International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:11, 2013

Comparison between Associative Classification and Decision Tree for HCV Treatment Response Prediction

Enas M. F. El Houby, Marwa S. Hassan

***Abstract*—**Combined therapy using Interferon and Ribavirin is the standard treatment in patients with chronic hepatitis C. However, the number of responders to this treatment is low, whereas its cost and side effects are high. Therefore, there is a clear need to predict patient's response to the treatment based on clinical information to protect the patients from the bad drawbacks, Intolerable side effects and waste of money. Different machine learning techniques have been developed to fulfill this purpose. From these techniques are Associative Classification (AC) and Decision Tree (DT). The aim of this research is to compare the performance of these two techniques in the prediction of virological response to the standard treatment of HCV from clinical information. 200 patients treated with Interferon and Ribavirin; were analyzed using AC and DT. 150 cases had been used to train the classifiers and 50 cases had been used to test the classifiers. The experiment results showed that the two techniques had given acceptable results however the best accuracy for the AC reached 92% whereas for DT reached 80%.

***Keywords*—**Associative Classification, Data mining, Decision tree, HCV, interferon.

I. Introduction

HEPATITIS C is an infectious disease affecting the liver, caused by the hepatitis C virus (HCV). The infection is often asymptomatic, but once established, chronic infection can progress to scarring of the liver (fibrosis), and advanced scarring (cirrhosis) which is generally apparent after many years [1]. An estimated 180 million people worldwide and nearly 4 million in the United States are chronically infected with HCV, leading to liver damage and increased risk of hepatocellular carcinoma [2]. In the United States, 10,000 deaths each year are attributed to chronic HCV infection [3]. The treatment for patients of HCV is combined therapy of pegylated Interferon alpha (PEG-IFN-α) and ribavirin (RBV), the number of responders to this therapy is low with high cost and unfavorable side effects. Different studies have proposed markers for predicting HCV patients' response to therapy. Markers may be based on viral factors such as viral load and genotype, host factors such as age, gender, body mass index (BMI), fibrosis and cirrhosis.

Data mining is the process of finding correlations or patterns among different fields in large databases. Application of data mining in the field of biology is considered an important research area. Data mining could be used for analyzing and finding hidden patterns inside patients' datasets. So, building a system for predicting patients' response for treatment of HCV is possible using different machine learning techniques. Many researchers had studied data mining using different machine learning techniques for analyzing and finding hidden patterns inside HCV patients' datasets. D. Wang et al. [4] had developed three models that predict virological response to therapy from clinical information. They compared accuracy of artificial neural network ANN, random forests (RF) and support vector machines (SVM). Lau-Corona et al. [5] had constructed Decision Trees (DTs) in patients with HCV. The recognition of clinical subgroups helps to enhance the ability to assess differences in fibrosis scores in clinical studies. Kurosaki M et al. [6], M. Hassan et al. [7] had developed DT model for predicting the probability of response to therapy with Peg-IFN and RBV in HCV patients. In [8], [9] Associative Classification (AC) had been used for developing models that predict patients' response to treatment of HCV from clinical information. In this research we compare the performance of the DT [7] and AC [8], [9] models.

Enas M.F. El Houby and Marwa S. Hassan are with the Systems & Information Dept., Engineering Division, National Research Centre, Dokki, Cairo, Egypt (e-mail: enas_mfahmy@yahoo.comeng, marwa_3@yahoo.com).

II. Material and Methods

A. Patients

Data from 200 Egyptian patients with hepatitis C virus who were treated with combined therapy PEG-IFN-α and RBV for 2 years, was collected at Cairo University Hospital. For each patient a record composed of 12 features including blood test features and patient characteristics features in addition to response feature, was constructed.

B. Data Preprocessing

In pre-processing phase, a series of data pre-processing steps were applied to clean, rank and select suitable features from patients' data to be in an appropriate form for applying the proposed machine learning techniques. The data included the following 12 features: Age; Gender; Body Mass Index (BMI); Albumin; Alanine Amino Transferase (ALT); Aspartate Amino Transferase (AST); Alfa-Feto Protein; Histology Activity Index (HAI); Viral load; Genotype; Fibrosis stage, and Cirrhosis.

For ranking these features, the value of each feature importance was calculated as (1 - P); where P is the value of the corresponding statistical test of association between the candidate feature and the target variable which is the response in our case. For categorical variables, the P values based on

Pearson's Chi-square were used, whereas for continuous variables the P values based on the F test were used [10]. The high ranked features which characterize the disease and effect the prediction had been selected to build our models. These selected features which are HAI, fibrosis stage and ALT features were collected together with response in a separated database to be in a suitable form for applying the proposed machine learning techniques. The data was partitioned randomly into training sets of 150 records and test sets of 50 records.

C. Data Mining Phase

Data mining is the process of finding patterns among different features in databases. In this study, decision tree and associative classification techniques had been used to build classification models which predict patients' response to treatment from selected features.

1. Decision Tree

The Classification and Regression Tree (CART) had been used to generate classification tree. Classification tree is built through a process known as binary recursive partitioning, which is an iterative process of splitting the data into partitions and then splitting it up further on each of the branches [11]. To partition the data at each stage of tree, a test is performed to select an attribute with lowest entropy. Information gain (IG) is used as a measure of entropy difference (H) when an attribute contributes the additional information about class C [12].

$$\text{Entropy} = H(C) = -\Sigma\, p(c) \log p(c),\ c \in C \quad (1)$$

$$\text{Remainder} = H(C|X_i) = -\Sigma\, p(x)\, \Sigma\, p(c|x) \log p(c|x),\ x \in X_i,\ c \in C \quad (2)$$

$$\text{Information gain} = IG_i = \text{Entropy} - \text{Remainder} \quad (3)$$

In (1), p(c) is the probability that an arbitrary sample belongs to class 'C'. Equation (2) shows the entropy after observing the attribute X_i for the class 'C' and p(c|x) is the probability that a sample in attribute branch X_i belongs to class 'C' [13].

2. Associative Classification

Associative Classification (AC) generates a set of Class Association Rules (CARs); which are learned and extracted from the available training data set. In this study PMA [14] had been applied to generate a set of CARs. CARs are generated from frequent rule items. A rule item is of the form <condset, class> where "condset" is a set of items; each item is (feature, value) pair. k-rule items, denote the patterns which condset has k items (where k=1, 2, 3). Once the frequent rule items are found, it is straight forward to generate CARs. The generated CARs are denoted as CAR_i (i= 1, 2, 3) according to number of items in L.H.S. The generated CARs are of the form:

$$\{\text{Feature name, value}\}_k \rightarrow \{\text{class, value}\}\ \text{(support, confidence)}$$

where the rule $X \rightarrow y$ has support s in dataset D if s% of the cases in D contains X and is labeled with class y, and confidence c if c% of cases in D that contain X is labeled with class y.

The highest precedence CARs have been selected and used in building classifier. The highest precedence rule selection is done by applying database coverage algorithm. Database coverage algorithm tries all generated CARs on training data, and selects highest precedence CARs that cover all data cases, any extra rules are redundant and useless rules, so it removes those rules and do not include them in the classifier [15], [16]. The classifier format:

$$<CAR_1, CAR_2, CAR_3, \ldots\ldots\ldots\ldots, CAR_n, \text{default class}>$$

III. Experimental Results

Extensive experimental studies had been tried to evaluate the AC and DT in predicting patients' response for treatment, 6 different classifiers had been built for each DT and AC by selecting randomly sets of 150 records for training and sets of 50 records for testing. By applying our techniques a great deal of statistical information was supplied to evaluate our models. This statistical information includes true positives (TP) and true negatives (TN), together with six performance measures which are sensitivity, specificity, positive predictive value, negative predictive value, Area Under Curve (AUC) and accuracy. In Tables I & II the performance of all DTs and ACs for different classifiers are recorded. As shown in Tables I & II the DT has sensitivity and specificity ranging from 54.5% to 88.9% and from 71.1% to 77.5%, respectively. While the AC has sensitivity and specificity values that diverse from 45.5% to 81.8% and from 89.3% to 100%. As for the positive predictive positive values, the values vary from 35.3% to 55.6%, for DT and from 76.9% to 100% for AC. Concerning the negative predictive values; they vary from 81.25 to 97.0% for DT and from 67.6% to 88.6% for AC. AUC values vary from 62.8% to 83.2% for DT whereas for AC they vary from 67.4% to 90.9%. The diagnostic accuracy for DT changes from 68% to 80%, whereas for AC changes from 70% to 92%. Figs. 1 and 2 show Receiver Operating Characteristic (ROC) curves for different DT and AC classifiers with their sensitivity and specificity values. Fig. 3 shows comparison between AUC for different DTs versus ACs whereas Fig. 4 shows comparison between accuracy for different DTs versus ACs classifiers. By comparing all performance measures for different ACs and DTs which are shown in Tables I and II, it is clear that AC outperforms DT by all performance measures.

The results including all performance measures of the best AC and DT which are AC6 and DT6 are indicated in Table III. By comparing AC6 and DT6 we can find that the maximum accuracy of AC reaches 92% whereas for DT reaches 80%. Comparing the ROC curves of DTs in Fig. 1 and for ACs in Fig. 2, it is clear that DT6 and AC6 are the closest to the top left of ROC curves, since DT6 and AC6 satisfy the highest values of sensitivity and specificity, and so satisfy the highest AUC from among the different DTs and ACs classifiers and that caused the increase in the accuracy of DT6 and AC6 to

their highest values from among the different DTs and ACs as indicated in Tables I and II. But still AUC for AC6 and so sensitivity, specificity and the accuracy are higher than those for DT6 as shown in Table III. Based on our results, we recommend using AC for the prediction of responders to HCV treatments. The development steps which were used in building the proposed models and comparing their performance are illustrated in Fig. 5. These steps started by cleaning data and ended by finding the best model for predicting patients' response to treatment.

IV. Conclusion and Future Work

The aim of this research is to compare the performance of AC and DT in the prediction of response to the treatment of HCV from clinical information. 200 patients treated with IFN and RBV; were analyzed and used to evaluate the two models. The experiment results showed that the two techniques had given acceptable results but AC outperformed DT. Although both models used the same features which are ALT, fibrosis score and HAI, AC model gave better performance measures than DT model. As sensitivity and specificity increase, AUC increases and so the accuracy. The best accuracy for the AC is 92 % whereas for DT is 80%. The proposed machine-learning techniques models benefit the patients by predicting the responders for HCV treatment which save patients from bad side effects and high cost.

In the future, we hope that we have more available data set to train our models and try more experiments and more analysis to the data. Also we hope to try many other techniques and compare our models with the other models to reach as high accuracy as possible.

TABLE I
Performances of Different DTs Models

Decision tree number	TP	TN	Positive predictive value %	Negative predictive value%	Sensitivity %	Specificity %	AUC %	Accuracy %
DT1	7	27	35.3	84.8	54.5	71.1	62.8	68%
DT2	7	28	38.9	87.5	60.0	71.8	66.9	70%
DT3	10	26	55.6	81.25	60.0	76.5	68.2	72%
DT4	8	28	44.4	90.6	70.0	74.4	72.2	74%
DT5	8	30	44.4	93.8	77.8	75.0	78.4	76%
DT6	8	32	47.0	97.0	88.9	77.5	83.2	80%

TABLE II
Performances of Different ACs Models

AC No.	TP	TN	Positive Predictive value	Negative Predictive value	sensitivity	specificity	AUC	Accuracy
AC1	10	25	76.9%	67.6%	45.5%	89.3%	67.4%	70%
AC2	9	29	81.8%	74.4%	47.4%	93.5%	70.5%	76%
AC3	11	29	78.6%	80.5%	61.1%	90.6%	75.9%	80%
AC4	12	29	85.7%	80.6%	63.2%	93.5%	78.4%	82%
AC5	13	31	86.7%	88.6%	76.5%	93.9%	85.2%	88%
AC6	18	28	100%	87.5%	81.8%	100%	90.9%	92%

TABLE III
Performance of the Best AC and DT

classifier number	TP	TN	Positive predictive value %	Negative predictive value%	Sensitivity %	Specificity %	AUC %	Accuracy %
AC6	18	28	100	87.5	81.8	100	90.9	92
DT6	8	32	47.0	97.0	88.9	77.5	83.2	80

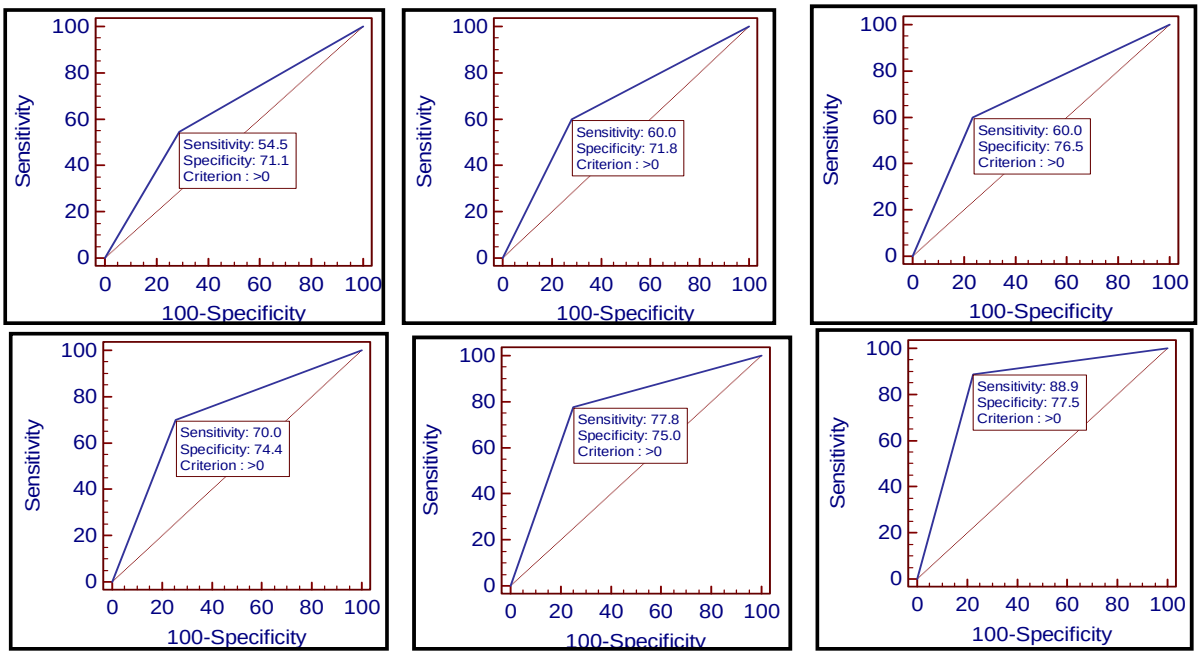

Fig. 1 ROC curves of six DTs with sensitivity and specificity values at the optimal cutoff points

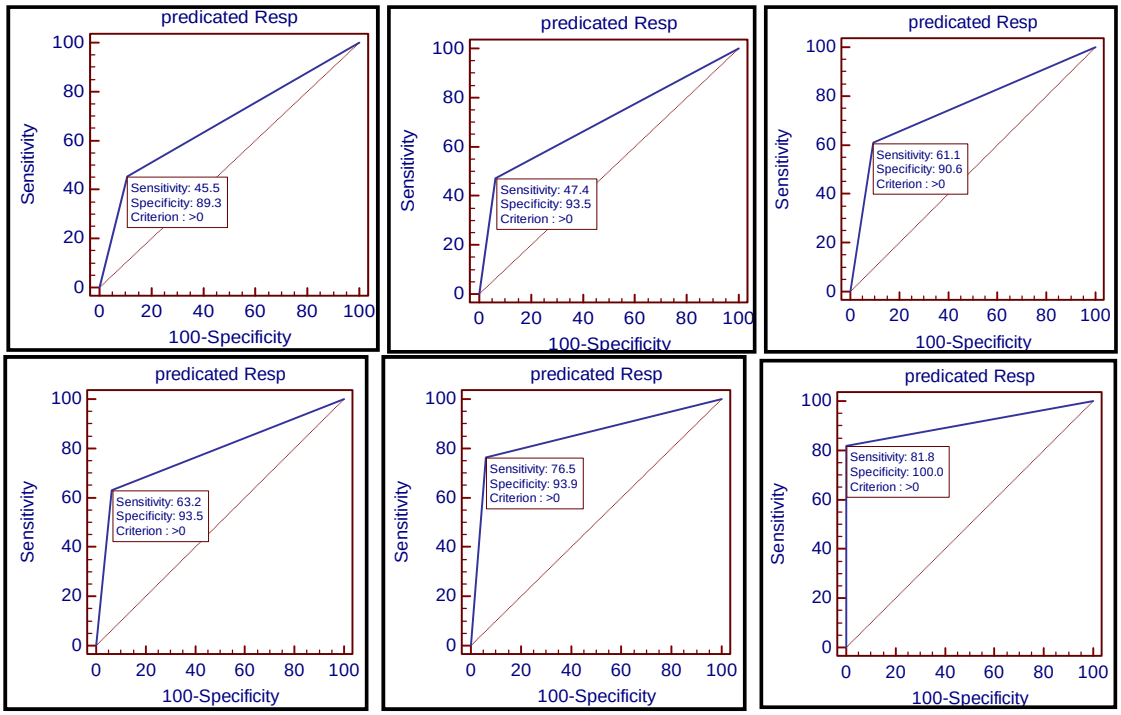

Fig. 2 ROC curves of six ACs with sensitivity and specificity values at the optimal cutoff points

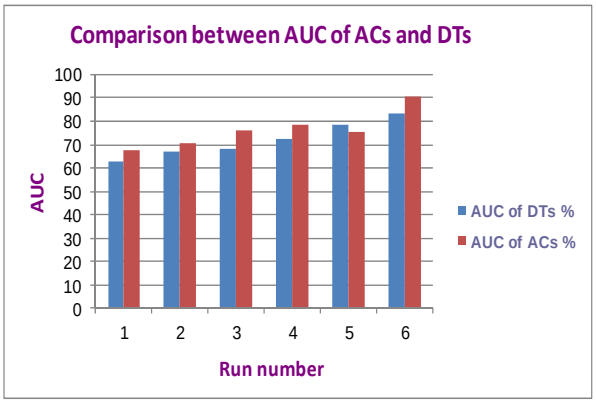

Fig. 3 Comparison between AUC for different DTs versus ACs

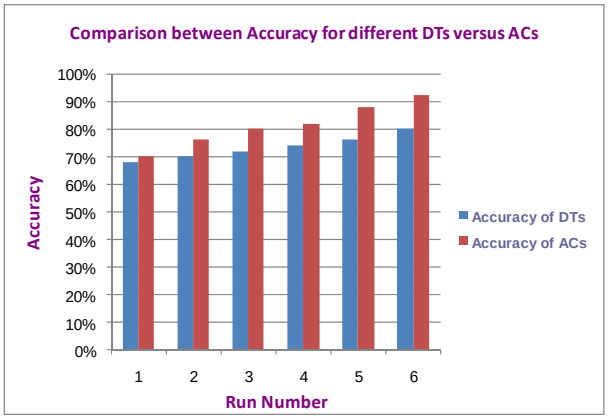

Fig. 4 Comparison between Accuracy for different DTs versus ACs

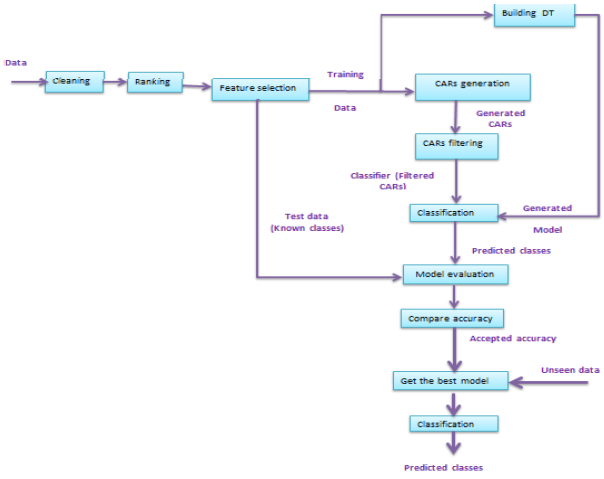

Fig. 5 The development steps of finding the best model for predicting patients' response to treatment

REFERENCES

[1] Roger Chou, MD; Elizabeth C. Clark, MD, MPH; Mark Helfand, MD and MPH," Screening for Hepatitis C virus infection: A review of the evidence for the U.S. preventive services task force", Ann Intern Med, vol.140, pp. 465–479, 2004.

[2] T. Asselah, I. Bieche, S. Narguet, A. Sabbagh, I. Laurendeau, M. Ripault, N. Boyer, D. Valla, M. Vidaud, and P. Marcellin, "Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C", NCBI, vol. 57,no.4, pp. 516-524 , 2008.

[3] Miriam J. Alter, Ph.D., Harold S. Margolis, M.D., Beth P. Bell, M.D., M.P.H., Steven D. Bice, M.Ed., Joanna Buffington, M.D., M.P.H., Mary Chamberland, M.D., Patrick J. Coleman, Ph.D., Beverley A. Cummings, M.P.H., and Catherine M. Dentinger, M.S., "Recommendations for prevention and control of Hepatitis C Virus (HCV) infection and HCV-related chronic disease", Center of Disease Control (CDC), Vol. 47, No. 19 , October 16-1998.

[4] D. Wang, B. Larder, A. Revell, J. Montaner, R. Harrigan, F. De Wolf, J. Lange, S. Wegner, L. Ruiz, M. Jésus J. Pérez-Elías, S. Emery, J. Gatell, A. Monforte, C. Torti, M. Zazzi, C. Lane, "A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy," Artificial Intelligence in Medicine, vol. 47, pp. 63-74, 2009.

[5] D. Lau-Corona, L. Pineda, H. Avilés, G. Reyes, B. Farfan-Labonne, R. Núñez-Nateras, A.Bonder, R. Martínez-García, C. Corona-Lau, M. Olivera-Martínez, M. Gutiérrez-Ruiz, G. Robles-Díaz, and D. Kershenobich, "Effective use of fibro test to generate decision trees in hepatitis C," Journal of Gastroenterology, vol. 15, no. 21, pp. 2617-2622, 2009.

[6] M. Kurosaki, K. Matsunaga, I. Hirayama, T. Tanaka, M. Sato, Y. Yasui, N. Tamaki, T. Hosokawa, K. Ueda, K. Tsuchiya, H. Nakanishi, H. Ikeda, J. Itakura, Y. Takahashi, Y. Asahina, M. Higaki, N. Enomoto, N. Izumi, "A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis", Hepatology research, vol. 40, no. 4, pp. 251-260, 2010.

[7] M. Hassan, M. I. Abdalla, S. R. Ahmed, W. Akil, G. Esmat, S.Khamis, M. ElHefnaw, "The decision tree mode for prediction the response to the treatment in patients with chronic hepatitis C", New York Science Journal, vol. 4, no. 7, pp. 69-79 ,2011.

[8] Enas M.F. El Houby, "Analysis of associative classification for prediction of HCV response to treatment", International Journal of Computer (0975 – 8887), New York, USA, Vol. 63– No.15, PP. 38-44, February 2013.

[9] Enas M.F. El Houby, Marwa S. Hassan, 2012. "Using associative classification for treatment response prediction", Journal of Applied Sciences Research, vol. 8, no.10, pp. 5089-5095, 2012.

[10] Floares A. G., Alexandru George Floares, "Artificial intelligence support for interferon treatment decision in chronic hepatitis B", World Academy of Science, Engineering and Technology, vol. 44, pp. 110-115, 2008.

[11] Binoy.B.Nair, V.P Mohandas, and N. R. Sakthivel, "A decision tree-rough set hybrid system for stock market trend prediction", International Journal of Computer Applications (0975 – 8887) Vol. 6, No.9, pp.1-6, September 2010.

[12] Basant Agarwal and Namita Mittal," Categorical probability proportion difference (CPPD): A feature selection method for sentiment classification", Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pp. 17–26, December 2012.

[13] T.G. Process, "CART algorithm," 1984, pp. 1-6

[14] Enas M.F. El Houby, 2010, "Mining protein structure class using one database scan", International Journal of the Computer, the Internet and Management (IJCIM), vol. 18, no. 2, pp. 8-16, 2010.

[15] Rattanakronkul, N. and K.Waiyamai, "Combining association rule discovery and data classification for protein structure prediction", The International Conference on Bio-informatics, 2002.

[16] Thabtah, F.A. and P.I. Cowling., "A greedy classification algorithm based on association rule", Applied Soft Computing, vol. 7, no. 3, pp. 1102-1111, 2006.