to work with. All these features combining together make the ORANGE tool, a comprehensive, component-based framework for machine learning and data mining. It can be easily used by both researchers of data mining and inexperienced users who want to develop and test their own algorithms. It gives advantage of reusing as much of the code as possible.

II. DATASET DESCRIPTION

The dataset used in this paper is from UCI Repository. There are three dataset, namely FERTILITY, HEPATITIS and BALLOON, used for analysis purpose with different number of attributes and instances so as to evaluate the behaviour of data mining tools with different parameters. The description of three dataset is shown in table1 below.

TABLE1 PARAMETERS OF DATASET

Dataset Name	**No. of Instances**	**No. of Attributes**	**Missing data**	**Dataset**
Hepatitis	155	19	Yes	Multivariate
Fertility	100	10	No	Multivariate
Balloons	16	4	No	Multivariate

The Parameter of Fertility Dataset are shown in table2

TABLE2 PARAMETERS OF FERTILITY DATASET

Parameter	**Category**

Season	Nominal
Age	Nominal
Childish_Diseases	Nominal
Accident	Nominal
Surgical_Intervention	Nominal
High_Fevers	Nominal
Frequency	Nominal
Smoking_Habit	Nominal
No. of Hours	Nominal
Output	Nominal

The Parameter of Hepatitis Dataset are shown in table3

TABLE3 PARAMETERS OF HEPETITIS DATASET

Parameter	**Category**
Class	Nominal
Age	Numerical
Sex	Nominal
Steroid	Nominal
Antivirals	Nominal
Fatigue	Nominal
Malaise	Nominal
Anorexia	Nominal
Liver_big	Nominal
Liver_firm	Nominal
Spleen_palpable	Nominal
Spiders	Nominal
Ascites	Nominal
Varices	Nominal
Bilirubin	Numerical
Alk_phosphate	Numerical
Sgot	Numerical
Albumin	Numerical
Protime	Numerical

The Parameter of Balloon Dataset are shown in table4

TABLE4 PARAMETERS OF BALLOON DATASET

Parameter	**Category**
Color	Nominal
Size	Nominal
Act	Nominal
Age	Nominal
Inflated	Nominal

III. ANALYSIS AND RESULTS

The four data mining tools are analyzed and their performances are compared on the basis of their working on three different dataset which are having different parameters and by using three different classification algorithm. The measure taken for comparing four data mining tools is classification accuracy. The study is divided into three test cases: on the basis of varying dataset and on the basis of varying classification algorithm.

Test case 1:

All three dataset having different number of instances and attributes are analysed separately on each data mining tool and classification algorithm.

The results of applying three different classification algorithms on FERTILITY dataset implementing it on all four data mining tool is shown in figure 1.all tools works almost same on this dataset.

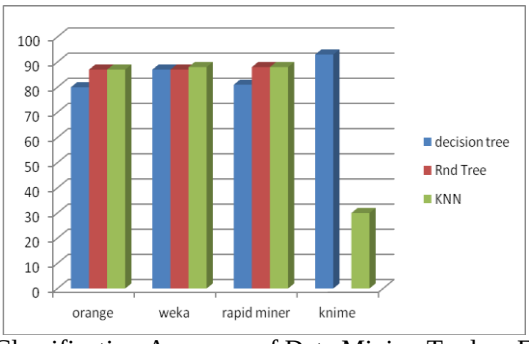

Fig 1: Classification Accuracy of Data Mining Tool on Fertility

On applying the same algorithms and data mining tools on BALLOON dataset which is having smaller number of attribute and instances, the behaviour is shown in figure 2 below. It can be observed that almost all tools do not work so well in small instances as the training data is lesser for the learning process.

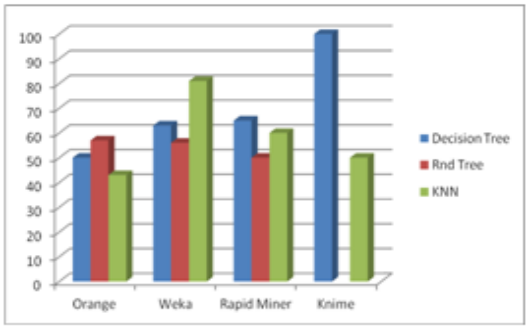

Fig 2: Classification Accuracy of Data Mining Tool on Balloon

On using HEPATITIS as a dataset, to evaluate the performance of data mining tools, it is observed that by increasing the dataset size by increasing the number of attributes and instances, the performance of tool is not as good as in intermediate dataset but much better than in small dataset. Knime is showing best classification accuracy.

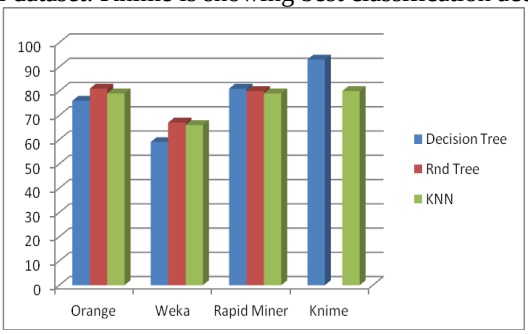

Figure 3: Classification Accuracy of Data Mining Tool on Hepatitis

Test case 2

In this test case, all three data mining tools are run using three classification algorithms namely, Decision Tree, Random Forest and KNN.

In figure 4, shows the classification accuracy of data mining tools using classification algorithm Decision Tree. KNIME is again showing unexpected behaviour in case of small dataset balloon by showing highest classification accuracy, otherwise all three data mining tools are showing worst classification accuracy in case of small dataset. The best classification accuracy is shown in case of intermediate dataset (FERTILITY). And they are working average while learning with large dataset (HEPATITIS).

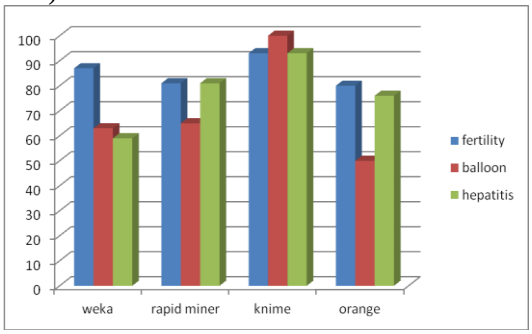

Figure 4: Performance of Tool on Decision Tree

Figure 5 shows the working of all data mining tool with classification algorithm Random Forest. Random forest is not supported by knime. All three data mining tools are working well in intermediate dataset. And the classification accuracy is reducing to large level when worked with less instances of dataset. The classification accuracy of large dataset is average in three data mining tool.

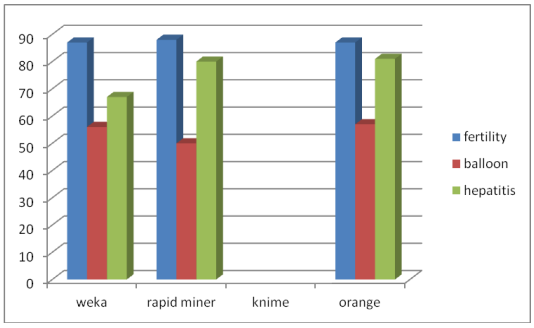

Figure 5: Performance of Tool on Random Forest

In figure 6, the performance of all four data mining tool is analysed using classification algorithm KNN. KNIME is showing unexpected behaviour in case of intermediate dataset, otherwise all the other data mining tools are working excellent in large dataset. The behaviour of all four data mining tool is good in case of large dataset that is HEPATITIS. And the classification accuracy is not satisfactory in case of BALLOON which has least number of instances and attributes.

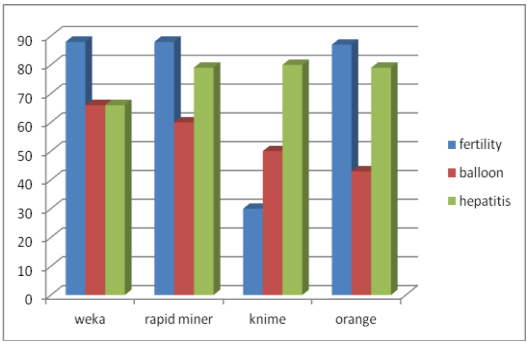

Figure 6: Performance of Tool on KNN

Test case 3

In this case, the data mining tools are run separately on each dataset and classification algorithm.

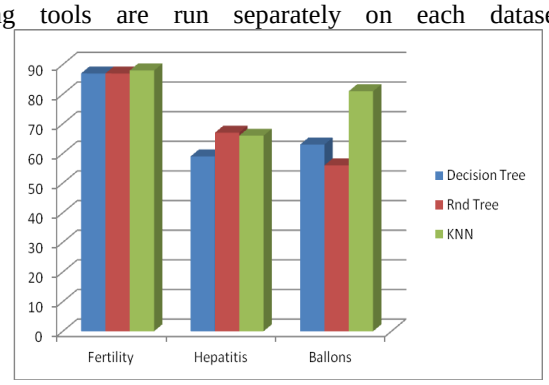

Fig8: performance of weka

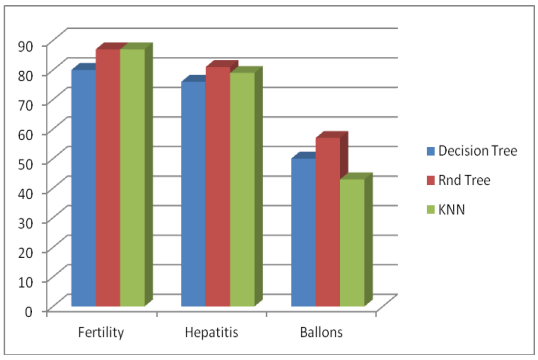

Fig9: performance of orange

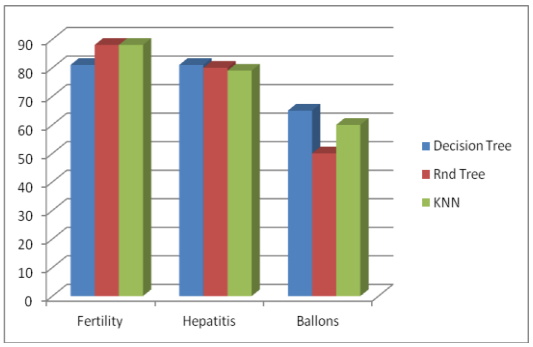

Fig10: performance of rapid miner

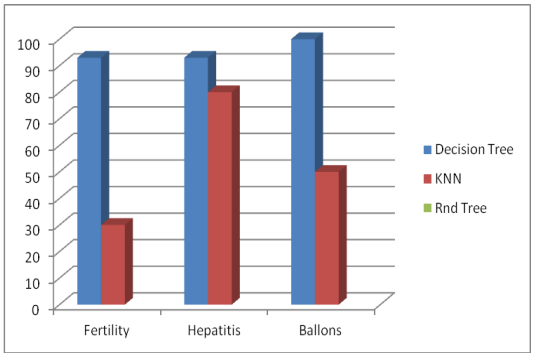

Fig 11: performance of knime

On the basis of three test cases, the behaviour of tools with different classification algorithm and different dataset with distinct number of attributes and instances is concluded in the form of tables shown below.

TABLE5 PERFORMANCE OF TOOLS USING DIFFERENT CLASSIFICATION ALGORITHM

Tools	**DecisionTree**	**KNN**	**RndTree**
Weka	Good	Excellent	Bad
Orange	Bad	Good	Excellent
RapidMiner	Excellent	Good	Bad
Knime	Excellent	Good	_

TABLE6 PERFORMANCE OF TOOLS USING DIFFERENT DATA SET

Tools	**Hepatitis(large)**	**Fertility(intermediate)**	**Balloon(small)**
Weka	Bad	Excellent	Good
Orange	Good	Excellent	Bad
RapidMiner	Good	Excellent	Bad
Knime	Excellent	Bad	Good

IV. CONCLUSIONS AND FUTURE SCOPE

Data mining provides with the techniques of database applications that look for hidden patterns or knowledge in a group of data that can be used to predict future behaviour of the newly entered data. Classification being one of the techniques of data mining consists of assigning a class label to a set of unclassified cases. The main objective of the work is to compare the data mining tools on the basis of their classification accuracy. The work has been divided into three test cases. Each case is observing the behavior of tools with different parameters by varying the size of dataset and varying the classification algorithms.

According to the result of three test cases used in this work, it has been observed that different data mining tools are acting differently on the application of different classification algorithm and different dataset. All the data mining tools are performing excellently with intermediate dataset instead of small or very large dataset. With small dataset the tools are not able to learn efficiently so the classification accuracy of prediction is not getting up to the mark. Larger training set can lead to over fitting. WEKA is showing best classification accuracy with algorithm KNN and KNIME with Decision Tree. The classification algorithm which is showing best results in ORANGE is Random Forest. The best classification accuracy of Rapid Miner can be seen with the classification algorithm Decision Tree. So it can be concluded although every data mining tool is behaving differently with each classification algorithm but WEKA is showing almost constant performance in each classification algorithm.

For the future scope, more dataset can be used and some different data mining algorithm such as clustering can be used in order to compare the performance of data mining tool. Some different measure can be considered for the evaluation purpose.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann, "Data Mining: Concepts And Techniques", Morgan Kaufmann Publishers, 3rd Edition, (2011).

[2] Gordon S. Linoff, Michael J. A. Berry, "Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management", Morgan Kaufmann Publishers, 3rd Edition, (2011).

[3] Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat," A Review Paper On Various Data Mining Techniques", International Journal Of Advanced Research In Computer Science And Software Engineering ,Vol:4,Issue 4, (2014)

[4] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, And Emad M. Al-Shawakfa," A Comparison Study Between Data Mining Tools Over Some Classification Methods", International Journal Of Advanced Computer Science And Applications (IJACSA), Special Issue On Artificial Intelligence.

[5] Kalpana Rangra, Dr. K. L. Bansal, " Comparative Study Of Data Mining Tools", International Journal Of Advanced Research In Computer Science And Software Engineering, Vol: 4, Issue 6, (2014).

[6] Mehmed Kantardzic, "Data Mining: Concepts, Models, Methods, And Algorithms", John Wiley And Sons Inc, 2nd Edition, (2011).

[7] Devi Prasad Bhukya, S. Ramachandram, "Decision Tree Induction: An Approach For Data Classification Using AVL-Tree", International Journal Of Computer And Electrical Engineering, Vol.2, No.4, (2010).

[8] Pooja Sharma,.Rupali Bhartiya,"Implementation Of Decision Tree Algorithm To Analysis The Performance", International Journal Of Advanced Research In Computer And Communicationengineering,Vol. 1, Issue 10,(2012).

[9] Ruijuan Hu, "Medical Data Mining Based On Decision Tree Algorithm"Computer And Information Science, Vol. 4, No. 5 ;(2011) .

[10] Trilok Chand Sharma, Manoj Jain," WEKA Approach For Comparative Study Of Classification Algorithm", International Journal Of Advanced Research In Computer And Communication Engineering,Vol. 2, Issue 4,(2013).

[11] Mahendra Tiwari, Manu Bhai Jha, Omprakash Yadav , "Performance Analysis Of Data Mining Algorithms In Weka", IOSR Journal Of Computer Engineering (IOSRJCE),Vol:6, Issue 3, (2012).

[12] Bernhard Pfahringer, Peter Reutemann, Ian H.Witten, Eibe Frank, Geoffrey, Holmes Mark Hall," The WEKA Data Mining Software: An Update", SIGKDD Explorations.

[13] Benjamin Schowe,"Feature Selection For High-Dimensional Data With Rapid Miner", Technical University Of Dortmund Artificial Intelligence Group, (2011).

[14] Christian Bockermann and Hendrik Blom, "Processing Data Streams With The Rapid Miner Streams Plug In", (2015).

[15] Stephan Beisken, Thorsten Meinl, Bernd Wiswedel, Luis F De Figueiredo, Michael Berthold And Christoph Steinbeck,"KNIME-CDK: Workflow-Driven Cheminformatics".

[16] C.Sieb, T.Meinl, M.R.Berthold, "Parallel And Distributed Data Pipelining With Knime", The Mediterranean Journal Of Computers And Networks, Vol. 3, No. 2, (2007).

[17] Lalit Kataria,"Implementation Of Knime-Data Mining Tool", International Journal Of Advanced Research In Computer Science And Software Engineering, Computers And Networks, Volume 3, Issue 11, (2013).

