AN OPTIMIZED MULTI-RELATIONAL CLASSIFICATION MODEL FOR THE PREDICTION OF HEPATITIS

Dr. *P. G. Sivagaminathan*[1], *Dr. C.R. Vijayalakshmi* [2], *Dr. M. Thangaraj*

ABSTRACT

Data mining techniques are commonly applied in different areas such as health-care, banking, manufacturing industry etc. Medical data are of immense use for predicting patients' health conditions. In health-care, normally, data mining algorithms are used for the analysis of various diseases, to predict the patients at risk of particular diseases and to suggest better medical services at reasonable cost. The present study is designed to provide an efficient multi-relational model for the prediction of ECML/PKDD'05 Hepatitis data. The main idea of this work is to find whether a patient is suffering from hepatitis, if he does what type (B or C) of it and the stages of liver fibrosis (from F0 to F4). The classification model was generated based on fuzzy rule-based classifier.

Keywords: Multi-relational data, Class label propagation, Genetic algorithm and Data Cleanser.

I. INTRODUCTION

Data mining methods are widely used for extracting significant patterns and regularities from big databases. The database contains multiple relations which are connected by primary/foreign keys. Nowadays, data mining techniques are frequently applied in clinical industry. The most demanding and fascinating task of data mining in clinical industry is predicting a disease. The data produced by healthcare industries are enormous and complex, which are stored in multiple database relations. The data contain details about hospitals, patients, medical claims, treatment cost etc. Moreover, data are too sparse and highly complex in nature. Hence, it is very difficult to analyze these data based on single-table method in order to make important decision regarding patient health. Therefore, there is a need to produce a powerful tool for analyzing and extracting important information from this complex multi-relational data.

In this paper, we propose an Optimized Multi-relational Fuzzy Rule System (OMRFRS) for hepatitis classification and stages of fibrosis. For this task, the multi-relational model of the dataset is considered that facilitates the analysis. The fuzzy rule-based decision system is used to find the patterns and associations among multiple database relations that characterize the patient behaviors to predict whether a patient has hepatitis or not. Next, the generated fuzzy rules are optimized by using genetic algorithm [3], [4].

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 outlines the proposed architecture which is being implemented. Section 4 provides the evaluation results and Section 5 discusses the conclusion.

[1]Assistant Professor, Dept.of CS,CA&IT, KAHE, Coimbatore

[2]Assistant Professor, Dept.of CS, MKUCA, Theni

[3]Head, Dept. of Computer Science Madurai Kamaraj University, Madurai.

II. RELATED WORK

A performance study was made for two classifiers namely Bayesian and lazy learners for hepatitis and thyroid dataset in [13]. The algorithms categorize the hepatitis patients into dead or alive class and thyroid data into hyperthyroid and hypothyroid. The Bayesian algorithm included two methods such as BayesNet and NaïveBayes. The lazy learners comprised IBK and KStar. The outcomes illustrated that the NaiveBayes method had superior accuracy for hepatitis dataset and BayesNet classifier achieved the best accuracy for thyroid data.

A fuzzy expert system based on genetic algorithm was proposed in [6] to diagnose Coronary Artery Disease. The parameters of the fuzzy membership functions were finetuned by a genetic algorithm. The decision trees were used for feature selection and discovery of rules. In [12], fuzzy framework was designed for cholera diagnosis and health monitoring system. It used Mamdani's inference engine, triangular method for fuzzification and center of gravity for defuzzification.

The hybrid approach was proposed in the paper [9] for analyzing cancer with the help of informative genes. This approach used the K-means clustering method based on statistical analysis-ANOVA for gene selection and support vector classifier (SVM) to classify the cancer diseases. The experimental results on micro-array data showed that the accuracy of K-means clustering with the combination of statistical analysis was better. In [1], Fei (et al)., proposed Particle Swarm Optimization-Support Vector Machines (PSO-SVM) model for analyzing arrhythmia cordis to ensure the health of humans and save human life. In this study, PSO was used to determine the parameters of SVM. This research was carried out with MIT-BIH ECG database. On the basis of the experimental results, it was proved that the accuracy of the proposed model was better than the accuracy of artificial neural network in the diagnosis of arrhythmia cordis.

III. PROPOSED SYSTEM

The architecture of the proposed OMRFRS is shown in Fig. 1. The motivation of this work is to design a decision- making system that will categorize whether the patient is affected by hepatitis disease, and if he does, its type and stage. The classification model of hepatitis disease is designed based on the system in [11] using fuzzy logic.

First, OMRFRS accepts Hepatitis data as input that contains a patient's details and medical tests. Next, class label propagation element is used to transmit the target class information from Biopsy relation to other relations in the dataset. The data cleanser module is used to remove the noises and inconsistencies present the data by applying Correlation-based Feature Selection method [10] followed by K-nearest neighbors [8].

After that, OMRFRS generates multi-relational classification rules by applying PART and fuzzy classification techniques. Fig.2 shows the fuzzification and defuzzification process by using triangular membership function and Centre of Gravity method. The generated rules are optimized with the help of genetic algorithm. Fig.3. shows the work flow of OMRFRS.

IV. PERFORMANCE ANALYSIS

The performance of the OMRFRS is measured, and the results are compared with the CrossMine system [15] using the following metrics. Table 1 shows the parameters used for this study. The experimentation is carried out on Intel Core i5 2.67 GHz with 4 GB RAM, running Windows 7, and implemented using Java based on WEKA [2]. A ten-fold cross validation is used for this evaluation.

The metrics are

- Accuracy of the classifier: Percentage of correctly classified tuple. Eq. (1).

$$\text{Accuracy} \quad \frac{\text{No. of tuples correctly classified}}{\text{Total noop tupies}} \qquad (1)$$

- Sensitivity: - True positive rate and it is specified as the number of people with the disease who will hold a positive result. Eq. (2).

$$\text{Sensitivity} \quad \frac{\text{No. of tuples that are truly possitives}}{\text{No. of true possitives t - no of false nagative}} \qquad (2)$$

- Specificity: - True negative rate, and it indicates the number of people without the disease who will hold a negative result.

$$\text{Specificity} \quad \frac{\text{No. of true negative tuples}}{\text{No. of true negatives - No of False positive}} \qquad (3)$$

- Run time: Classifier's Induction time in seconds
- Rule Set: Number of rules generated by the classifier

A. The Hepatitis Dataset

This dataset is a customized form of the PKDD'02 Discovery Challenge database and was proposed by Frank et al. [5]. In this dataset, Biopsy is the target relation that contains 206 records of Hepatitis B and 484 records of Hepatitis C. The attribute type is used as a class label for classifying patients with types (B or C). The Fibrosis attribute is the class label for categorizing patients with fibrosis stages. The number of non-target tables in the modified database is 4. This dataset has several tables [14] and its relational model [7] is shown in the Fig. 4.

The predictive performance of OMRFRS based on sensitivity and specificity is shown in Table 2. From this table, we observe that the OMRFRS gets maximum specificity and sensitivity value for both types of classifications when compared to CrossMine.

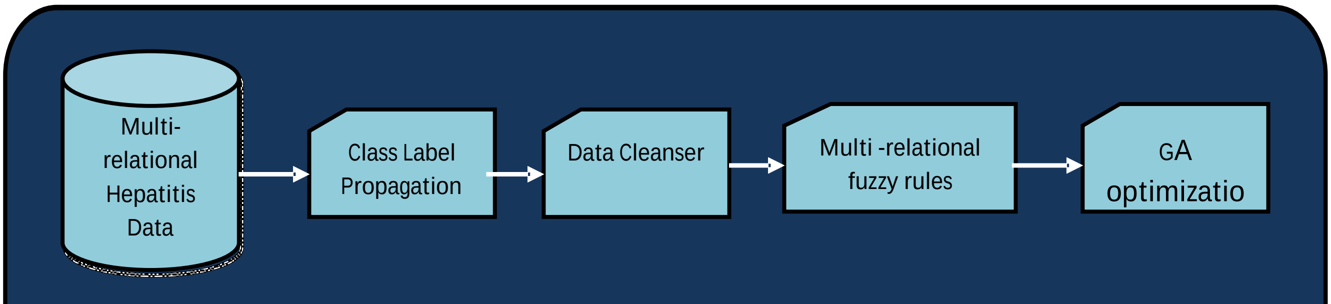
Multi-relational Hepatitis Data
Class Label Propagation
Data Cleanser
Multi -relational fuzzy rules
GA optimizatio

Fig. 1. Architecture of OMRFRS

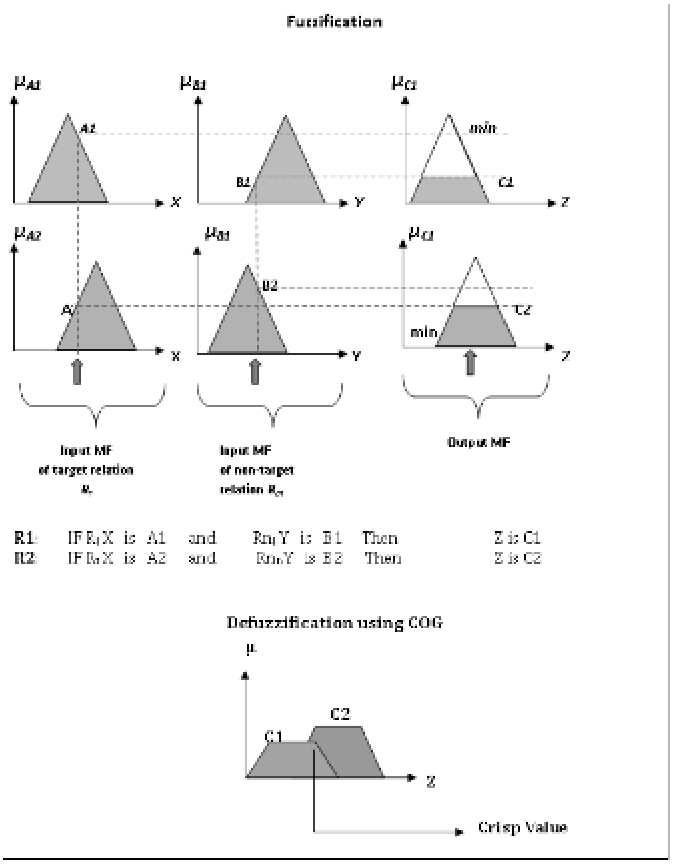
Fuzzification
μA1
μB1
μC1
A1
B1
min
C1
X
Y
Z
μA2
μB2
μC2
A2
B2
C2
min
Input MF of target relation Rt
Input MF of non-target relation Rn
Output MF
R1: IF Rt.X is A1 and Rn.Y is B1 Then Z is C1
R2: IF Rt.X is A2 and Rn.Y is B2 Then Z is C2
Defuzzification using COG
μ
C1
C2
Z
Crisp Value

Fig. 2. Process of Fuzzy Classifier

The runtime of the OMRFRS against CrossMine system is shown in Fig. 5. This figure indicates that the OMRFRS is quicker than CrossMine for Hepatitis dataset. This is because in OMRFRS, rules are produced in parallel by fuzzy classifier, whereas in CrossMine rules are generated in a sequential manner.

The accuracy of OMRFRS and CrossMine system is shown in Fig.6. This graph illustrates that the accuracy of OMRFRS is superior to CrossMine for both types of classification. The OMRFRS gets higher accuracy compared to CrossMine for Fibrosis stages.

The rule giving comparison between OMRFRS and CrossMine is shown in Fig. 7. This graph explains that OMRFRS generates fewer number of rules than CrossMine, because it generates merely optimized rules with the help of genetic algorithm.

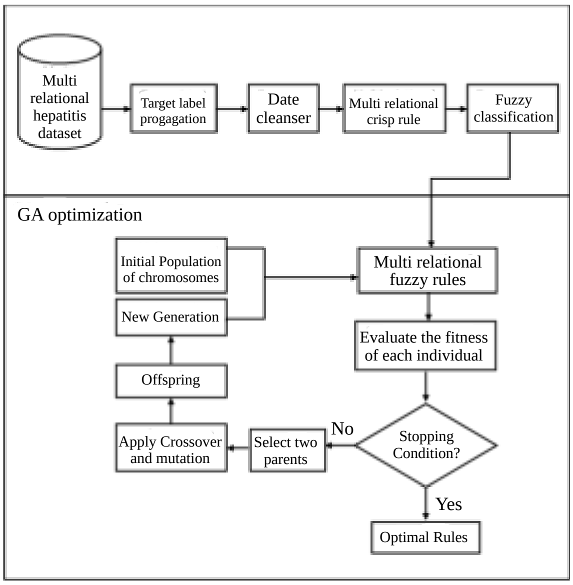
Multi relational hepatitis dataset
Target label progagation
Date cleanser
Multi relational crisp rule
Fuzzy classification
GA optimization
Initial Population of chromosomes
New Generation
Offspring
Apply Crossover and mutation
Select two parents
No
Multi relational fuzzy rules
Evaluate the fitness of each individual
Stopping Condition?
Yes
Optimal Rules

Fig.3. Workflow of OMRFRS

Table 1. Parameter of Genetic algorithm

Parameters	Values
Number of generations	50
Crossover probability	0.8
Number of crossover points	3
Mutation probability	0.01
Number of individual in the generation	30

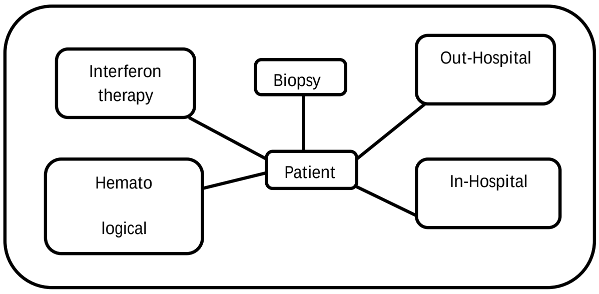

Fig.4. Relational Model of Hepatitis

Table 2. Predictive performance of OMRFRS

Method	Hepatitis Type	Hepatitis Stages
CrossMine		
Accuracy	77.2	81.2
F-measure	0.76	0. 79
OMRFRS		
Accuracy	86.7	90.0
F-measure	0.85	0.89

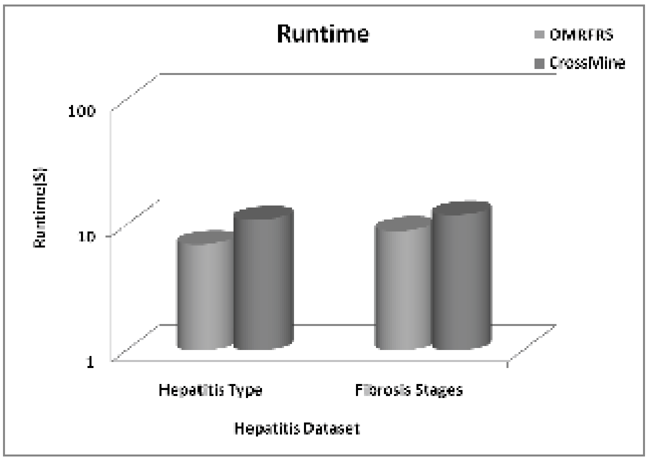

Fig. 5. Runtime of OMRFRS on Hepatitis Dataset

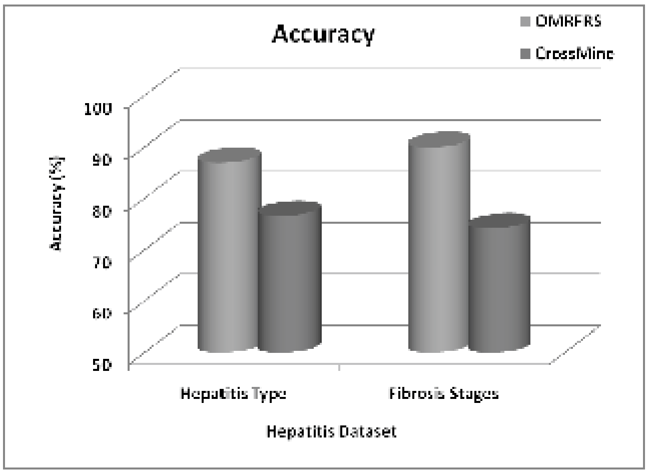

Fig. 6. Accuracy of OMRFRS on Hepatitis Dataset

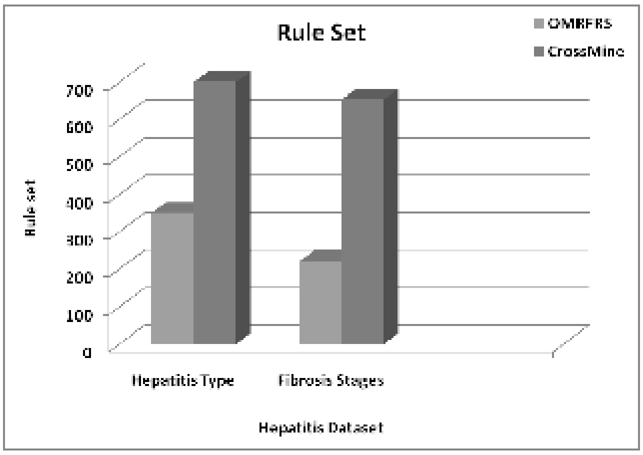

Fig. 7. Rule set generation on Hepatitis dataset

In this paper, an optimized multi-relational classification model for Hepatitis disease prediction was designed using ECML/PKDD 2005 dataset. The performance of OMRFRS was measured using accuracy, sensitivity and specificity metrics. The integration of clinical decision support system and rule-based methods with fuzzy logic minimized the time consumed for the classification of hepatitis disease. It also produced optimized fuzzy rules based on genetic algorithm.

VI. REFERENCES

[1] Fei, S.W., 2010, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine," Expert Systems with Applications, 37(10), pp. 6748-6752.

[2] Ian, H.W., and Frank, E.,2000, " Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," Morgan Kaufmann Publishers.

[3] Lo, C.H., Chan,P.T., Wong, Y.K., Rad, A.B., and Cheung, K.L., 2007, " Fuzzy genetic algorithm for automatic fault detection in hvac systems," Applied Soft Computing, 7(2), pp. 554-560.

[4] Mankad, K., Srinivas Sajja, P., and Akerkar, R., 2011, "Evolving rules using genetic fuzzy approach - an educational case study," Int. J. on Soft Computing, 2(1).

[5] Neville, J., Jensen, D., Friedland, L., and Hay, M, 2003, " Learning relational probability trees, " In S. Dzeroski and N. Lavrac, editors, Proc. of the ninth ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, pp. 625-630

[6] Niranjana Devi, Y., and Anto, S., 2014. "Investment risk analysis an evolutionary fuzzy expert system for the diagnosis of coronary artery disease," Int. J. Advanced Research in Computer Engineering and Technology, 3(4), pp. 1478-1484.

[7] Pizzi, L., Ribeiro, M., and Vieira, M., 2005, "Analysis of hepatitis dataset using multi-relational association rules," Proc. ECML/PKDD 2005 Discovery Challenge.

[8] Scheel, I., Aldrin, M., Glad, I.K., Sørum, R., Lyng, H. and Frigessi, A. (2005) ' The influence of missing value imputation on detection of differentially expressed genes from microarray data', Bioinformatics, Vol. 21, 4272-4279. doi:10.1093/bioinformatics/bti708.

[9] Soliman, T.H.A., Sewissy, A.A., and Latif, H.A, 2010, "A gene selection approach for classifying diseases based on microarray datasets," Proc. Of the 2nd Int. Conf. on Computer Technology and Development (ICCTD), Cairo. IEEE.

[10] Tan, F. (2007), Improving feature selection techniques for machine learning (Ph.D. thesis), Computer Science Dissertations: Paper 27, Retrieved from http://digitalarchive.gsu.edu/cs diss/27.

[11] Thangaraj., M and Vijayalakshmi, C.R "An efficient multi-relational framework using fuzzy rule based classification technique", International Journal of Data Mining, Modelling and Management (In press).

[12] Umoh, Ntekop, U.A., and Mfon, M., 2013, "A proposed fuzzy framework for cholera diagnosis and monitoring," Int. J. of Computer Applications, 82(17), pp. 0975-8887.

[13] Vijayarani, S., Janani, R., and Sharmila, S., 2015, "Data mining classification algorithms for hepatitis and thyroid data set analysis," Proc. Int. Conf. on Computing and Intelligence Systems, volume 04 of Special Issue, pp. 1270-1275.

[14] Yamada, Y., Suzuki, E., Yokoi, H., and Takabayashi, K., 2003, " Decision-tree induction from time-series data based on a standard-example split test," Proc. of 20th Int. Conf. on Machine Learning (ICML'03), pp. 840- 847. Morgan Kaufmann.

[15] Yin, X., Han, J., and Yu, P.S., 2006, " Efficient classification across multiple database relations: A crossmine approach,: IEEE Transactions on Knowledge and Data Engineering, 16(6).