Analysis of Associative Classification for Prediction of HCV Response to Treatment

Enas M.F. El Houby
Systems & Information Dept., Engineering Division, National Research Centre, Dokki, Cairo, Egypt.

ABSTRACT

The objective of this research is the analysis of predicting the response for treatment in patient with hepatitis C virus. The Interferon Alfa (IFN) in combination with ribavirin (RBV) is used as a standard therapy for chronic hepatitis C (CHC), it is very expensive and accompanied with great side effects, with that it fails in more than half cases.

For the prediction of treatment response, a knowledge discovery framework including two main phases: pre-processing and data mining was developed. In pre-processing phase, the cleaning and selection of suitable features from patients' data were done. In data mining phase the selected patients' features were mined using Associative Classification (AC) technique to generate a set of Class Association Rules (CARs). The most suitable rules from the generated CARs were selected to build a classifier, which predicts patients' response for treatment. Using our classification model, 220 patients treated with IFN plus RBV were analyzed, 92 patients resulted responders and 128 non-responders at the end of treatment and during the follow up. 170 cases had been used to train our intelligent models and 50 patients had been used to test the models. The experiment results showed that the proposed technique is an effective classification technique with high prediction accuracy reach up to 90%.

General Terms

Knowledge discovery, Data mining, Bioinformatics.

Keywords

Associative classification, Class association rules, hepatitis C virus, interferon, ribavirin.

1. INTRODUCTION

Chronic hepatitis C is very common and often progressive form of chronic liver disease. Hepatitis C Virus (HCV) leads to liver fibrosis, cirrhosis and hepatocellular carcinoma [1]. Combined therapy using Interferon Alfa (IFN) and Ribavirin (RBV) represents the standard treatment in patients with chronic hepatitis C. However, the percentage of responders to the treatment is low, while its cost and side effects are high. Therefore, the possibility to predict patient's response to the treatment is very important. The progress in the field of informatics has led to the development of new techniques related to the artificial intelligence [2].

Extensive analysis have identified several factors that correlate with Sustained Virological Response (SVR), from these factors gender, age, viral load and degree of fibrosis on liver biopsy. Several of these predictive factors may be helpful clinically in guiding therapy, in providing advice about the likelihood of a response and in determining the duration of therapy [3, 4]. Such information has been used to predicate patients' response, understand the difference between responders and non responders. Furthermore early prediction of virologic response would allow a more appropriate selection of candidates helping to identify patients who are unlikely to have a sustained response and saving patients the side effects and cost of additional therapy [5].

Developing prediction models from various data sources is possible using a knowledge discovery in data or data mining techniques, and the prediction accuracy of the resulted intelligent systems could even reach high accuracy. Data mining could be used for analyzing and finding hidden patterns inside patients' datasets. Such patterns have been used to challenge the way to treat patients, understand the difference between responders and non responders. So, an intelligent system for predicting patients of HCV response for treatment can be built and be effective.

Classification and association rule are important data mining techniques. Using association rule to construct classification systems is known as Associative Classification (AC). AC is a promising classification approach, which is more accurate than traditional classification approaches. Building a prediction model using AC consists of two steps: generate CARs and building a classifier from the generated CARs. CARs generation is done by focusing on a special subset of association rules whose right-hand-side is restricted to the classification class attribute. The most suitable CARs are selected to build a classifier [6, 7, 8].

The aim of this research is the analysis of associative classification in predicting the response to the treatment with IFN plus RBV in patients with chronic hepatitis C. AC has been used for building a classifier using biological information obtained from patients' blood test. The used blood test is collected after treating patients for two years. The proposed model could help to predict treatment response from clinically relevant information.

This paper is organized as follows: In section 2, an overview of the previous works related to our research is presented. In section 3, the application of knowledge discovery technique for predicting HCV patients' response for treatment is described. In section 4, experiments are implemented to demonstrate the prediction accuracy results of the proposed technique. In section 5, conclusions and future works are drawn.

2. RELATED WORK

During the last decade, a huge amount of issues related to Hepatitis C virus (HCV) was investigated. The treatment of HCV is one of the most important issues. Many researchers have tackled clinical researches in the area of treatment of HCV and new information appears frequently. S. Wasik *et al.* [9] proposed a method for early-stage HCV patients' assessment, under which predictions can be made about efficiency of a treatment. Asselah, T. *et al.*, [10] presented the mechanism of non-response to overcome it and to identify factors that can help to predict the response to anti-HCV therapy. Berenguer M., et al. [11] developed a model based on pre-and/or early post-transplantation variables, which may predict progression to severe HCV disease recurrence. Moucari R., et al. [12] evaluated the efficiency of peginterferon alfa-2b and ribavirin in unselected consecutive patients with chronic hepatitis C, treated outside of trials, who were responders or non-responders to interferon and ribavirin combination.

Many researchers have studied data mining using different machine learning techniques for analyzing and finding hidden patterns inside HCV patients' datasets and predicting response of HCV patients to treatment. Wang, D. *et al* [13] 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, D., *et al*. [14] 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 in [15], Hassan, M. *et al.* [16] developed DT model for predicting the probability of response to therapy with Peg-IFN and RBV in HCV patients.

Associative classification has been successfully used in many applications. [6, 17] used AC to predict protein structure class from the other protein's features. In [18] they developed approach, using AC for pattern recognition on the NS5A protein and its motifs to find biomarkers for response prediction. In [19] AC had been used to predict response of HCV patients' to treatment, however in the current research extra 20 patients data have been added to our data set. A full description of our model and a detailed analysis of the results have been presented.

3. MATERIAL AND METHODS

3.1 Patients

The study included 220 Hepatitis C patients with genotype 4 at Cairo University Hospital who were treated with combined therapy interferon-Alfa and ribavirin for 2 years. Patients that showed clearance of the virus after 2 years were considered as responder and those who didn't show clearance of the virus were considered as non responder.

3.2 Knowledge Discovery Technique for HCV Patients' Response to Treatment

In this research, we applied knowledge discovery technique to predict HCV patients' response to treatment, according to a set of features. Our framework consists of two phases which are pre-processing and data mining. The pre-processing phase was done by a set of steps (e.g. data cleaning, features selection) to prepare data for data mining phase. The data mining phase was done by applying AC technique. AC generated a set of CARs; CARs were learned and extracted from the available training dataset. The most suitable rules were selected to build a classifier which predicts patient's response to treatment from a set of features.

3.2.1 Pre-processing phase

A database of 220 Egyptian cases was constructed from patients with hepatitis C virus genotype 4, who were treated with combined therapy IFN and RBV for two years. For each patient a record composed of 12 features was registered, in addition to response feature. These features include:

Blood test features are 9 features; which are Albumin, genotype, Alfa-Feto Protein, viral load, cirrhosis, Alanine Amino Transferase (ALT), Aspartate Amino Transferase (AST), fibrosis stage and Histology Activity Index (HAI)

Patient characteristics features are 3 other features; which are gender, age and Body Mass Index (BMI).

Response registers outcome based on PCR result of patient after treating for two years; the response feature takes either (0 or 1) value, 0 for non-responder patients and 1 for responder patients. Table1 contains the description of patients' features.

Table 1: List of HCV Patients' features

Features name	Values
Fibrosis stage	0-6
Cirrhosis	0=No (189) ; 1=Yes (31)
Age	years
Genotype	Genotype 4= 173; Non-Genotype 4 = 47
Gender	1= M (176) ; 2 = F (44)
Viral load	0.006 - 5.05 copies/ml
Body Mass Index (BMI)	16.6 - 43.2 Kg/(cm)^2
Histology Activity Index (HAI)	1 - 15
Alanine Amino Transferase (ALT)	0. 8 - 7.05
Albumin	2.5 – 5.4 g/L
Alfa-Feto Protein	0.02 – 3.25
Aspartate Amino Transferase (AST)	0.01 – 0.3
Response	0 = non-responder (128) , 1= responder (92)

For ranking the features, a simple method which considers one feature at a time was used, to test how well each feature alone predicts the target feature. For each feature, the value of its 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 feature. For categorical features, the P value was based on Pearson's Chi-square. For the continuous features, P values based on the F statistic were used. The importance of these features (1-P) was calculated from P and sorted by P value in ascending order or by (1-P) in descending order [20]. Table2, table3 show the categorical features and continuous features respectively sorted by P-value in ascending order and by (1-P) value in descending order. In order to construct AC model, the major features which characterize the disease and affect prediction had been identified in order to define the specific inputs for our model. So, 2 categorical features which are HAI, fibrosis stage and one continuous feature which is ALT had been selected according to their (1-P) value as inputs for our AC model. The 3 selected features and the response feature were collected together in a database file to be in a form suitable for applying the data mining technique. These features represent key information required from a patient who had had treatment, in order to build a model which could predict patient response in future cases. A sample of the selected features collected in a database file is shown in table4. It represents the output of the pre-processing phase and the input for our AC model.

Table 2: Chi square test for categorical features

Features	P-value	(1-P)
Fibrosis stage	0.001	0.999
Histology Activity Index (HAI)	0.001	0.999
Cirrhosis	0.03	0.97
Age	0.17	0.83
Genotype	0.362	0.638
Gender	0.75	0.25

Table 3: F test for continuous features.

Features	P-value	(1-P)
Alanine Amino Transferase (ALT)	< 0.0001	> 0.999
Viral load	< 0.0001	> 0.999
Body Mass Index (BMI)	< 0.0001	> 0.999
Albumin	< 0.0001	> 0.999
Alfa-Feto Protein	< 0.007	> 0.993
Aspartate Amino Transferase (AST)	< 0.02	> 0.98

Table 4: A sample of the generated database during pre-processing phase.

HAI	fibrosis stage	ALT	Response
6	2	3.6	1
9	4	3.6	1
7	4	1.1	0
7	4	1.1	0
7	4	1.2	1

3.2.2 Data mining phase

Data mining is the main phase of knowledge discovery. It is the process of finding patterns among different features in databases. In this study, we applied AC on training data to generate CARs, in “CARs generation” step. The generated CARs were used to build a classifier in “building classifier” step. This classifier has been used to predict patient’s response to treatment from the selected features.

3.2.2.1 Class association rules generation

The aim of “CARs generation” step was to generate CARs, which relate patient’s response to treatment with HAI, fibrosis stage and ALT features. CARs were generated from frequent ruleitems. A ruleitem is of the form < condset, class > where “condset” is a set of items; each item is (feature, value) pair. k-ruleitems, denote the patterns which condset has k items (where k=1, 2, 3). Once the frequent ruleitems were found, it is straight forward to generate CARs from them. The generated CARs are denoted as CARi (i= 1, 2, 3) according to number of items in L.H.S. A CAR is said to be large or frequent if its support s is greater than or equal a given minimum support threshold σ. The CARs that have a support and confidence greater than a given threshold are called strong rules. Where a rule X→ y holds in dataset D with confidence c if c% of cases in D that contain X is labeled with class y. The rule X → y has support s in D if s% of the cases in D contains X and is labeled with class y.

To generate CARs, the algorithm called PMA [17] had been adapted to be suitable for processing patients’ data which is numerical data instead of string data as was in the initial version of PMA. The adapted PMA had been applied to the training data set. The algorithm needed only one database scan to generate all rule items which relate patient features with the associated class (patient’s response). Frequent rule items which greater than support threshold were considered. The CARs had been generated directly from frequent rule items, where response represented right hand side of the rule (class 0 or 1).

A sample of generated CAR3, CAR2, CAR1 respectively are shown below:

{HAI, 7}, {fib_stag, 4}, {ALT, 1.1} → {response, 0}
(Support=2) (Confidence=100%)
{HAI, 7}, {fib_stag, 4} → {response,0}
(Support =2) (Confidence=66%)
{fib_stag, 4}→ {response, 1}
(Support =2) (Confidence=50%)

3.2.2.2 Building a classifier

In this step, a classifier had been built using a set of generated CARs in “CARs generation” step. Database coverage pruning selected highest precedence CARs for building classifier, CARs had been arranged from highest precedence to lower. The database coverage heuristic removed training objects covered by every evaluated rule and ensured that each rule must cover at least one training object to be part of the final classifier. In cases where the evaluated rule covers no training objects, it is redundant and it had been discarded. Precedence value was considered from confidence value, support value, number of items in the rules, and prior rule as shown below:

If rule1's confidence value is more than rule2's confidence
Then rule1 has higher precedence value
If rule1's confidence value is equal to rule2's confidence and rule1's support value is higher than rule2's support
Then rule1 has higher precedence value
If rule1's confidence value is equal to rule2's confidence and rule1's support value is equal to rule2's support and rule1's number of items is higher than rule2's
Then rule1 has higher precedence value
Otherwise the prior rule has higher precedence value

The classifier format: <CAR1, CAR2, CAR3, …………, CARn, default class> [6, 8, 21]. Figure 1 depicts the whole process of building a classifier using AC technique.

4. EXPERIMENTAL RESULTS

In order to test the performance of the AC in the prediction of patient’s response to treatment, a data set which includes 220 records of patients who were treated with IFN plus RBV had been divided into two parts: 170 cases for training and 50 cases for testing. At the end of follow up 92 (41.8%) of these patients resulted responders and 128 (58.2%) were non-responders. Since data size was very small we supposed that minimum support threshold was 2 in all our experiments, also minimum confidence threshold was supposed to be 50%. Extensive experimental studies had been tried to evaluate the AC in predicting patients’ response for treatment, 6 different classifiers had been built by selecting test cases randomly and using the remaining data for building classifiers. By applying our model a great deal of statistical information was supplied including true positives (TP), false positives (FP), true negatives (TN), false negatives (FP) together with six performance measures which are sensitivity, specificity, positive predictive value, negative predictive value, accuracy and Area Under Curve (AUC). These performance measures had been used to evaluate our model.

$$\text{Sensitivity} = \frac{\text{Tp}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{TN}{\text{TN} + \text{FP}}$$

$$\text{Positive Predictive value} = \frac{TP}{\text{TP} + \text{FP}}$$

$$\text{Negative Predictive value} = \frac{TN}{\text{TN} + \text{FN}}$$

$$\text{Accuracy} = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Table 5 shows the performance of different AC models for different runs using these statistical information values. Sensitivity and specificity varied from 36.4% to 76.5% and from 79.4 % to 97%, respectively. Concerning the positive predictive and negative predictive values, they varied from 61.1% to 93.3% and from 65.9% 88.9 %, respectively. The accuracy varied from 70% to 90 % and AUC varied from 66.4% to 86.7%. Figure2 shows Receiver Operating Characteristic (ROC) curves for different classifiers with sensitivity and specificity values. Figure3 shows Area under the curve (AUC) for different classifiers. Figure4 shows accuracy of different classifiers. Figure5 shows the relation between accuracy and AUC for different classifiers.

Table 5: The Performance of AC models

Run No.	TP	FP	TN	FN	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy	AUC
AC1	13	1	32	4	76.5 %	97%	92.9%	88.9 %	90 %	86.7%
AC2	14	1	30	5	73.7 %	96.8%	93.3%	85.7 %	88 %	85.2%
AC3	16	3	26	5	76.2 %	89.7 %	84.2 %	83.9 %	84%	82.9%
AC4	11	2	30	7	61.1 %	93.7 %	84.6 %	81.1 %	82 %	77.4%
AC5	11	7	27	5	68.7 %	79.4 %	61.1%	84.4 %	76%	74.1%
AC6	8	1	27	14	36.4%	96.4%	88.9%	65.9%	70%	66.4%

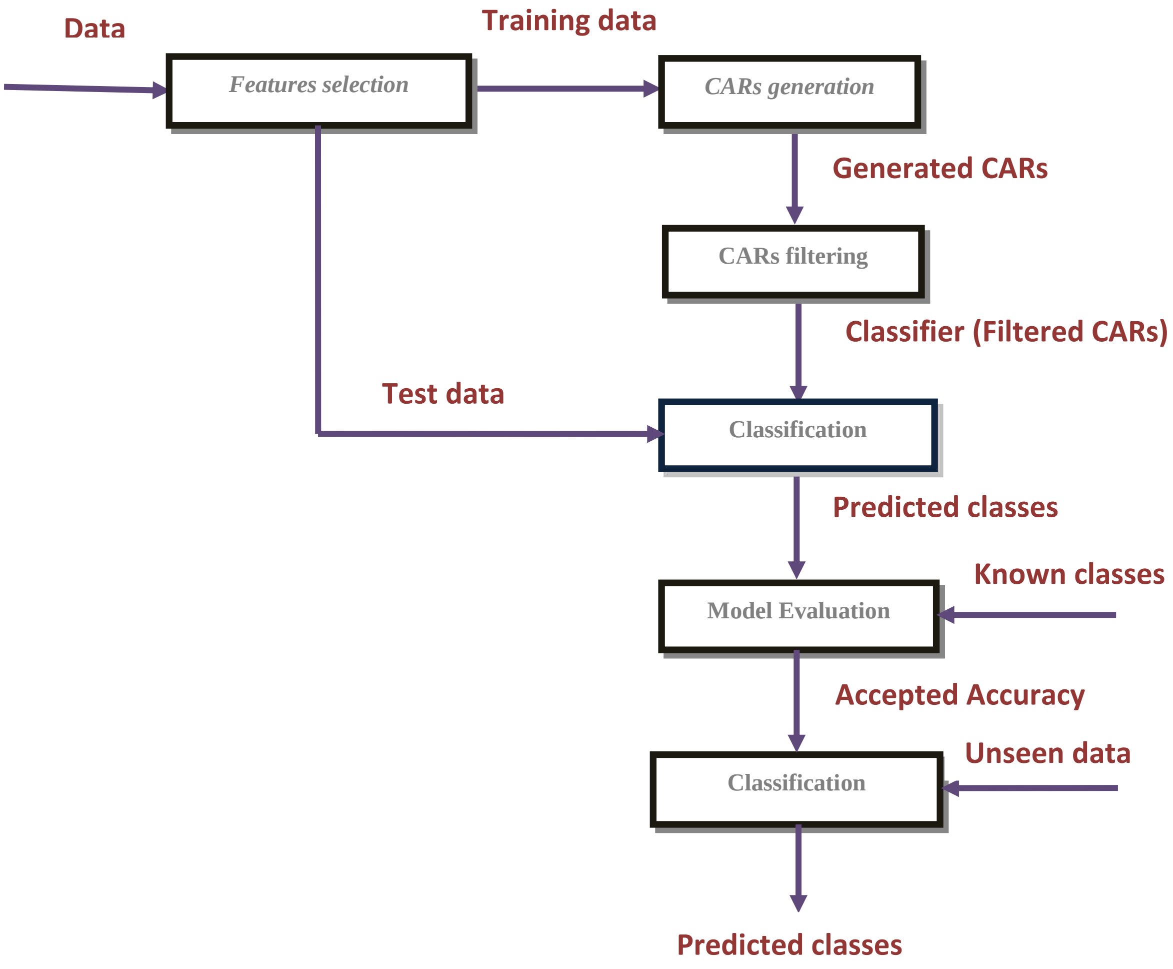


Figure 1: The whole process of building a classifier using AC technique.

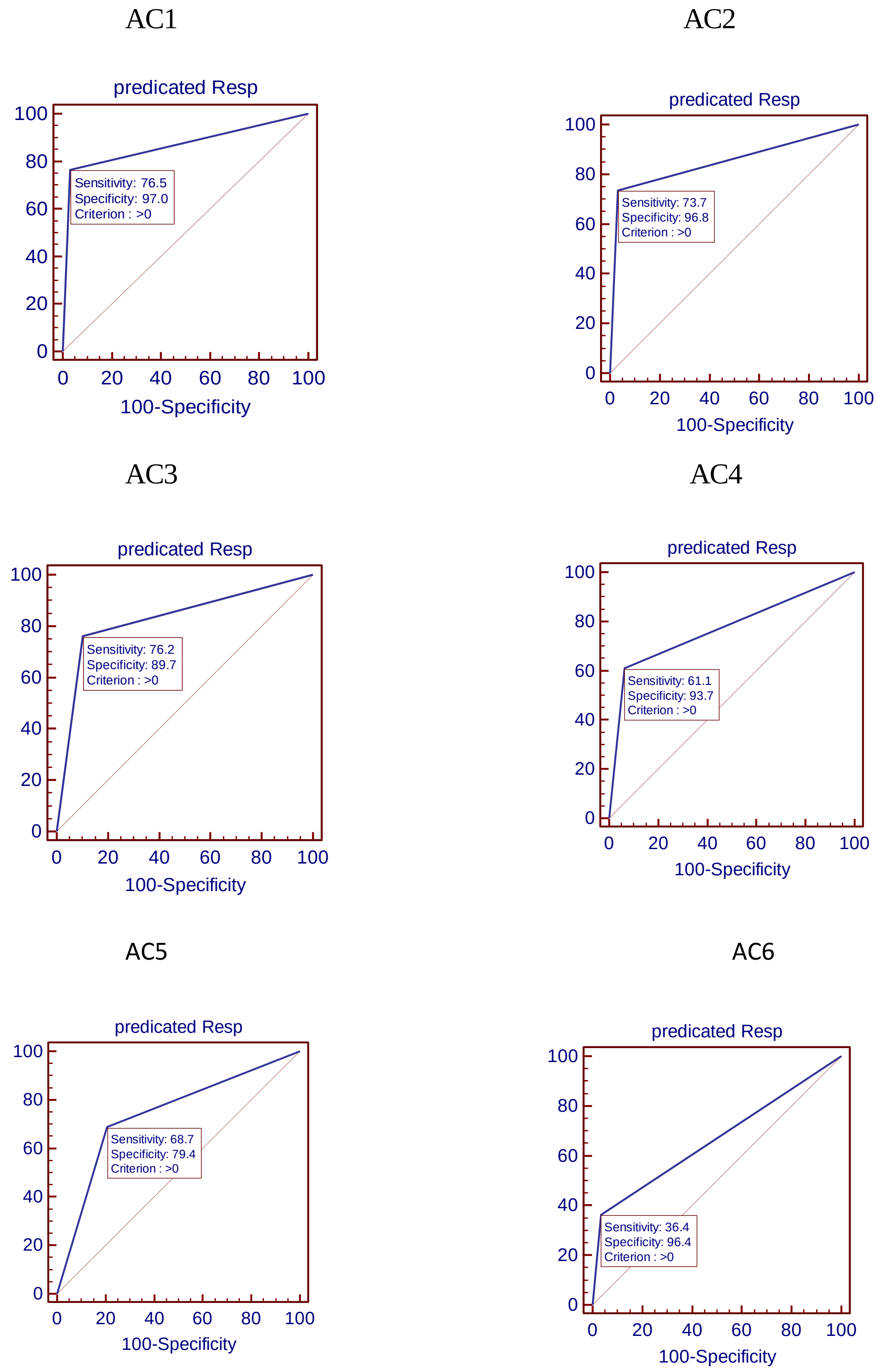


Figure 2: ROC curves of different classifiers with sensitivity and specificity values

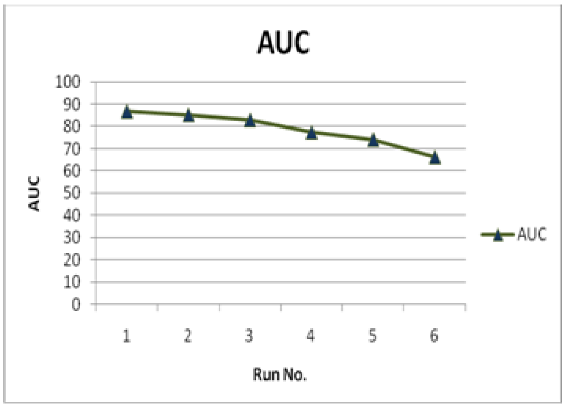
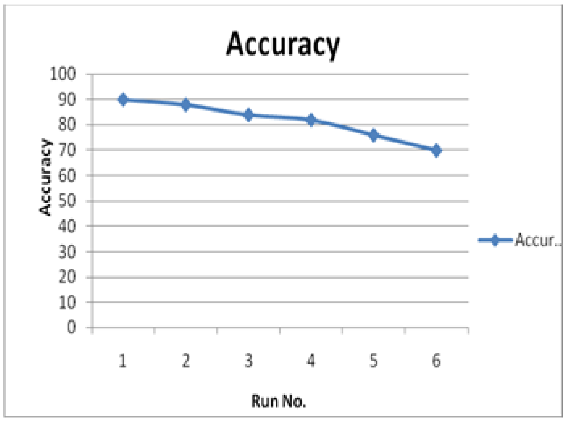


Figure 3: Area under the Curve of different AC models.

Figure 4: Accuracy of different AC models.



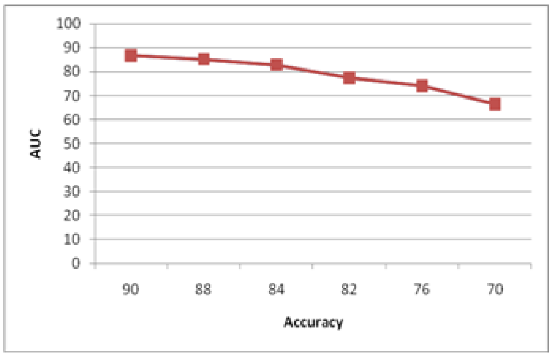


Figure 5: the relation between accuracy and AUC for ACs.

5. CONCLUSION AND FUTURE WORK

In this paper, associative classification had been used to predict response to treatment in patients with hepatitis C virus (HCV) from patients' blood test. AC technique had been used to generate a set of CARs. The most suitable CARs were selected to build a classifier which predicts patient's response to treatment. A data set which included only 220 records of patients who were treated with IFN plus RBV had been used to build and evaluate our model. Extensive analysis for the results had been done. The accuracy of the algorithm is high and reaches up to 90%.

However we need extra data to learn and extract more patterns for HCV patients. So in the future, we hope that we have more available data set to train our model and also we hope to try many other techniques and compare our model with the other techniques to reach as high accuracy as possible.

6. REFERENCES

[1] Seeff, L.B., "Natural history of chronic hepatitis C", Hepatology, 36, 2002: S35-S46.

[2] Maiellaro, P.A., et al., "Artificial Neural Networks for the Prediction of Response to Interferon Plus Ribavirin Treatment in Patients with Chronic Hepatitis C", Current Pharmaceutical Design, Vol. 10, No. 17, 2004, pp. 2101-2109.

[3] McHutchison JC, et al. "Predicting response to initial therapy with interferon plus ribavirin in chronic hepatitis C using serum HCV RNA results during therapy", J Viral Hepat 2001; 8: 414-20.

[4] McHutchison JG, Hoofnagle JH, 2000. "Therapy of Chronic Hepatitis C", In: Liang TJ, Hoofnagle JH eds. Hepatitis C. Biomedical Research Reports. San Diego (CA): Academic Press 2000; 203-239.

[5] Poynard T, et al., "Randomised trial of interferon alpha2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus", International Hepatitis Interventional Therapy Group (IHIT). Lancet 1998; 352: 1426-32.

[6] Rattanakronkul, N. and K.Waiyamai,"Combining Association Rule Discovery and Data Classification for Protein Structure Prediction", The International Conference on Bio-informatics (INCOP'2002).

[7] Thabath, F.,"A review of associative classification mining", Knowledge Engineering Review, 22(1),2007, 37-65.

[8] Thabtah, F.A. and P.I. Cowling, "A greedy classification algorithm based on association rule", Applied Soft Computing, 7(3), 2006, 1102-1111.

[9] Szymon, W.,"Towards prediction of HCV therapy efficiency", Computational and Mathematical Methods in Medicine, 11(2) 2010, 185-199.

[10] Asselah, T. *et al.*, "Hepatitis C: viral and host factors associated with non-response to pegylated interferon plus ribavirin", Liver International, 30, 2010, pp. 1259-1269.

[11] Berenguer M., et al., "A Model to Predict Severe HCV-Related Disease Following Live Transplantation", HEPATOLOGY, Vol. 38, No. 1, 2003, PP. 34-41.

[12] Moucari R., et al, "High predictive value of early viral kinetics in retreatment with peginterferon and ribavirin of chronic hepatitis C patients non-responders to standard combination therapy", Journal of Hepatology 46, 2007, PP. 596–604.

[13] Wang, D. *et al.*, "A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy," Artificial Intelligence in Medicine, 47, 2009, 63-74.

[14] Lau-Corona, D., *et al.*, "Effective use of fibro test to generate decision trees in hepatitis C," Journal of Gastroenterology, 15, 2009, pp. 2617-2622.

[15] Kurosaki, M. *et al*., "A predictive model of response to peg interferon ribavirin in chronic hepatitis C using classification and regression tree analysis", Hepatology Research, 40, 2010, pp 251-260.

[16] Hassan, M. et al., "The Decision tree Mode for Prediction the Response to the Treatment in Patients with Chronic Hepatitis C", New York Science Journal, 4(7), 2011, pp. 69-79.

[17] El-Houby E. M.F., "Mining Protein Structure Class Using One Database Scan", International Journal of the Computer, the Internet and Management (IJCIM), 18(2), 2010, pp 8-16.

[18] ElHefnawi, M., *et al*., "Prediction of prognostic biomarkers for Interferon-based therapy to Hepatitis C Virus patients: a metaanalysis of the NS5A protein in subtypes 1a, 1b, and 3a", Virology Journal 2010, 7: 130.

[19] El-Houby E. M.F., Hassan M. S., "Using Associative Classification for Treatment Response Prediction", Journal of Applied Sciences Research, 8(10): 5089-5095, 2012, ISSN 1819-544X.

[20] 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, 2008, pp. 110-115.

[21] Liu, B., *et al*., "Integrating Classification and Association Rule Mining", the proceedings of the International1998 Int. Conf. KDD'98. New York, USA, pp: 80-86.