Biological Forum – An International Journal **8**(2): 103-107(2016)

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques

Shabia Shabir Khan and S.M.K. Quadri
Department of Computer Science,
University of Kashmir Srinagar, (Jammu and Kashmir), India

(Corresponding author: Shabia Shabir Khan)

(Received 08 June, 2016, Accepted 27 July, 2016)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: **Data mining is the process of uncovering the fluctuating hidden patterns or trends in the data that is not immediately apparent by just summarizing the data. It can help in predicting the future (predictive analytics) in addition to explain the current or past situation (descriptive analytics). After the interpretation of information, knowledge can be extracted by identifying relationships among patterns. Various data mining (machine learning) algorithms have been provided for extracting the nuggets of knowledge from medical datasets in the field of diagnostics. This paper discusses various machine learning techniques that have been evaluated using heart disease dataset for the prediction of class i.e. angiographic disease status (diameter narrowing). The main aim is to search a model that accurately predicts the class of the unknown records. The evaluation has been performed using WEKA software tool that helps in comparing the various techniques on the basis of certain important evaluation measures.**

Keywords: Data mining, Machine Learning, Angiographic disease status, classification.

INTRODUCTION

The data from different operational data sources is heterogeneous, huge (with respect to dimension as well as size) and scattered all over the network. It is near to impossible for human intelligence to discover potentially useful information from such a large amount of data, so we need a system that would extract the nuggets of knowledge and help us in strategic decision making. The various issues are resolved within the process of knowledge discovery using various data mining or machine learning techniques.

Data mining works almost in an opposite way of statistics wherein the first step does not start with the null hypothesis. Rather we just have a data set and we don't really know what and which pattern we are looking for. So, here we start by applying the interestingness criteria (notion) over the dataset in an attempt to get some interesting patterns forming the basis of the hypothesis thus the name "Hypothesis discovery" (Fayyad *et al.*, 1996).

Efforts are being made towards the exploration of knowledge by providing improved scalable interactive methods. The main aim is to be able to find certain patterns or trends in the data and forecast the future values of the data. Investigating data mining process, user interface issues, database topics, or visualization has always been a point of concern in the research area.

METHODOLOGY

Inspired by Machine learning and Statistics, the process of data mining has been provided that extracts the nuggets of knowledge (potentially useful information) from the huge amount of complex data. DM helps in finding out the unknown patterns in the data set that help in predicting something that we don't know. Data mining, considered to have been originated from three branches of artificial intelligence -neural networks, machine-learning and genetic algorithms leads us to such advancement (Agrawal and Srikant, 1995; Schumaker *et al.*, 2010; Prati *et al.*, 2004; Trueblood and Lovette, 2001; Han *et al.*, 2006). The various steps of knowledge discovery are shown in Fig. 1.

Given steps below represent basic principle of working for each of these classifiers which is same:

(i) Provide the training set that consists of the training records along with their associated class label.

(ii) The Classification model is built by applying the learning algorithm used in respective technique.

(iii) Finally the model built is applied on the test set that consists of the tuples that do not have the associated class label.

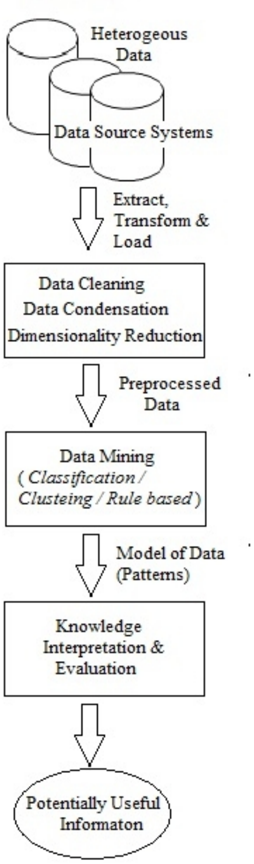

Fig. 1. Steps of knowledge discovery.

As far as training data is concerned, we can go for the cross-validation that involves the partitioning of the training data into mutually exclusive and same-sized subsets.

Data mining extracts the knowledge by grouping data into same cluster if they are similar enough or by grouping the data in separate classes if they are different enough (Hinton and Sejnowski, 1999) (Duda *et al.*, 2001) (Zhu *et al.*, 2003). However, the aim is same i.e. knowledge discovery.

Classification (supervised technique) is defined as a technique of building a model from the class-labelled predetermined dataset. The technique uses learning algorithms that generates the model which best fits the relationship between the predictors (attributes for prediction) and the prediction (class attribute) (Roiger and Geatz, 2003; Farahmandian, 2015). The main aim is to assign a correct label to the new arrived unlabelled instance. The technique of classification is implemented using certain algorithms like Naive Bayes, decision trees, Artificial Neural Network etc. Analysis is performed on one subset which is termed as the training set and the validation of the analysis is done using the other subset termed as the validation set or testing set. This is the case of simple one round cross-validation; however we can go for multiple rounds or fold cross validation that can be performed using different partitions in an attempt to reduce the variability.

EXPERIMENTAL EVALUATION MECHANISM

WEKA open source software consists of a collection of data mining learning algorithms and data pre-processing (transforming of dataset using filters) tools. The data set used has been described below:

Medical Data Set For Evaluation:

Title: Heart Disease Databases

Source Information:	UCI Machine Learning Repository (URL upload hakank.blogg)
Creator & Donor:	Dr. Andras Janosi, Dr. William Steinbrunn, Dr. Matthias Pfisterer, Dr. Robert Detrano (Creators) and David W. Aha (Donor)
Date:	July, 1988
No. Instances:	303 (Cleveland dataset)
Attribute Information:	14 (including Class attribute)
Class Angiographic disease status	-Value 0: < 50% diameter narrowing -Value 1: > 50% diameter narrowing
Relevant Information:	This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.

In Fig. 2, the sample dataset (Heart disease dataset 'HD.arff') was directly loaded in the WEKA software using the 'open URL' option.

In order to test the efficiency of our learning models we use training and test sets.

The training set, which is used to build a predictive model, consists of the predictor attributes as well as the prediction (class label) attribute. On the other hand we have the unseen test set, which is without any class label and is used to check the performance of the model trained. So, the next step is to split the "HD.arff" dataset into 30% testing set and 70% training set. For this we use the WEKA filter – "Randomize" Filter so as to create a random permutation.

Further, another filter "Remove Percentage" is applied two times. First by keeping option "invert Selection" as 'false' and then 'true' so as to keep the 30% of the dataset saved as a test set and rest as the training set , respectively.

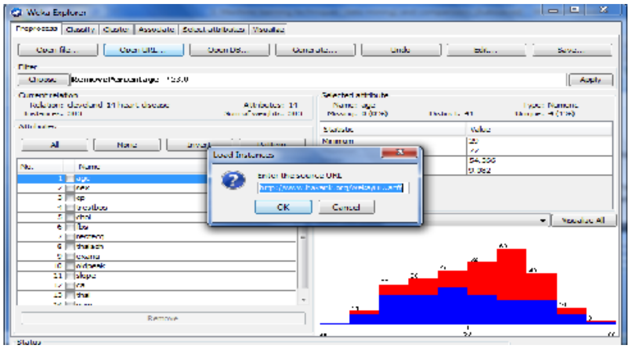

Fig. 2. Uploading 'HD.arff' Dataset using URL option

By following these above steps, we get two datasets:

The "Trainingdataset.arff" with 70% of the instances in the original datasets

The "Testdatset.arff" with 30% of the instances in the original data

As the number of instances in original dataset were 303.So, the training subset will contain 212 instances while as the test set will contain 91 instances.

RESULTS AND DISCUSSION

In WEKA, we can go for cross-validation process where same sized disjoint sets are created so as to train the model fold wise. In n-fold cross validation, data is being randomly divided into equal sized 'n' subsets or folds. Training is being done on n-1 subsets/folds and the left one fold is used for testing. The whole process is repeated n times in an attempt to use all the folds for testing thus allowing the whole of the data to be used for both training and testing (Keller, 2002). The Validation set is used to validate the trained model i.e. estimate how good the trained model is. In our experiment, 10 fold cross validation option is being selected.

Next to use our sets in the experiments we choose the training set and move to the "Classify" panel and choose the procedure that we have to use and start the experiment.

After that we apply the same procedure on our testing set to check what it predicts on the unseen data. For that, we select "supplied test set" and choose the testing dataset that we created. We run the algorithm again and we notice the differences in the confusion matrix and the accuracy.

As far as classification is concerned, rule mining is one of the most effective techniques and is considered as similar to the tree-based classification (Akeel, 2004). Apart from this, we have algorithms that extract the antecedent –consequent form of rule wherein the consequent is applied only if the antecedent proves to be true. Some of them include Finite automata, Neural Networks and Fuzzy controllers (Han *et al.*, 2006). Here, we shall focus on some rule generating classifiers available in weka, evaluate them using heart disease dataset and find out which classifier best predicts the class of the data instance based on several evaluation measures.

The rule based algorithms are:

(i) Zero Regression based Algorithm: The rule behind this pseudo regression algorithm is the consideration of the majority or common class of training data set to be taken as the real Zero R prediction. This algorithm predicts a value on the basis of training set average value. So, it relies on the target prediction and ignores all predictors. There is no predictability power of Zero R algorithm

(ii) One Regression Algorithm: The rule behind the algorithm is to find the single attribute that best predicts the class of the data. It generates a one-level decision tree and infers accurate rules that are easy to interpret. It works by creating one rule for each attribute in the training data and selects among them the best /one rule with the smallest/ lowest error rate.

(iii) Decision Tree Classifier: This classifier is the hierarchical structure, consisting of nodes and the directed edges that organizes series of questions about the predictors (attributes) and their possible answers in an effective way. The kind of the attribute determines the test condition in this classifier. Further, we can build a rule based (IF-Then) classifier by tracing the paths of the tree from root towards the leaf nodes (class labels) (Han *et al.*, 2006) (Shabia. *et al*, 2013). WEKA uses J4.8 algorithm for implementing C4.5 decision tree learner.

(iv) Random Forest. Random Forest Classifier is an ensemble technique wherein the predictions from multiple decision trees (base classifiers).

(v) Artificial Neural Network or Multilayer Perceptron: Artificial Neural network Multilayer Perceptron (MLP) has the ability to adapt and train through historical data, perform in a parallel processing, and work with multivariable system. All this makes it much similar to human brain.

The back propagation learning mechanism in neural networks is based on gradient descent technique and least square estimation (Werbose, 1974). Neural Network has the ability to data even if the knowledge about the data and relationships between the features or attributes is very less. Apart from this Neural Network can be applied on the dataset with continuous values.

It comprises of Input layer, hidden layer (can be more than one) and the output layer that are interconnected along with associated connection weights, to each other. The network learns by adjusting the connection weights so as to be able to predict the correct class label of the input vector. The input values are normalized within range [0, 1] for each attribute in the training set. This helps in speeding up the learning phase (Han et al., 2006).

In the software tool 'WEKA', the performance of the data can be checked using 2 important measures/ metrics:

Accuracy= (No. of correct predictions) / (Total No of Predictions)

Error rate= (No. of wrong predictions)/ (Total No of Predictions)

Following table 1 below shows the comparison of various evaluation measures obtained from different classifiers over heart disease dataset:

Table 1: Comparison between the Classification algorithms.

Evaluation Measures	**ZeroR**	**OneR**	**J48**	**Random Forest**	**Neural Network (MLP)**
Time taken to test model on supplied test set:	0.00 sec	0.03 sec	0.14 sec	0.02 sec	0.02 sec
Correctly Classified Instances	43	71	72	74	76
Incorrectly Classified Instances	48	20	19	17	15
Kappa statistic	0	0.5602	0.5837	0.6275	0.6705
Mean Absolute Error (MAE)	0.2044	0.0879	0.1086	0.1075	0.0727

Above comparison of some important evaluation measures proved that Multilayer perceptron (MLP) can provide better class prediction results. The classifier resulted in kappa statistic of 0.6705 which is higher than kappa statistic values of other classifiers and is much nearer to the value '1'. Further, the Mean Absolute error (MAE) value of 0.0727 has been obtained from MLP classifier which is lowest of all thus resulting in highest value of correctly classified instances than other classifiers (Table 1).

CONCLUSION

An overview has been presented to summarize the various data mining techniques that can help in efficient prediction for early medical diagnosis. This paper has experimentally proved that, for the same dataset, different algorithms work in different ways. As far as accuracy is concerned, the comparison between various rule based classifiers concluded neural network as an optimal model for classification in complex heart disease dataset. This is evident from the various evaluation measures like correctly or incorrectly classified instances, Kappa statistics, and mean absolute error, wherein the values obtained are better for neural network than any other classifier. This neural network model would help in accurately predicting theangiographic disease status which is the class attribute indicating percentage of diameter narrowing in diseased patients. This would in turn help in early medical diagnostics.

REFERENCES

Agrawal R, Srikant R (1995). Mining sequential patterns. In: Yu P, Chen A (Eds) Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, pp 3–14, March 1995.

Akeel Al- Attar (2004). White Paper: Data Mining - Beyond Algorithms, (2004).

Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification. Wiley-Interscience, 2nd Edition, (2001).

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. (1996). "The KDD process for extracting useful knowledge from volumes of data." *Communications of the ACM 39.11* (1996): 27-34.

Han, J., Kamber, M., and Pei, J., Data Mining: (2006). Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2nd ed., 2006.

Hinton, G. E. and Sejnowski, T. J. (1999). Unsupervised Learning: Foundations of Neural Computation, MIT Press, Cambridge, MA, MIT Press Publishers, 1999.

Keller, F., (2002). Evaluation Connectionist and Statistical Language Processing, Computer linguistik, Universitat des Saarlandes (2002).

Prati, R.C., Monard, M.C, Carvalho, (2004). Looking for exceptions on knowledge rules induced from HIV cleavage data set, *Genet Mol Biol.*, **27**(4): 637-43.

Roiger, R. and Geatz, M., (2003). Data mining: a tutorial-based primer, Boston, Massachusetts, Addison Wesley, (2003).

Schumaker, R.P. *et al.*, (2010). Sports Data Mining Methodology, Sports Data Mining, Integrated Series In:Information Systems 26, Springer Science+Business Media, LLC 2010.

Shabia S.K, Peer M. A., (2013). "Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume **3**, Issue 6, June 2013.

Trueblood, R.P. and Lovett, J.N., (2001). Data Mining and Statistical Analysis Using SQL, USA, Apress, 2001.

Werbos, Paul. (1974). "Beyond regression: New tools for prediction and analysis in the behavioral sciences." (1974).

Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty.(2003). "Semi-supervised learning using gaussian fields and harmonic functions." ICML. Vol. **3**. 2003.