International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-1, Issue-3, August 2012

Simulation of Various Classifications Results using WEKA

Shilpa Dhanjibhai Serasiya, Neeraj Chaudhary

Abstract: ***In this paper, we focused on the construction of class association rules and classification model. In knowledge discovery process association rule mining and classification are two important techniques of data mining and widely used in various fields. In order to mine only rules that can be used for prediction, we modified the well known association rule mining algorithm - Apriori to handle user-defined input constraints. The paper tries to explain the basics of class association rule mining and classification through WEKA. This article presents how problems of classification and prediction can be solved using class association rules. In the simulation on WEKA, we have used selected classification techniques to propose the appropriate result from our training dataset. Thus, by using the simulated results, we suggest the classification using association rules.***

Keywords: Association rule, class association rules, classification, Data mining.

I. INTRODUCTION

Generally, association rule mining [1] and classification rule mining [2] are the two most popular techniques in data mining. Both of association rules and classification rules are represented as *if-then* type rules. However, there are some differences between them. Association rules are generally used as descriptive tools, which give the association relationships to the specific application experts,

While classification rules are used for predicting the unseen testing data. Therefore, the evaluations of the two type of rules are different. Association rules are typically evaluated by the application experts, while classification rules are evaluated by the classification accuracy of testing data.

In order to discover the strongly correlated rules, many kinds of measures have been proposed to evaluate the interestingness of patterns, such as famous support and confidence [1]. However, there are so many measures proposed and different measures have different properties which usually lead to different and conflicting results. Some studies investigate that there is no optimal measure which is better than others in all applications [3], [4]. Therefore, given a specific application, finding the appropriate measure becomes the essential problem in data mining. Data mining algorithms have well taken up challenges for data analysis in large database. Association rule mining is one of the key data-mining tasks in which associability of the items is discovered in training database [1]-[3]. Classification is another data mining tasks. The objective of classification is to build a model in training dataset to predict the class of future objects whose class label is not known[4],[5].

The idea of use association rule mining in Classification was first introduced in 1997 by[4],[6] and it was named as class association rule mining or associative classification. The first classifier based on association rules was CBA [7] given by Liu et al. in 1998. Later, some improved classifiers were given. More research is going on to design even improved classifiers.

Class association rule mining process can be decomposed in *three* parts.

1. First find frequent item sets and frequent class association rules.
2. Second we find the strong class association rules by pruning the weak rules.
3. Design a classifier [2].

Various methods [4],[6],[7] are common to accomplish the class association rule. In our work, we have simulated class association rules and classifier in the WEKA framework. WEKA is a data mining system developed at the University of Waikato and has become very popular among the academic community working on data mining. We have chosen to develop this system in WEKA as we realize the usefulness of having such a classifier in the WEKA environment.

II. ASSOCIATION RULE MIMING

Association rule mining is a widely-used approach in data mining or knowledge discovery. Association rules are capable of revealing all interesting relationships in a potentially large database. The abundance of information captured in the set of association rules can be used not only for describing the relationships in the database, but also for discriminating between different kinds or classes of database instances [8]. An association rule is defined as

"Let $I = \{i_1, i_2, \ldots . i_m\}$ be a set of literals, called items. Let D be a set of transactions(database), where each transaction T is a set of items such that $T \subset I$. TID indicates a unique transaction identifier. An association rule is an implication of the form $X \rightarrow Y$, $X \subset I$ and $Y \subset I$ and $X \cap Y = \Phi$. X is called antecedent while Y is called the consequence of the rule."

There two measurements in association rule mining are *support* and *confidence*. The support corresponds to the frequency of the pattern while confidence indicates rule's strength.

Revised Manuscript Received on 30 August 2012.
* Correspondence Author
Ms. Shilpa Dhanjibhai Serasiya*, M.Tech Student, Department of Computer Science Engg., Rajasthan Collage of Engineering for Women, RTU, Jaipur, India.
Prof. Neeraj Choudhary, Reader, Department of CSE, Rajasthan Collage of Engineering for Women, RTU, Jaipur, India.

$$\text{Support } A \rightarrow B = \frac{\text{tuples containing both A and B}}{\text{total tuples}}$$

$$\text{Confidence } A \rightarrow B = \frac{\text{tuples containing both A and B}}{\text{tuples containing A}}$$

Rules that satisfy both the minimum support threshold and minimum confidence threshold are said to be **strong**. Suppose dataset contain 4 transaction had shown in Table I.

Table I. A dataset

TransactionID	ItemsBought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support is 50% and the minimum confidence is 50 % then the then we have rules,

$A \Rightarrow C$ (50%, 66.6%)
$C \Rightarrow A$ (50%, 100%)

III. CLASSIFICATION

Classification is another data mining tasks, can be defined as learning a function that maps (classifies) a data instance into one of several predefined class labels. When classification models are constructed from rules, often they are represented as a decision list. Classification rules are of the form P → C, where P is a pattern in the training data and C is a predefined class label (target).

The objective of classification is to build a model in training dataset to predict the class of future objects whose class label is not known[9]. There are two major issues in classification,

1. Preparing the data for classification and prediction
2. Comparing classification and prediction methods

There are commonly used classification techniques which extract relevant relationship in the data are Artificial neural networks, Decision trees, Bayesian Method[10] etc.

Association and classification rules are represented as *if-then* type rules. However, there are some differences between them. Association rules are generally used as descriptive tools, which give the association relationships to the specific application experts, while classification rules are used for predicting the unseen testing data. However, a major problem in association rule mining is its complexity. The result of an arbitrary association rule mining algorithm is not the set of all possible relationships, but the set of all interesting ones. That is an important issue of the mining process, but the quality of the resulting rule set is ignored. On the other hand there are approaches to investigate the discriminating power of association rules and use them according to this to solve a classification problem [11][12].

IV. CLASS ASSOCIASTION RULE MINING

Class association rule mining is a special case of association rule mining and associative classification finds a subset of class association rule set to predict the class of previously unseen data (test data) as accurate as possible with minimum efforts. This subset of class association rule set is called associative classifier or simply a classifier. The problem states of finding a subset of an association rule set of the X → C, where X is association of some or all object features and C is class label of that object.

The algorithmic approach for classification using association rules can be divided into association rule mining and pruning and from that, getting classification rules. Figure 4.1 provides a graphical overview of the entire process.

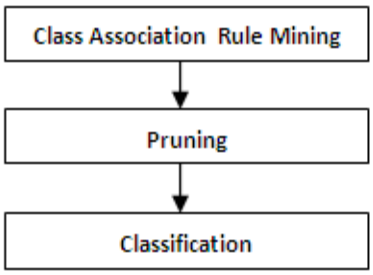

Figure 4.1

Let we illustrate the class association rule mining with the training data shown in Table II. It consists three attributes X (X1, X2, X3), Y (Y1, Y2, Y3), Z (Z1, Z2, Z3) and two class labels (C1, C2). We assume the min_sup = 30% and min_conf = 70%. Table III shows the strong class association rules along with their support and confidence. It also represents a classifier as the rules are sorted according to confidence they hold.

Table II

Training database				
TID	X	Y	Z	Class
1.	X2	Y2	Z1	C1
2.	X1	Y2	Z2	C2
3.	X1	Y3	Z3	C2
4.	X3	Y1	Z2	C1
5.	X1	Y1	Z3	C2
6.	X2	Y3	Z1	C1
7.	X3	Y3	Z2	C1
8.	X1	Y1	Z1	C1
9.	X2	Y3	Z1	C1
10.	X1	Y1	Z1	C2

Table III

Strong class association rule set			
Class association rule		Support	Confidence
Antecedent	Consequent		
X2	C1	3/10	3/3
Y3	C1	3/10	3/3
X2Z1	C1	3/10	3/3
X1	C2	4/10	4/5
Z1	C1	4/10	4/5

As explained above, the mining of association rules is a typical data mining task that works in an unsupervised manner. A major advantage of association rules is that they are theoretically

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

capable of revealing all interesting relationships in a database. But for practical applications the number of mined rules is usually too large to be exploited entirely. This is why the pruning phase is stringent in order to build accurate and compact classifiers. The smaller the number of rules a classifier needs to approximate the target concept satisfactorily, the more human-interpretable is the result. In addition the classification time for unseen instances is less when using a smaller rule set. In order to compare rule mining strategies, the pruning steps can be simplified. Therefore differences in the mining methods affect the classifier directly.

V. SIMULATION

WEKA is a data mining system developed at the University of Waikato and has become very popular among the academic community working on data mining. The researcher has chosen to develop this system in WEKA as it realized the usefulness of having such a classifier in the WEKA environment. Weka is an open source machine learning environment with many useful data mining and machine learning algorithms.

Many other classification systems have been built based on association rules. In the research paper, there is an implementation of an association rule-based classifier system in the WEKA framework. The researcher has selected the dataset given in the Table IV which depicts the information about different possibilities of the play to occur on the basis of weather. Thus in the Table IV outlook, temperature, humidity and windy are *antecedent* and play is *consequence.*

Table IV

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

One such work includes merging the Apriori implementation in Weka for rule generation and J4.8 – Weka's implementation of the C4.5 decision tree - algorithm for classification. As the researcher wants to classify the data using association mining, in which the class attribute (consequence) should remain common for the data mined rules. But by applying the simple Apriori algorithm for association rule mining, one gets different values of consequence in association rules given in the Figure 5.1

As researcher want to study class association rule mining and the same can be done by Apriopri algorithm with an option of car in WEKA. If car is selected or enabled then classification rules are mined instead of general association rules. As required class attribute in the class association rule mining, one can define it by the field class index in the given apriopri algorithm. Class index is the index of the class attribute. If it is set to (-1) the last attribute is taken as class attribute. Minimum support can be set in lower bond minimum support field. The above procedure of fields under the Apriopri algorithm is shown in the Figure 5.2 By using the above fields under Apirori algorithm, the researcher gets the same value of consequence of association rules forming class association rules (Figure 5.3) as mention in the class Index field of Figure 5.2.

Figure 5.4 shows different testing option for classification available in WEKA framework. In that first user training set option takes whole training data set as a testing. Second option sets the testing dataset which have same attribute and class value. Then next option divides dataset in number of given folds and test on each by generating rules from all other folds. And last option divides the dataset in some percentage and takes one for training and other for testing. Table V shows correctly and incorrectly classify instance with different testing options.

Table V Testing options

Testing option	Correct Classify Instance %	Incorrect Classify Instance %
Training set	100	0.0
Supplied test set	55.5	44.5
Cross validation folds=10	57.14	42.85
Percentage split (40%)	62.5	37.5

From the above table V the Evaluation graphs for correctly and incorrectly classify instance is shown in Figure 5.5

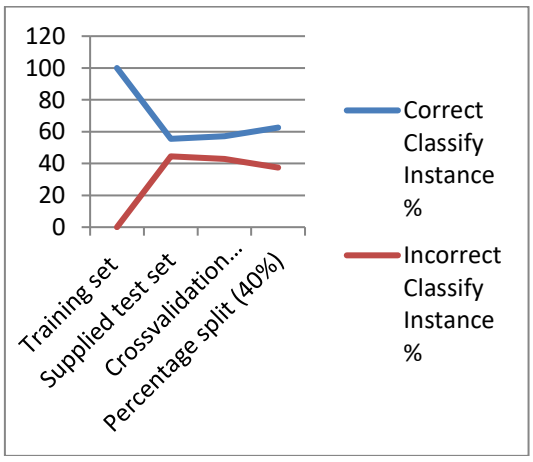

Figure 5.5 Evaluation Graph

Here researcher uses J48 - Weka's implementation of the C45 decision tree algorithm - for classification.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The classification result of J48 algorithm with 10 folds cross validation testing option is shown in the Figure 5.6. As the classification result gives more pruned rules then class association rules.

Thus we can get classifier, using class association rules by applying some pruning algorithm (e.g. Finding subset of the rules with greater confidence value) on the result of class association rule mining.

VI.CONCLUTION

Use of association rule for classification is novel approach in data mining. Classification rules are just subset of association rules. In this paper first association rules and then class association rules are discovered they are pruned to get qualitative and sufficient classification rules. The approach described in this paper can be very helpful and efficient if there is an application where both kind of knowledge is required (association among attributes and classification of objects). Further, this approach can be compared with other classification approaches like decision tree, neural network, rule based classifiers etc. in terms of accuracy and efficiency.

REFERANCE

1. R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", Proceedings of the 20th International Conference on Very Large Data Bases (VLDB-94), Morgan Kaufman Publishers, Santiago de Chile, Chile, September 1994, pp. 487-499.
2. P. R. Pal, R. C. Jain, CAARMSAD: "Combinatorial Approach of Association Rule Mining for Sparsely Associated Databases". Journal of Computer Science, Tamilnadu India, Vol. 2, No 5, pp 717, July 2008.
3. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", proceedings of international conference on management of data (ACMSIGMOD'00), pp 1-12, Dallas, TX, May 2000.
4. R. Bayardo, "Brute-force mining of high-confidence classification rules",Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press, Newport Beach, CA, United States, August 1997, pp. 123-126.
5. J. Quinlan, C4.5, "Programs for machine learning", San Mateo, CA: Morgan Kaufmann, 1993.
6. K. Ali, S. Manganaris, and R. Srikant, "Partial Classification using Association Rules", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press, Newport Beach, CA, United States, August 1997, pp.115-118.
7. B. Liu, W. Hsu, and Y. Ma, "Integrating Classification and Association Rule Mining", Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press, New York City, NY, United States, 1998, pp. 80-86.
8. Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. SIGMOD Record (ACM Special Interest Group on Management of Data), 29(2):1{?, 2000.
9. Tom Johnsten and Vijay V. Raghavan, "Impact of Decision- Region Based Classification Mining Algorithms On Database Security", supported in part by a grant from the U.S. Department of Energy.
10. Hui Yin1,2, Fengjuan Cheng2, Chunjie Zhou1, "An Efficient SFL-Based Classification Rule Mining Algorithm", Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education, pp. 969 – 972.
11. Xianneng Li, Shingo Mabu, Huiyu Zhou, Kaoru Shimada and Kotaro Hirasawa, "Analysis of Various Interestingness Measures in Classification Rule Mining for Traffic Prediction", SICE Annual Conference 2010 ,August 18-21, 2010, pp. 1969 – 1974.
12. Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classi_cation and association rule mining. In Knowledge Discovery and Data Mining Integrating, pages 80-86, 1998.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

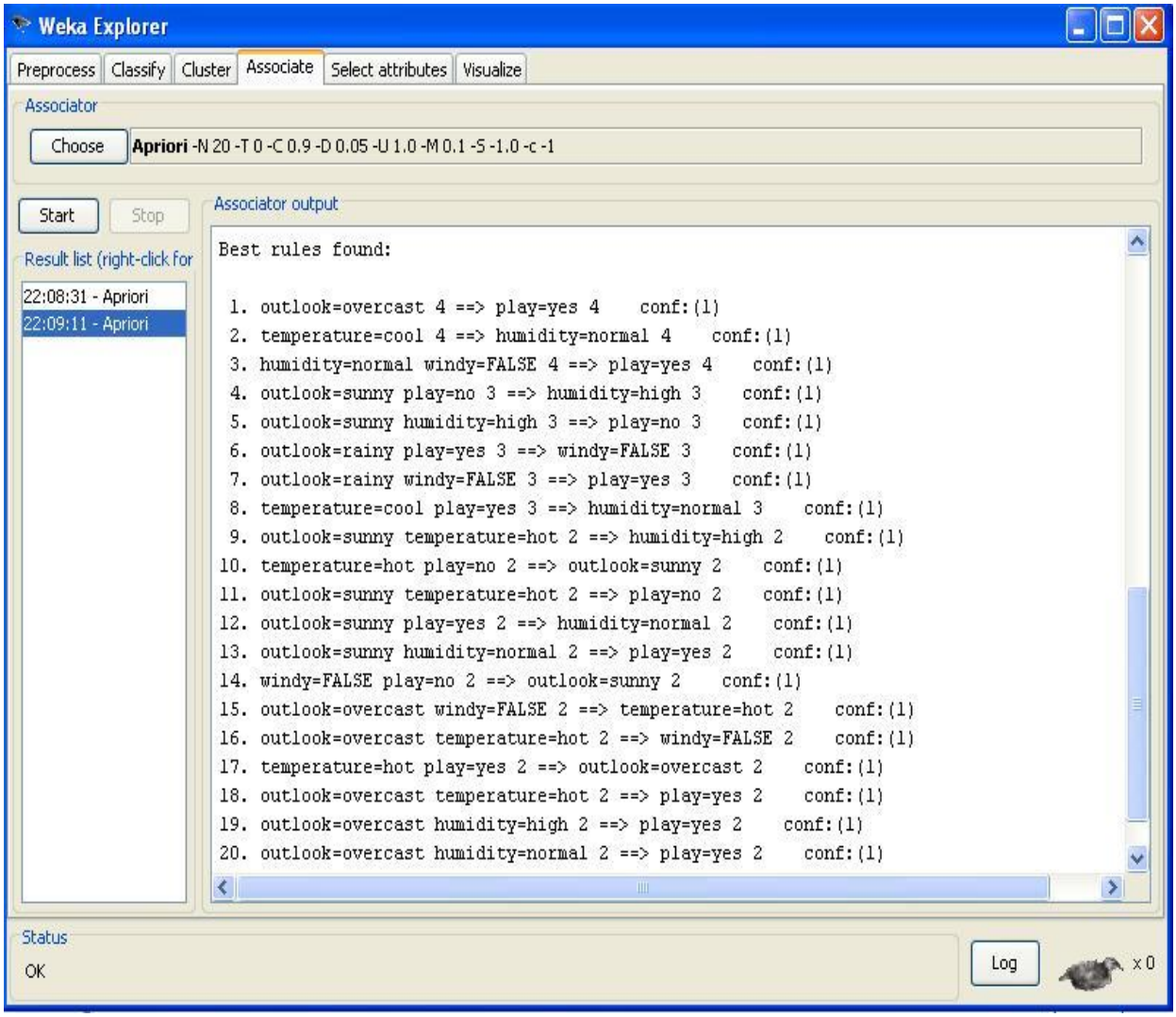

Figure 5.1

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

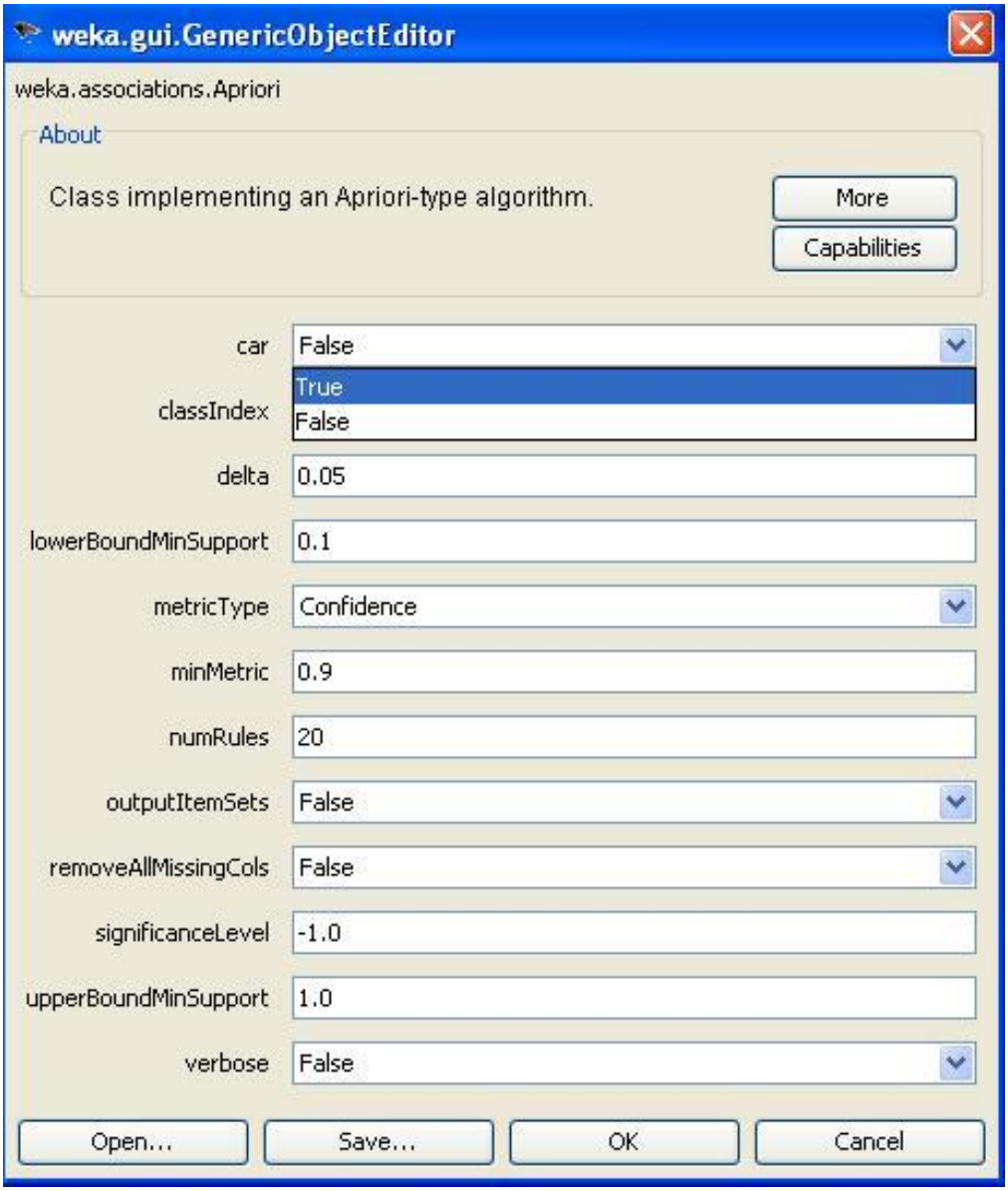

Figure 5.2

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

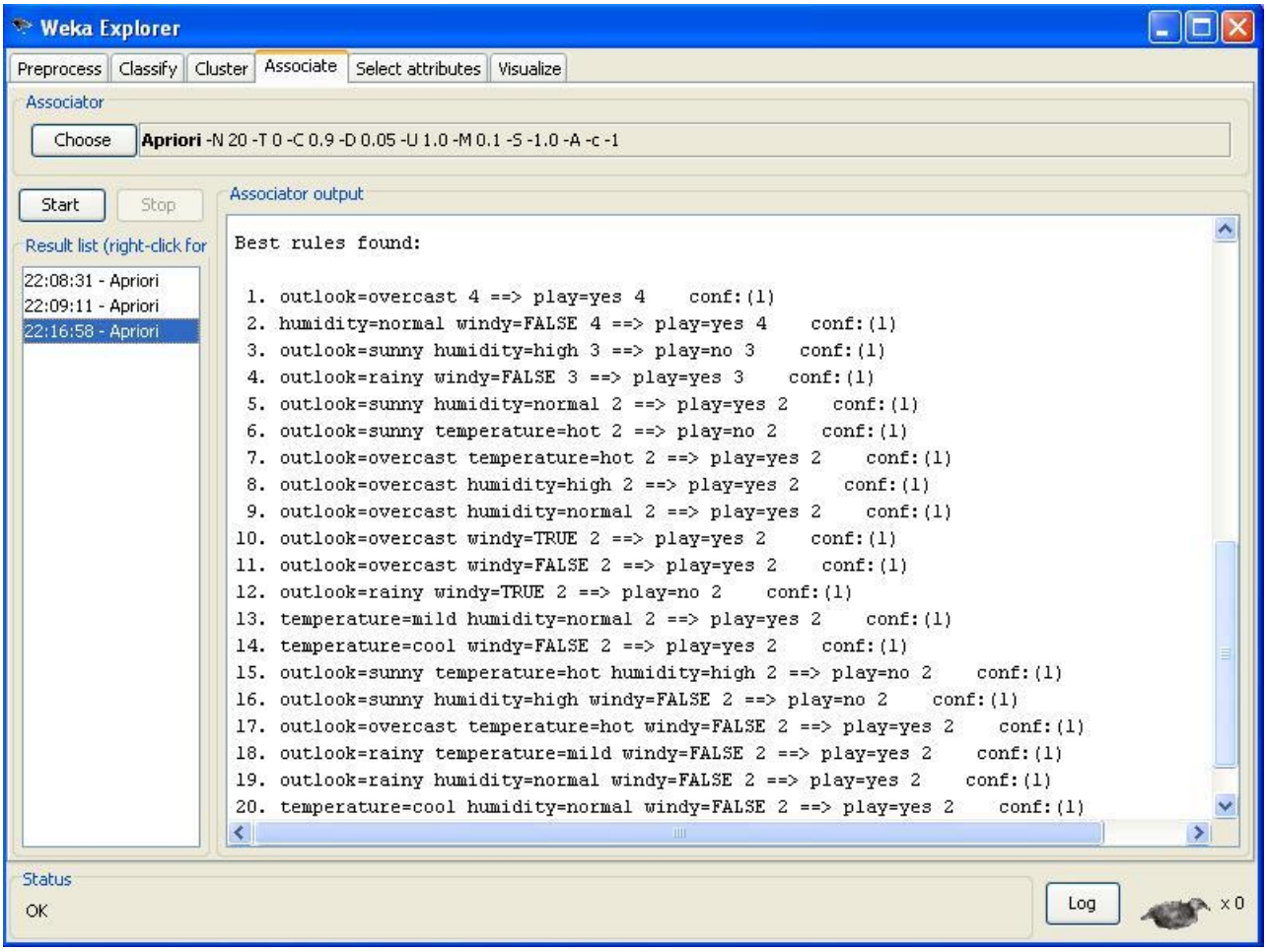

Figure 5.3

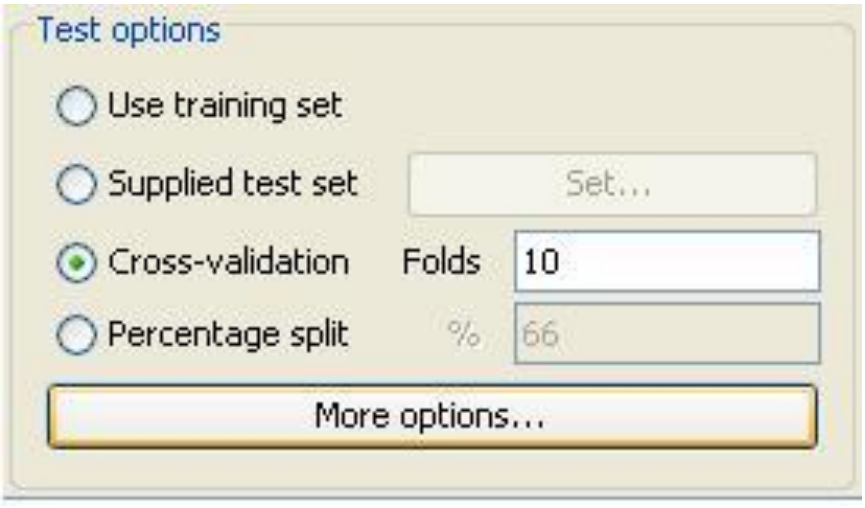

Figure 5.4

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

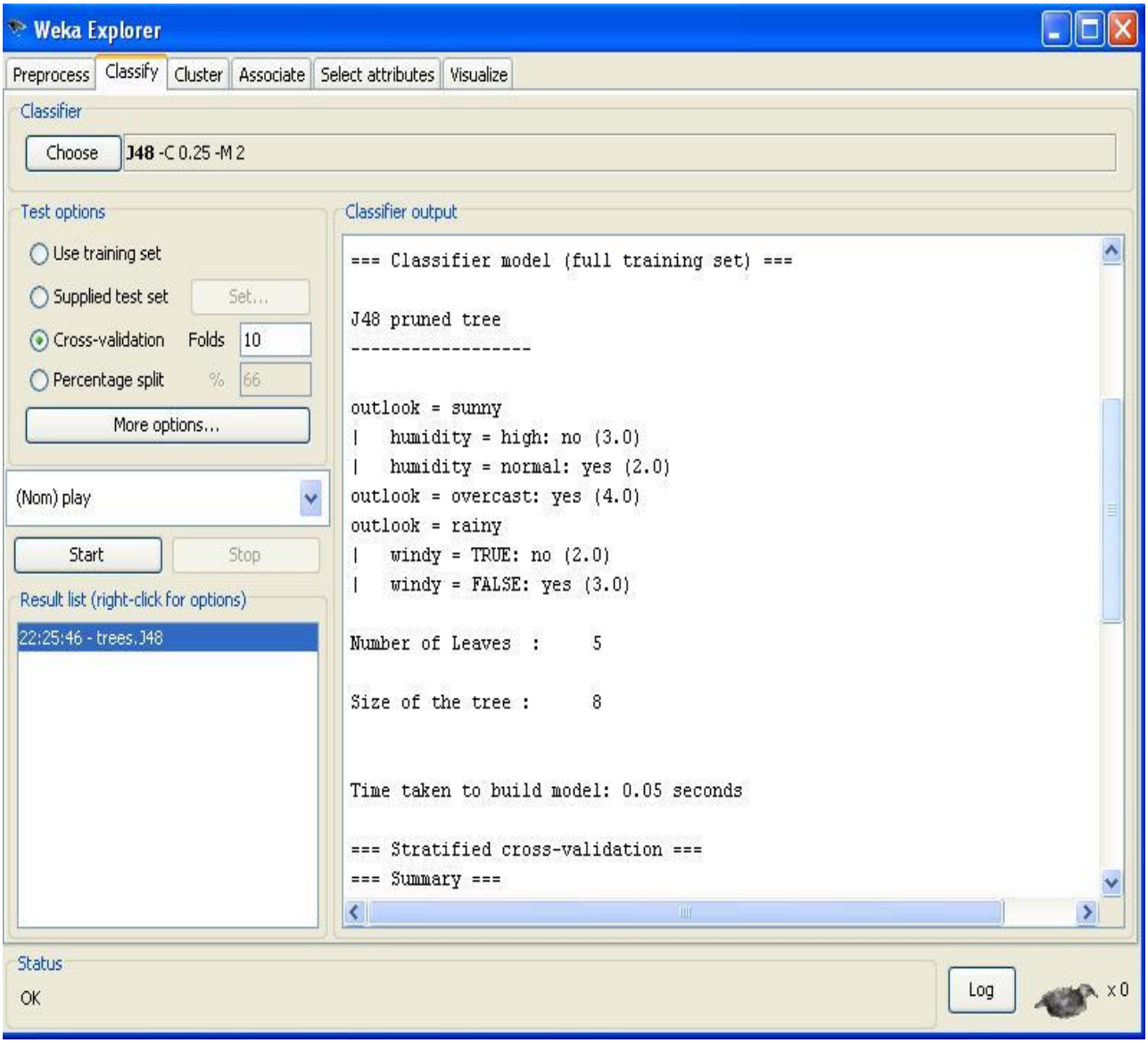

Figure 5.6

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)