International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 4 Issue 11, November-2015

Comparative study of Different classification algorithms using varying parameters in WEKA Tool

Yogesh Saharan[1], Yogesh[2]

[1,2]Dept. of Computer Science, UIET, MDU, Rohtak

ABSTRACT

Classification is a classic data mining technique based on machine learning. Data mining refers to the finding of relevant and useful information from data bases .data mining and knowledge discovery in the database is a new interdisciplinary field, merging ideas from statistics, machine learning, database and parallel computing. The work is performed on two data mining algorithms: C4.5 and Multilayer Perceptron. The algorithms have been applied to the medical datasets. The experiments are conducted in WEKA (Waikato Environment for Knowledge Analysis).The analyses are performed as follows: Each of the algorithms is calibrates with the use of several parameters available in the WEKA and For each of the parameter settings of individual algorithm analyses are performed on each of the datasets. Then the outcomes are measured with the use of the metrics .The results are presented in a form of tables and graphs. This thesis shows that the ultimate step encompasses the comparison of the performance of each of the data mining models.

Keywords: Data Mining, multilayer perceptron, WEKA, Machine Learning, C4.5.

1. INTRODUCTION

Data mining is the search for the relationships and global patterns that exist in the large databases but are hidden among vast amount of data, such as the relationship between patient data and their medical diagnosis. This relationship represents valuable knowledge about the database and the objects in database, if the database is a faithful mirror of the real world registered by the database.

Although the data mining is a quite young discipline it is popular due to successful applications in telecommunication, marketing and tourism. In recent years the usefulness of the methods has been proven also in medicine. Data mining aims at describing specific patterns (dependencies, interrelations, various regularities) which may be present in data. These patterns, discovered in historical data, may be used to support future decisions concerning diagnosing of new cases. Such knowledge may also have an enormous value for decision making in treatment planning, risk analysis and other predictions. Prior to the mining process it is essential to gain sufficient amount of data. This may require integrating data from multiple heterogeneous information sources and transforming it into a form specific to a target decision support application. Afterwards the data has to be prepared for knowledge extraction (e.g. by selecting proper records and attributes).

HOW DATA MINING HELPS IN MEDICAL FIELD?

The practise of using concrete data and evidence to support medical decisions (also known as evidence-based medicine or EBM) has existed for centuries. John Snow[8] considered being the father of modern epidemiology, used maps with early forms of bar graphs in 1854 to discover the source of cholera and prove that it was transmitted through the water supply. Snow counted the number of deaths and plotted the victim's addresses on the map as black bars. He discovered that most of the deaths clustered towards a specific water pump in London.

Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector. Data mining and its application to medicine and public health is a relatively young field of study. In 2003, Wilson [8] began to scan cases where KDD and data mining techniques were applied in health data. The

preceding factors remind us of an incident in the Philippines at the Rizal Medical Centre in Pasig City in October 2006. Failing to implement strict sanitation and sterilization measures the hospital contributed to the death of several new-born babies due to neonatal sepsis (bacterial infection). No one really knew what was going on until the deaths became more frequent. Upon examining hospital records, the Department of Health (DOH) found that 12 out of 28 babies born on October4, for example, died of sepsis. With an integrated database and the application of data mining databases Cheng, cited the use of classification algorithms to help in the early detection of heart disease, a major public health concern all over the world. Another study used the K-means clustering algorithm to analyse cervical cancer patients and found that clustering found better predictive results than existing medical opinion.

2. WEKA

Waikato Environment for Knowledge Analysis (WEKA) Waikato Environment for Knowledge Analysis, called shortly WEKA, is a set of state-of-the-art data mining algorithms and tools to in-depth analyses. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License. These complex algorithms may be applied to data set in the aim of detailed analyses and evaluation of data mining examination. There are three main ways of WEKA use. First is analysing data mining methods' outputs to learn more about the data; next is generation of model for prediction of new instances and finally the last but most important for this master's thesis feature, comparison of data mining methods in order to choose the best one as a predictor e.g. in Medical Decision Support System.

WEKA consists of four user interfaces out of which three are graphical and one command-line. The main interface is called Explorer. It is graphical interface built of menu section and six panels connected to various data mining methods. It enables data pre-processing, classification, cauterization, and mining associations among attributes. Furthermore there is a possibility to select attributes with the attribute evaluator and search method. The last option is visualization plotting the dependencies among attributes.

The next graphical interface, Knowledge Flow is dedicated to selecting components from the tool bar and placing them on the special canvas, connecting them into directed graph than processing and analysing. Furthermore the data stream data processing can be designed and executed with the usage of this interface. To compare performance of data mining algorithms it is useful to choose third graphical interface called Experimenter. This module allows one to evaluate how well various data mining methods perform for given datasets. This process is automated and statistics can be saved. This module is a most important part of the experiment. It makes in-depth statistics which are useful in case of medical datasets. After the selection of various methods, their parameters and datasets, it is possible to prepare statistic which are priceless in case of medical diagnosis support.

Experimenter and Explorer are two mainly used interfaces during master's thesis experiments. WEKA allows analysing the data sets saved in the **.arff** files what can be easily achieved by converting **.txt** files in the way presented in Figure 1.1 The file with data has a structure of decision table, it begins with the name of the table, than names and types of attributes are declared, finally observed attributes' values are typed. This uncomplicated document structure allows one to upload to the environment prepared in this way own dataset and analyse it.

```
@relation diabetes

@attribute pregnant real
@attribute plasma real
@attribute diastolic real
@attribute triceps real
@attribute insuline real
@attribute mass real
@attribute pedigree real
@attribute age real
@attribute diabetes {1,0}

@data
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
```

Figure 1.1 Sample .arff file for WEKA

The comprehensive and deep analysis is possible with the use of WEKA environment. It is the purpose of selection it to analyse medical data sets. The availability of WEKA and its documentation allows one to conduct similar studies to presented in the master's thesis and compare her results with presented in this document. The state-of-art techniques implementations make analyses accurate and precise.

3. C4.5 CLASSIFIER

The C4.5 algorithm is an extension of the well-known ID3 algorithm. The extension includes avoiding data over fitting by determining how deeply a tree can grow. The C4.5 algorithm is capable of handling continuous attributes, which are essential in case of medical data (e.g. blood pressure, temperature, etc.). Other very common aspect – missing values – was also taken into consideration in C4.5. Moreover the algorithm handles attributes with differing costs.

DEFINITION

"Given a hypothesis space H, a hypothesis h E H is said to over fit the training data if there exists some alternative hypothesis h' E H, such that h has smaller error than h' over the training examples, but h' has a smaller error than hover the entire distribution of instances." The inductive C4.5 algorithm generates rules from a single tree. It can transform multiple decision trees and create a set of classification rules. Such feature of this algorithm may be used to scale up the rule generalization, time of learning and size and number of rules. The usefulness of C4.5 algorithm was widely proven in medicine. This algorithm suits medical data because it copes with missing values. What is more the algorithm handles continuous data which are common among medical symptoms. The efficiency of C4.5 was shown e.g. in breast cancer and prostate cancer classification to generate a decision tree and rules which may be helpful in medical diagnosing process.

3.1 MULTILAYER PERCEPTRON:

In WEKA environment Multilayer Perceptron is equipped with additional graphical interface which allows modification of the network. It is possible to add nodes and connections among them. The user may decide how long the net should be trained with the parameter number of epoch and continue the training by denying the obtained results or break the training by accepting the presented error per epoch. WEKA's Multilayer Perceptron algorithm has several parameters which can influence its performance. The hidden Layers parameter sets if the hidden layers are present in the network and, if yes, how many hidden nodes each layer contains. The hidden Layers value is a list of integers separated with commas. The hidden Layer equals 0 means that the hidden layer is absent. Moreover, there are four predefined values of this parameter: o - number of hidden layers equals number of class values, i – number of hidden layers equals number of attributes, a – number of hidden layers equals average of i and o, and t number of hidden layers equals sum of i and o. The next option auto Bild lets to add hidden layers and their connections. With this option off there would be no hidden layer in the net. The other parameters are learning Rate and Momentum available also in graphical presentation of neural networks. The decay is a ratio of starting value to epoch number. In the text interface reset option let one to begin re-training of the network with lower learning rate. The other parameter training Time corresponds to a number of epochs. The alternative to this parameter is validation Set Size which allows stopping the training phase. The phase is stopped while performance of this set starts to worsen. The Validation Threshold sets the number of times when validation set error performance is worsen before the end of the training phase. There are also filters to improve neural network's performance like normalize Numeric Class and normalize Attributes useful in case of numeric values.

4. PROPOSED WORK

In this paper, we attempt to analyses of calibration of individual algorithms. The purpose of these analyses is to determine what parameters settings yield the best models. The calibration aims at finding optimal settings which maximize the performance of each of the algorithms. The tests are done with the use of various dataset splits and n-fold cross validations (for several n's) – testing configuration. For comparative purposes also the entire training set is used for testing. These results are skipped while commenting the results of the analyses, though. All the work is done medical datasets which is downloaded from UCI repository. We have divided our work into four steps. These are as follows:

- In the first step each of the algorithms is calibrates with the use of several parameters available in the WEKA.
- In the second step for each of the parameter settings of individual algorithm analyses are performed on each of the datasets.
- In the third step the outcomes are measured with the use of the metrics and the results are presented in a form of tables and graphs.
- The ultimate step encompasses the comparison of the performance of each of the data mining models.

5. EXPERIMENTAL EVALUATION

We have performed our experiment on computer running Intel core i3 processor have windows 7, 256GB RAM, 1GB hard disk, and WEKA tool installed on the machine.

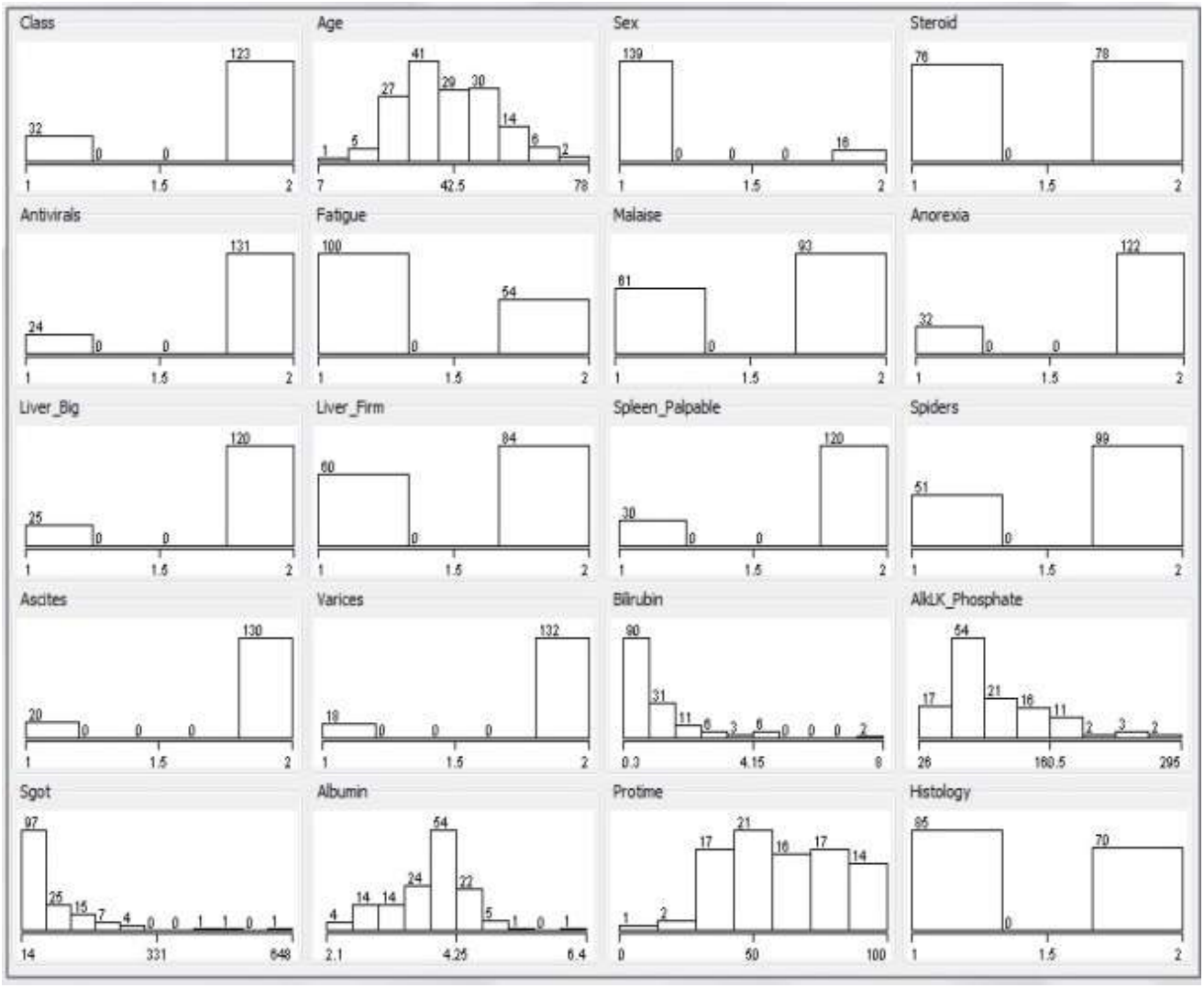

Figure 1.1: Distribution of the attributes of the Hepatitis data

Here in figure 1.1 hepatitis database consists of seventeen conditional attributes. The four of them are multi-valued, the rest is binary. The decisional attribute die takes values 0 or 1. As presented in the Figure 1.1 the distributions of values of the attributes. The distributions of the attributes: bilirubin, and s got decrease with the increase of the values. The distribution of values of the attributes: age and albumin is of bell-shape.

a) C4.5 algorithm for hepatitis database

After the generation of the decision tree its accuracy is verified similarly to the previous databases. The experiment gave good results than in the case of the breast cancer database. The average number of correctly classified instances does not exceed 73% (Table 4.1.7). The best in this regard is the 50% split. The errors were high for all of the configurations. The highest was the Root Relative Squared Error which reaches nearly 95%.

Table 1.1: Performance of the C4.5with the respect to a testing configuration for the Hepatitis database

Testing Method→	**Training set**	**10-fold cross validation**	**5- fold cross validation**	**15- fold cross validation**	**30 % split**	**50 % split**	**66 % split**
Correctly Classified instances	67.7%	76.7%	75.4%	73.5%	66.6%	80.5%	79.2%
Incorrectly Classified Instances	32.2%	23.2%	24.5%	26.4%	33.3%	19.4%	20.7%
Kappa statistic	45.2%	52.7%	49.6%	45.3%	32.6%	61%	58.2%
Mean Absolute error	32.4%	30%	29.5%	31.5%	33.4%	28.7%	27.2%

Root Mean Squared error	43.9%	44.9%	45.4%	46.74%	52.8%	41.32%	41.5%
Relative Absolute error	61.9%	61.5%	60.5%	64.7%	67.4%	58.07%	55%
Root Relative Squared error	91.8%	91.1%	92%	94.6%	96.3%	80.4%	82.2%
TP Rate	97.7%	76.8%	75.5%	73.5%	66.7%	80.5%	79.2%
FP Rate	30.4%	23.6%	25.8%	28.5%	34.4%	19.4%	21.1%
Precision	68.2%	77%	75.5%	73.4%	73.8%	80.6%	79.2%
Recall	67.7%	76.8%	75.5%	73.5%	66.7%	80.5%	79.2%
F-Measure	63.8%	76.9%	75.5%	73.5%	63.7%	80.5%	79.2%
AUC	68.9%	70.1%	73.3%	68.3%	70.1%	80.9%	83.7%

When it comes to the True Positive rate the best results were gained for the 66% split with decent False Positive rate. The rest of the configurations had the True Positive rates low. Good results were gained for the Precision, Recall, F-measure and AUC. The worst configurations in terms of the Precision are the 30%.

b) Multilayer Perceptron for hepatitis database

The Table 1.2 shows the results gained for the Multilayer Perceptron for the hepatitis database. The number of hidden layers equals the average of the number of class values and attributes. The results in terms of the percent of correct classifications are worse. None of the models exceeded the level of 75% of correct classifications (Table 1.2). Slightly good results were gained in terms of the errors. In this respect the best testing configuration the 66% split turned out to be. Poor results were also gained with regard to the True Positive rate (TP rate). The best results were obtained for the 30% split. However, when it comesto the False Positive rate (FP rate) the 66% split gave the best classifier. Finally, in terms of the Precision, Recall, F-measure and AUC the results varied considerably. When the Precision is considered the average value is about 63%, with the 66% split being the best, reaching 70%. The Recall and the F-measure metrics had the greatest values for the 10-fold cross-validation (about 98%, while the rest do not exceed 70%). Varying values were also gained for the AUC, where the best classifier was the one built with the 66% split testing configuration. The rest configurations delivered models whose AUC was less than or equal about 75%.

Table 1.2: Performance of the Multilayer perceptron with the respect to a testing configuration for the Hepatitis database

Testing Method→	**Training set**	**10-fold cross validation**	**5- fold cross validation**	**15- fold cross validation**	**30 % split**	**50 % split**	**66 % split**
Correctly Classified instances	78%	73.5%	74.1%	74.8%	69.4%	69.4.1%	75.4%
Incorrectly Classified Instances	22%	26.4%	25.8%	25.1%	30.5%	26.8%	24.5%
Kappa statistic	76%	45.3%	46.3%	48.6%	38.3%	46%	50.4%
Mean Absolute error	26.2%	27.9%	27%	25.3%	32.3%	28.4%	26.3%
Root Mean Squared error	48.2%	48.8%	46.9%	46.3%	52.2%	45.9%	45.03%
Relative Absolute error	57.4%	57.2%	55.5%	51.9%	65.1%	53.3%	53.3%
Root Relative Squared error	94%	98.9%	95.1%	93.9%	95.3%	94.7%	89.1%
TP Rate	68.1%	73.5%	74.2%	74.8%	79.4%	68.1%	72.5%
FP Rate	26.8%	28.5%	28.5%	25.9%	31.5%	24.4%	25.3%
Precision	71.8%	73.4%	74%	75%	75.7%	72.26%	75.6%
Recall	72.2%	73.5%	74.2%	74.8%	69.4%	73.3%	75.5%
F-Measure	73%	73.5%	74%	74.9%	67.2%	72%	75.3%
AUC	76.61%	78.3%	76.9%	71.3%	78.5%	74.1%	79.1%

The results from the Table 1.2 have been also displayed in the Figure 1.2. The figure clearly shows the diversity of values of particular metrics for different testing configurations.

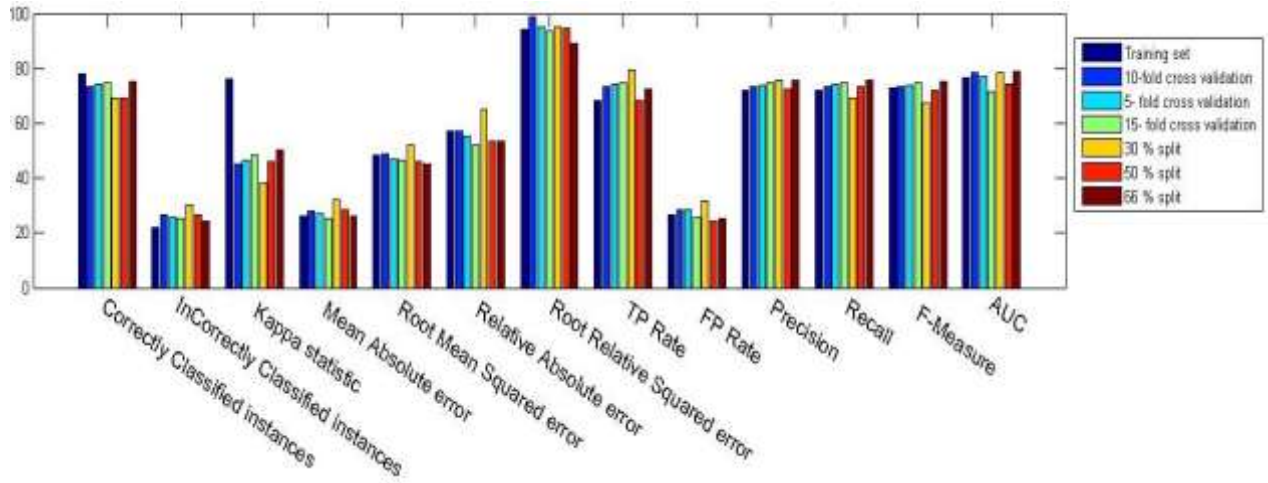

Figure 1.2: Relation between the performance measures and the testing configurations of the Multilayer Perceptron for the hepatitis database

6. EXPERIMENTAL RESULT

We are taking medical dataset .This section is dedicated to analysis of the results obtained during the calibration of the algorithms. Here the comparison of the algorithms in terms of performance is also done. This part of the examinations was conducted in the Experimenter graphical interface of the WEKA environment. The results obtained for the individual testing configuration do not make it possible to select the best one. The outcomes for individual metrics were very similar for all of the configurations in most of the cases. However, the results show that the 10-fold cross validation and the 66% split have a slight advantage over the rest of the testing configurations. However, there are cases where they yield one of the worst results. On the other hand, very often 50% split turned out to give bad results. Nevertheless, it is impossible to say which of the testing configurations gives the best results. Thus for the final evaluation of the algorithms the 10-fold cross-validation has been chosen. The reason for this is high popularity of this configuration.

The results of the comparison of the algorithms are presented in the Table 1.1. The table shows the performance of the algorithm in case of each of the performance measures and databases. The unquestionable leader in majority of cases is the C4.5. Then worst results this algorithm gained for the breast cancer databases. Nevertheless, overall performance was always better in comparison to other algorithms. When it comes to the Naïve Bayes, it wins the second place in terms of the performance. For most of the databases and metrics the results gained by this algorithm were slightly worse than for the Multilayer Perceptron in most of the cases. The worst results this algorithm delivered for the hepatitis data. Finally, the worst results were yielded by the Multilayer Perceptron decision tree. Its results were the worst in terms of both errors and AUC in comparison to all of the algorithms. The reason for this may the nature of medical data. Its complexity and heterogeneity of values of attributes can hinder data mining. In this case the Multilayer Perceptron and the Naïve Bayes may be over trained.

However, the results show that the 10-fold cross validation and the 66% split have a slight advantage over the rest of the testing configurations. However, there are cases where they yield one of the worst results. On the other hand, very often 50% split turned out to give bad results. Nevertheless, it is impossible to say which of the testing configurations gives the best results. Thus for the final evaluation of the algorithms the 10-fold cross-validation has been chosen. The reason for this is high popularity of this configuration.

CONCLUSION

In this paper, we examined various algorithms of classification. The main goal of the research was to identify the most common data mining algorithms, implemented in modern Medical Decision Support Systems, and evaluate their performance on several medical datasets. Three algorithms were chosen: C4.5, Multilayer Perceptron and Naïve Bayes. For the evaluation five UCI databases were used: heart disease, dermatology diseases, hepatitis, breast cancer and diabetes datasets. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, AUC, Precision, Recall, F-measure and a set of errors. The underlying reason for such a research was the fact that no work was found which would analyse these three algorithms under identical conditions. The variety of Medical Decision Support Systems makes it difficult to choose the most common data mining algorithms. Often a complete documentation of a system is not publicly available and it is difficult to know the algorithms implemented. Sometimes a system may be in a test phase and some part of its functionality may not be working yet.

REFERENCES

[1]. Aftarczuk K., Kozierkiewicz A., The method of supporting medical diagnosis based on consensus theory. Report of Institute of Information Science & Engineering, University of Technology. Wroclaw, 2006 Series PRE No. 1.
[2]. Aftarczuk K., Kozierkiewicz A, Nguyen N. T., Using Representation Choice Methods for a Medical Diagnosis Problem. Knowledge-Based Intelligent Information & Engineering Systems 2006, 805-812.
[3]. Alter S.L. Decision Support System: Current Practice and Continuing Challenge. Addisson-Wesley, 1980.
[4]. Autio L., Juhola M., Laurikkala J., on the neural network classification of medical data and an endeavor to balance non-uniform data sets with artificial data extension. Computers in Biology and Medicine, 2007, vol. 37, no. 3, 388-397.
[5]. Banfield R.E., Hall L.O., Bowyer K. W., Kegelmeyer W.P., A Comparison of Decision Tree Ensemble Creation Techniques. IEEE Computer Society, vol. 29, 2007.
[6]. Berrar D., Bradbury I. and Dubitzky W., Avoiding model selection bias in small-sample genomic datasets. Oxford University Press, 2006.
[7]. Berry J., Linoff G., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 2004, 2nd edition, Indianapolis, Wiley.
[8]. Brin S., Motwani R., Ullman J. D., Tsur S. Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Record, 1997, vol. 26, no. 2, 255–264.
[9]. Cancer Facts & Figures 2007. Retrieved on 26 04 2007.
[10]. Chae Y. M., Kim H. S., Tark K. C., Park H. J., Ho S. H., Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications, 2003, 167–172.
[11]. Child Ch., Stathis K., The Apriori Stochastic Dependency Detection (ASDD) Algorithm for Learning
[12]. Stochastic Logic Rules. J. Dix, J. Leite, and P. Torroni (eds), 2004, Proceedings of the 4th International Workshop on Computation, 201-216.
[13]. Cios K., Moore G., Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine, 2002, vol. 26, 1-24.
[14]. Comak E., Arslan A., Turkoglu I., A decision support system based on support vector machines for diagnosis of the heart valve diseases. Elsevier, 2007, vol. 37, 21-27.
[15]. Cosic D., Loncaric S., Rule-based labeling of CT head image. Lecture Notes in Artificial Intelligence, Berlin, Germany, Springer-Verlag, 1999, vol. 1211, 453–456.
[16]. Cunningham P., Carney J., Jacob S., Stability problems with artificial neural networks and the ensemble solution. Artificial Intelligence in Medicine, 2000, vol. 20, no. 3, 217–225.
[17]. Duch W., Adamczak R., Grabczewski K., Zal G., Hayashi Y., Fuzzy and crisp logical rule extraction methods in application to medical data. Computational Intelligence and Applications, Berlin, Germany, Springer-Verlag, 2000, vol. 23, 593–616.
[18]. Fayyad U. M., Data mining and knowledge discovery: Making sense out of data. IEEE Expert, 1996, vol. 11, no. 5, 20-25.