International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Comparison of Data Mining Tools using the implementation of C4.5 Algorithm

Divya Jain

School of Computer Science and Engineering, ITM University, Gurgaon, India

Abstract: ***This paper presents the implementation on a healthcare dataset using data mining tools to find important parameters that reflect the effect of diabetes on kidney of patients. This is done with the use of Kidney Function Tests (KFT). The data mining tools used are Tanagra and Weka with the application of C4.5 Algorithm which is based on decision trees. This paper compares the result given by Weka and Tanagra. The outcome of both the tools is analyzed and conclusion is drawn that both the tools are able to work well on dataset but Tanagra is more efficient and less error-prone in terms of the performance of the classifier. The effective usage of data mining tools enables us to find important parameters that reflect the effect of diabetes on kidney. Additionally, it is found that the performance of Weka is best when used with "Use Training Set" mode than with cross validation followed by percentage split mode for training the classifier.***

Keywords: Weka, Tanagra, Diabetes, Classification, Kidney

1. Introduction

Data mining [1] is one of the most important domains which help in management of healthcare data. It also helps to discover new trends from healthcare data collected from various hospitals. The data mining tools and techniques help in analyzing data collected from different hospitals and summarizing it into useful information [2]. There are huge applications of data mining in healthcare sector like providing effective treatment, customer relationship management; detecting fraud and. Diabetes [3] is a disease which can lead to other diseases like kidney disease**,** heart disease, etc. The effect of diabetes on kidneys is very substantial. Classification and prediction techniques [4] have been found to be successful in finding the effect of diabetes on kidney of patients.

2. Methodology

2.1 Kidney Function Tests (KFT)

Kidneys play vital role for proper maintenance of health. Kidneys are essential for filtering wastes from blood and removing them from body as urine [5]. Kidney Function Tests are done to find various aspects related to kidney and to have a check on kidney disorders [6]. These tests help us to know whether our kidneys are working properly or not. These tests give us indication of the performance of kidneys in the removal of wastes from human body. When a person wants to check the functioning of kidneys, they go for Kidney Function Tests (KFT). Diabetes has a significantly great effect on the working of kidneys. High blood glucose due to diabetes can damage kidneys severely and can even stop their proper functioning if its effect is not reduced on time. Long term association with diabetes can lead to kidney disease called "Nephropathy" [7]. According to the literature, around one third of people suffering from diabetes for 15 years will definitely be suffering from kidney disease [7]. If we keep our blood sugar and blood pressure in control, we can prevent the occurrence of diabetic kidney disease. There are various tests to check kidney function tests:

- Blood Urea
- Serum Creatinine
- Uric Acid
- Total Protein
- Albumine
- Blood Sugar

2.2 Algorithm Used

The algorithm used to implement classification technique using data mining tools is C4.5 Algorithm [8]. This algorithm is used to generate decision trees from the dataset. Decision tree induction is a powerful method for classifying datasets and extracting rules from huge databases [9]. C4.5 Algorithm is named as J48 Algorithm in Weka for its implementation [10]. There are several applications of classification like weather forecasting, diagnosis of various faults, recognition of patterns etc.

2.3 Weka Tool

Weka [11] is an open source tool for the implementation of various data mining algorithms. It is based on java application and was first given by University of Waikato in New Zealand [12]. It is named after the bird "Weka" which is found in New Zealand. Weka toolkit consists of a large number of machine learning algorithms written in java. Weka implementation [13] of C4.5 Algorithm is named as J48 Algorithm. We can use this software through interactive GUI (Graphical User Interface) as well as through command line. It provides an influential interface for the construction of decision trees. Weka provides fairly good solutions to many problems. Through this software, several experiments are implemented by researchers to get knowledge of different methods and algorithms.

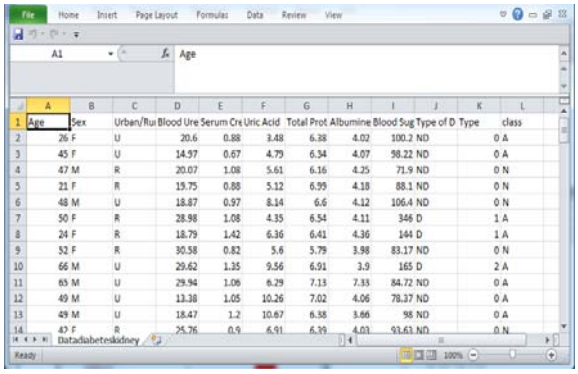
2.4 Tanagra Tool

Tanagra [14] is an open source data mining tool which has wide applications in research area. It is a simple, easy to use and understand software. It is a freely available machine learning tool given by Ricco Rakotomalala [15]. This machine learning framework is used commonly by students and researchers because of its simplicity and interactive GUI associated with it. This tool can be used extract knowledge from huge databases. It has a strong ability to mine data effectively to get useful and required information. It is an academic tool which supports the implementation of different algorithms in data mining.

3. Implementation on Tools

3.1 Dataset Description

The dataset consisting of records of 100 patients is collected from **Jyoti Diagnostic and Research Centre, Gurgaon**. The dataset consists of 12 attributes. Some of the attributes are related to Kidney Function Tests, while some are related to diabetes. As we are applying classification technique, the last attribute is "class" which has 2 values – A (Affected) and N (Not Affected). With the help of classification using decision trees, the diabetic effect on kidney is found out. Data mining tools are used to accomplish this task. Both data mining tools (Weka and Tanagra) are given learning using classification technique creating a learning model. For this, we apply classification algorithm called C4.5 Algorithm in both Weka & Tanagra. This algorithm is named as J48 algorithm in Weka (java implementation of C4.5 Algorithm). The algorithms are applied in both the tools and decision trees are generated using supervised learning finding the effect of diabetes on kidney of patient.



Age	Sex	Urban/Rur	Blood Ure	Serum Cre	Uric Acid	Total Prot	Albumine	Blood Sug	Type of D	Type	class
26	F	U	20.6	0.88	3.48	6.38	4.02	100.2	ND		0 A
45	F	U	14.97	0.67	4.79	6.34	4.07	98.22	ND		0 A
47	M	R	20.07	1.08	5.61	6.16	4.25	71.9	ND		0 N
21	F	R	15.75	0.88	5.12	6.99	4.18	88.1	ND		0 N
48	M	U	18.87	0.97	8.14	6.6	4.12	106.4	ND		0 N
50	F	R	28.98	1.08	4.35	6.54	4.11	346	D		1 A
24	F	R	18.79	1.42	6.36	6.41	4.36	144	D		1 A
52	F	R	30.58	0.82	5.6	5.79	3.98	83.17	ND		0 N
66	M	U	29.62	1.35	9.56	6.91	3.9	165	D		2 A
65	M	U	29.94	1.06	6.29	7.13	7.33	84.72	ND		0 A
49	M	U	13.38	1.05	10.26	7.02	4.06	78.37	ND		0 A
49	M	U	18.47	1.2	10.67	6.38	3.66	98	ND		0 A
47	F	R	25.76	0.9	6.91	6.39	4.03	93.61	ND		0 N

Figure 1: Dataset

3.2 Classification in Weka

- First, we open Weka, then select explorer option from right hand side. After that, we use preprocess tab to import our dataset which is in csv format. Weka provides filters for preprocessing tasks. But as J48 Algorithm works well with a mixture of both categorical and continuous attributes, it is not required in our implementation. This presents all attributes from the dataset as shown in Fig. 2

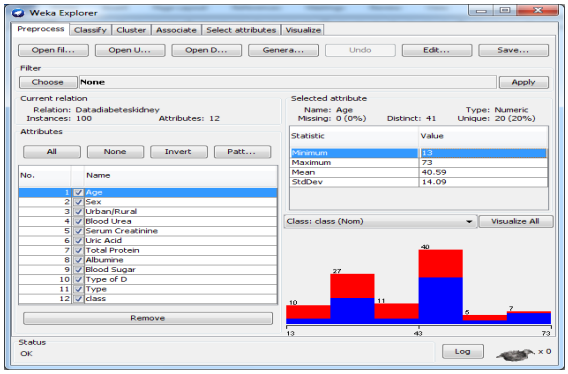


Figure 2: Opening Page

- After that, we click on classify tab. Then we choose J48 Algorithm from the left side under trees option. Then, we click on the textbox present on the right of "choose" button. We work with default values of this algorithm. The screen appears as in figure 3.

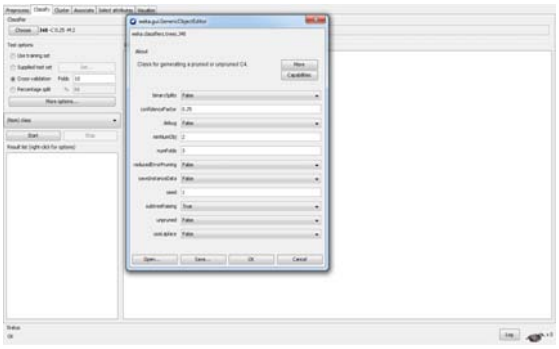


Figure 3: Selection of Algotithm & choosing Parameters

- Then using cross validation with 10 folds, classification is performed by clicking on start button. It would divide dataset into ten parts. With ten folds, it would apply training on first 9 parts and testing on last part. The result window is shown in Fig. 4 & 5. We can right click in the result window to visualize tree separately as shown in Fig. 13.
- In Fig 4, classifier output shows the decision tree generated by Weka. According to the tree, it takes "Serum Creatinine" as the root node i.e. Out of all the attributes, "Serum Creatinine" is the most important parameter that reflects the greatest effect of diabetes on kidney. Class 'A' (Affected) and class 'N' (Not-Affected) is taken as decision attributes.
- The result window illustrates the classifier performance in Weka. The accuracy is coming out to be 75% and computed error rate is 25%. It means we need to work more to get more accurate model. Mean absolute error is 29%. The confusion matrix is also shown in Fig. 5.

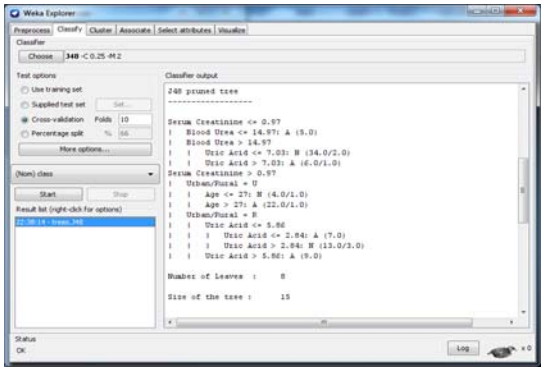


Figure 4: Generated Decision Tree

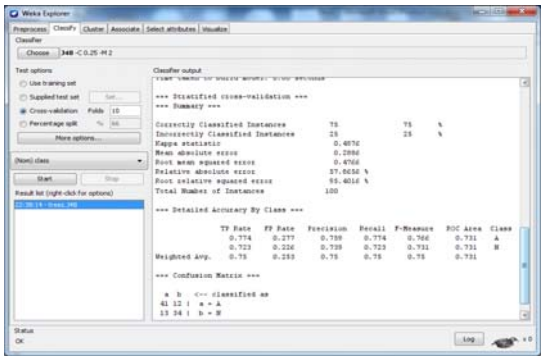


Figure 5: Interpreting Classifier Performance

- In Table 1, we can interpret the performance of Weka using different test options. The performance is interpreted in terms of the accuracy and error rate. It is found that the performance of Weka is best when tested with "Use Training Set" followed by "Cross validation" with 10 folds than with "Percentage Split" option.

Table 1: Performance of Weka Under Different Test Options

Test Options	*Accuracy*	*Error Rate*	*Kappa Statistic*	*Mean Absolute Error*
Use Training Set	92 %	8%	0.840	0.134
Cross Validation (10 folds)	75%	25%	0.497	0.288
Percentage Split (66%)	58.8%	41.1%	0.167	0.423

3.3 Classification in Tanagra

- Open Tanagra and then load the dataset in txt format. The dataset appears in Tanagra as shown in screenshot in Fig. 6. Tanagra detects the variable types automatically. It can be seen that there are 100 examples (records) and 12 attributes out of which there are 4 discrete attributes and 8 continuous attributes.

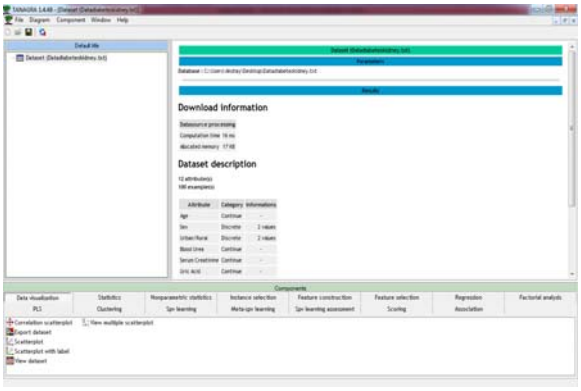


Figure 6: Opening Page

- Then, from the "View Dataset" component present inside the Data Visualization Tab, a pop-up menu appears. On choosing view menu, the data set would be displayed from Tanagra.

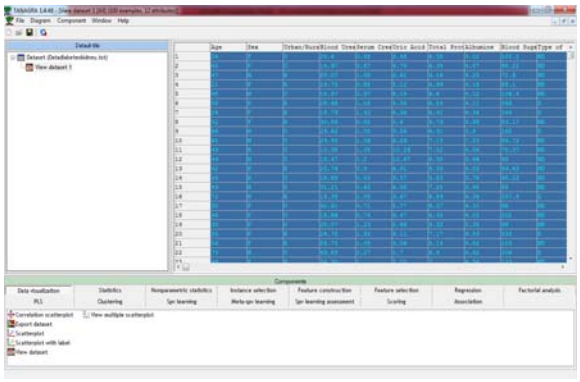


Figure 7. Viewing Dataset

- Then from the Feature Selection Tab, select "Define Status" component. Then we do the selection of parameters. We select all attributes as input except the last attribute - class. As we are interested in knowing the class (Affected or Non-affected), we set class as target.

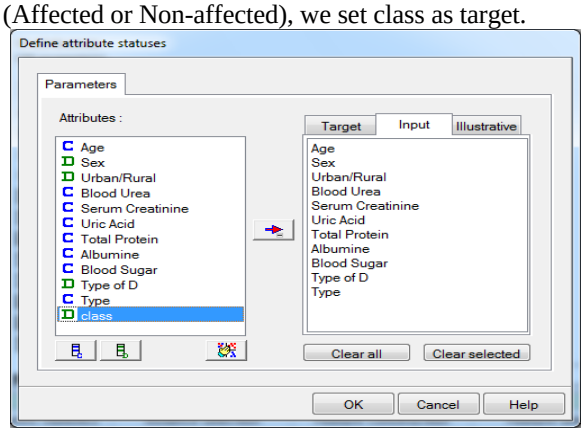


Figure 8. Selection of Input parameters

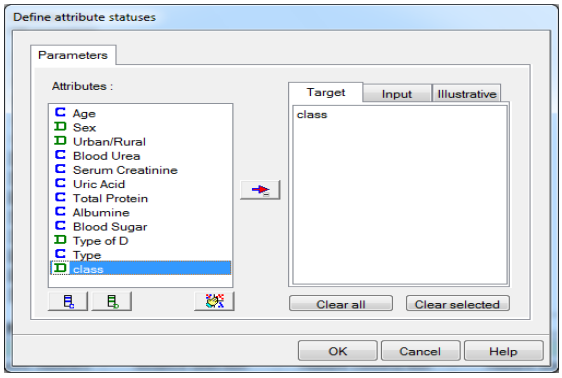


Figure 9. Selection of Output parameters

- Now we provide supervised learning using C4.5 Algorithm. For this, we add the "Supervised Learning" component present inside the Meta-Spv Learning in which we insert the "C4.5" learning algorithm (from Spv-Learning palette). On executing it, the result would be displayed. The result is shown in Fig. 10 and Fig. 11.
- Fig. 10. shows the generated decision tree in Tanagra. The root node is taken as "Total Protein" attribute and class 'A' and 'N' as the decision nodes. The tree shows that "Total Protein" is the most important attribute in the dataset that reflects greatest effect of diabetes on kidney.

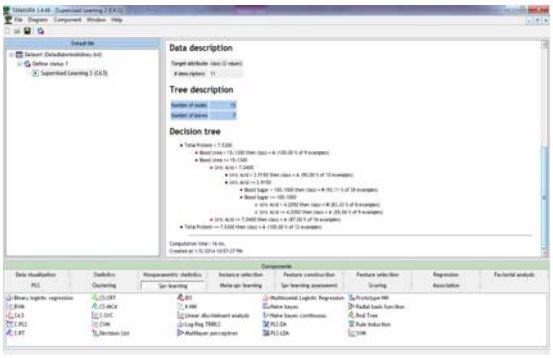


Figure 10: Generated Decision Tree

- Fig. 11. illustrates the classifier performance in Tanagra. It shows the confusion matrix and concludes that the resubstitution error rate is very less. This value is quite good for decision tree model.

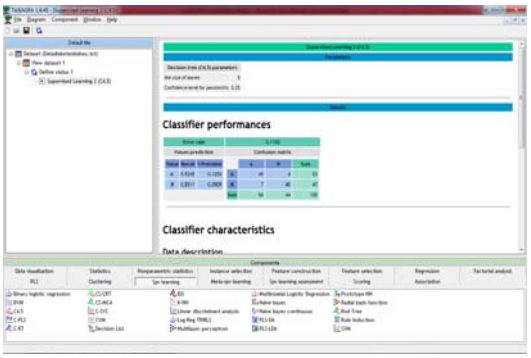


Figure 11: Interpreting Classifier Performance

- After the learning method we add a "Cross- Validation" component (from Spv Learning Assessment). We work with 10 folds and set number of repetitions to 1. We do not change the default parameters as shown in Fig. 12. The computed error rate is coming out to be 28%.

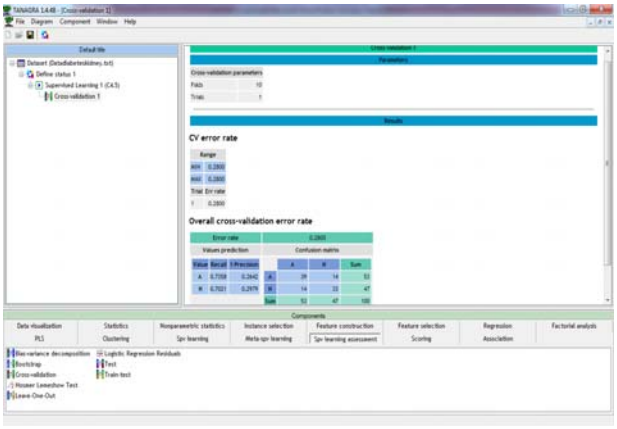


Figure 12: Supervised Learning Assessment

3.5 Comparison of classification in Tanagra & Weka

In this paper, a comparative study is made between Weka and Tanagra based on decision trees. The decision trees are generated using the application of C4.5 Algorithm that is used to generate rules signifying the effect of diabetes on kidney. The performance of classifier in both the tools is compared in terms of its accuracy, computation time and error rate.

- **Weka**

In Weka, the implementation of J48 Algorithm generates decision trees using 10-fold cross validation. Cross-validation is an efficient method for the estimation of error rate.

In Fig. 13, the decision tree has root node as "Serum Creatinine". According to the tree, Serum Creatinine determines the first decision. The numbers in parenthesis signifies the number of examples in the leaf node. The numbers after slash gives the number of misclassified examples. The decision tree includes 8 leaves and time taken to build tree model is 0.05 seconds. The error rate is 25%.

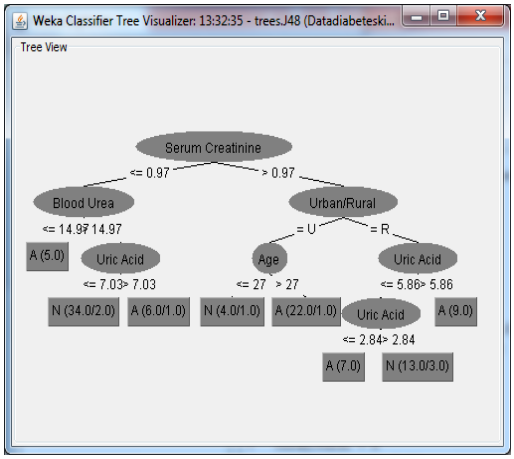


Figure 13: Decision Tree in Weka

- **Tanagra**

In Tanagra, the decision tree is generated by providing Supervised Learning using J48 Algorithm. According to the tree, "Total Protein" is taken as the root node i.e. this

attribute determines the first decision to find the diabetic effect on kidney. The tree model has 13 nodes and 7 leaves. The computation time is 0 ms. The error rate of the classifier is 11% which is lesser than Weka. So, Tanagra is more error-free than Weka.

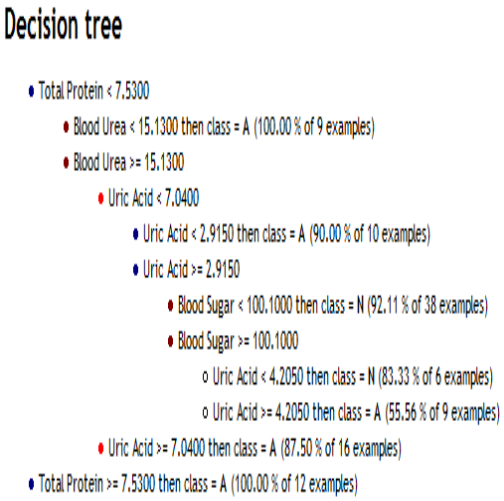


Figure 14: Decision Tree in Tanagra

4. Results and Conclusion

This research has conducted a comparative study on a dataset between two data mining toolkits (Weka and Tanagra) for classification purposes. After analyzing the results of both the tools, we found that both are able to generate tree model in very less time. Both the tools are very efficient in generating decision trees. However, in terms of classifiers` applicability, we conclude that the Weka tool is better in terms of the ability to run the classifier. However, the performance of classifier is better in Tanagra than Weka in terms of error rate. Also, Tanagra is faster than Weka in tree generation as its internal structure is organized in columns in memory. In addition, Weka tool has attained the highest performance in terms of accuracy when used with "Use Training Set" test mode than "Cross Validation" test mode followed by "Percentage Split" test mode. Through this comparative study, we conclude that Tanagra is better tool than Weka. Also, we found that c4.5 algorithm works well in decision tree induction. In future, we can implement this algorithm with more data and larger set of patient records to produce better results..

References

[1] A. Bonnaccorsi, "On the Relationship between Firm Arun K. Pujari, Data Mining Techniques
[2] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[3] http://en.wikipedia.org/wiki/Diabetes_mellitus
[4] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers,second Edition, (2006).
[5] http://www.healthline.com/health/kidney-function-tests
[6] http://www.britannica.com/EBchecked/topic/317431/kidney-function-test
[7] http://www.diabetes.ca/diabetes-and-you/living/complications/kidney/
[8] http://en.wikipedia.org/wiki/C4.5_algorithm
[9] Veronica S. Moertini,"Towards the use Of C4.5 Algorithm for Classifying Banking Dataset", INTEGRAL,Vol 8. No. 2, October 2003..
[10] Jay Gholap,"Performance Tuning of J48 Algorithm for Prediction of Soil Fertility", Innovative Journal of Medical and Health Sciences, Vol 2, No 8 (2012) .
[11] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 April 2011).
[12] http://wwww.samdrazin.com/classes/een548/project2report.pdf
[13] I.H. Witten and E. Frank, Data Mining Practical Machine Learning Tools and Techniques, Second Edition, Elsevier Inc., 2005
[14] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[15] http://en.wikipedia.org/wiki/Tanagra_(machine_learning