ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

An Effective model for mutagenesis prediction using Multi relational Fuzzy Tree

[1] Dr. C.R.Vijayalakshmi, [2] Dr. P.G Sivagaminathan, [3] Dr.M.Thangaraj
[1] Assistant Professor, Dept.of CS, MKUCA, Theni,TN, [2] Assistant Professor, Dept.of CS,CA&IT, Karpagam academy of Hr. Education,Coimbatore,TN, [3] Associate Professor, MK University, Madurai, TN

***Abstract*: -- Most of the real world applications such as Loan approval, Credit card fraud detection etc uses relational databases which contain multiple relations that are inter-linked with the help of primary and foreign keys. It is very tricky to examine these applications with the help of traditional classification methods such as RIPPER and RIDOR. These methods are suitable for single relation and generate simple and comprehensible rules. But it cannot handle uncertainties and noises present in the real dataset. This paper presents a novel method for generating multi-relational classification model for mutagenesis prediction. The classifier is constructed based on the fuzzy extension of the decision tree. The experimental results show the efficiency of the proposed method compared to the existing algorithms.**

Index Terms: - CrossMine, Multi relational classification, Mutagenesis, target relation, tuple id propagation.

I. INTRODUCTION

Today, the data collected from various real world applications are vast and stored in relational databases. The databases contain several relations which are connected by primary key and foreign key links. Many existing data mining algorithms are not appropriate for handling these databases because it focuses on single relation. Therefore, the development of Multi Relational Data Mining (MRDM) has attracted many researchers. MRDM is suitable for variety of areas including bioinformatics, environmental science or engineering and healthcare, business data analysis, text and Web mining. The bioinformatics applications comprise of drug design, predicting mutagenicity and carcinogenicity and predicting protein structure and function, including genome scale prediction of protein functional class. One important technique of MRDM is Multi Relational Classification (MRC). The MRC categorize the data in target relation with the support of vital information stored in the remaining non-target relations. The MRC either convert multi relational data into single flat data by using propositionalization method or it uses improved version of the existing methods known as Upgrading method to handle multiple database relations [1], [2]. The earlier method has numerous drawbacks. In this paper, a new method called Multi Relational model is constructed for mutagenesis prediction based on Fuzzy decision Tree (MRFT). The proposed system is an improvement of fuzzy tree for treating multi relational data based on [3] and tuple- id propagation [4]. Fuzzy decision tree is an enhancement of classical decision tree based on fuzzy set theory [5], [6]. The fuzzy decision tree is an effective method for extracting knowledge presents in uncertain classification problems.

A. Research Contributions

We propose a method for generating Multi relational model for predicting mutagenesis data with the help of fuzzy tree. The research contributions are as follows

- We improve the fuzzy decision tree for predicting mutagenesis multi relational data.
- The architecture includes an efficient method known as Correlation based Feature Selection method to reduce the dimension search space size.
- To improve the predictability of the classifier, K-nearest neighbor method is used to impute the missing values in the data. The rest of the paper is structured as follows. Section 2 provides a summary of related works. Classification using multi relational fuzzy decision tree is discussed in section 3. Section 4 illustrates the experimental results of MRFT. Finally, section 5 concludes the paper.

II. RELATED WORK

Fuzzy decision tree is applied in various areas for different prediction problems. In [7], fuzzy decision tree has been applied for estimating recurrence of cancer disease and predicting survival of patient. This experiment was carried out using SEER breast cancer data set and the results were compared with Decision tree. From the performance results, it was found that, for cancer prognosis, hybrid fuzzy

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

decision tree classification was more robust and balanced than independently applied crisp classification algorithm. The fuzzy decision tree for analysis of gene expression data has been proposed in [8]. The proposed algorithm was compared with Radial basis function, Naïve Bayes classifier etc. The study was performed on five different data sets and outcomes show that fuzzy tree outperforms the classical algorithms on these data sets. The protein model prediction based on extended fuzzy decision tree with spatial neighborhood features has been proposed in [9] and it achieves 90% accuracy on training data. A novel fuzzy tree for predicting protein active sites was proposed in [10]. Later, the proposed model was used for determining the functions of the protein molecules.

III. PROPOSED SYSTEM

The proposed system is shown in the Fig. 1. This system, first accepts the raw mutagenesis data from the user. And then it joins the target relation in the multi relational database with non-target relations by using class label propagation module. This module transmits the class label value from the target relation to non-target relations in the database with the help of foreign key links.

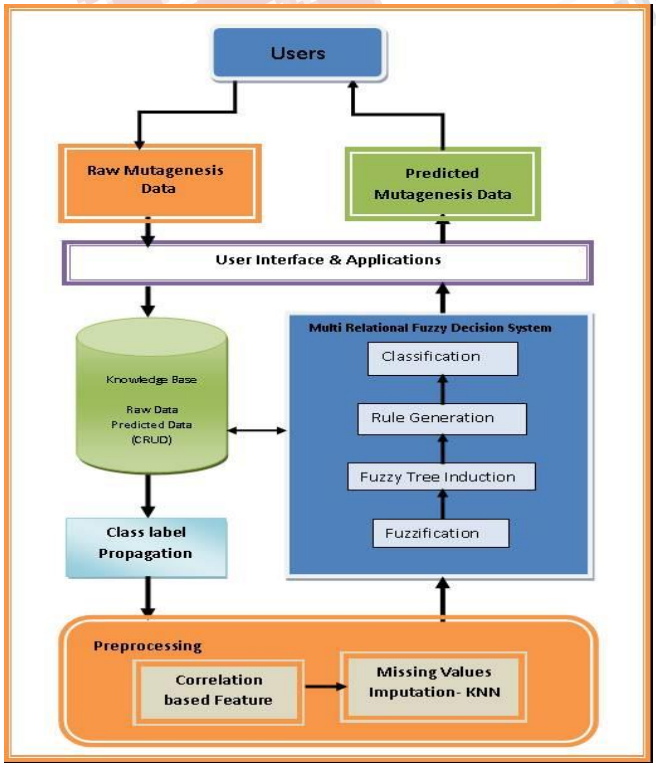


Fig 1: Framework of MRFT System

Next, the data processing module is used to refine the data by eliminating noises and inconsistencies in the dataset. By using correlation feature selection method [11], it chooses the most essential attributes in the dataset and removes the immaterial and redundant data. Next, the missing values are filled with K-nearest neighbor method (KNN) [12]. Next, this reduced and complete data set is given as an input to the fuzzy tree phase. The multi relational fuzzy decision tree induction process consists of the following modules:

A. Multi relational Fuzzy tree induction

Before inducing the tree, the numerical values in the dataset are fuzzified into linguistic terms by using triangular membership function [13]. This is process is known as fuzzification. After fuzzification, the algorithm for the multi relational fuzzy tree shown in the Fig. 2 is applied for predicting mutagenesis data. It generates the fuzzy tree with the help of class label propagation element. Next, the rules are extracted from this tree by tracing the path from root to leaf which is in the form of If- Then rules. Then it used to classify the previously unknown data

Algorithm: MRFT (D, R_t)
Input : A multi relational database S with a target Relation R_t. Each tuple has p continuous attributes A_1 , A_2 ,.....A_p.
Output : A fuzzy decision tree *F* for predicting class labels of target tuples.
Parameters: S, D - Database, C- Class label, – attribute F_{i1}, F_{i2}, F_{im}- fuzzy sets for attribute Ai that has m different values, D^k - fuzzy subset in D whose class is C k |D| - the sum of the membership values in D, θ_r , θn - threshold, N, N1N2 – tree nodes, t- tuples in S - Fuzzy membership value for tuples t.
Procedure :
1. Set N → Set of tuples **t** S with = 1 // Create a Root node that has a set of fuzzy data with membership value 1
If |Rt| < MIN_SUP then return
2. If node N with a fuzzy set of data D satisfies the following conditions then it is a leaf node and assigned by the class name
a. If |R| < θn or // The number of a data set is less than θn
b. If $|_D^{CK}| / |_D| > \theta_{er}$ // The proportion of a data set of a class $C_K >= \theta r$
c. There is no A D for more classifications then return n as leaf
// If it does not satisfy the above conditions, it is not a leaf node, and the new sub node is generated as follows:
3. Evaluate all or R linked with via foreign Keys based on information gain
A_{max} = Attribute with max. Information gain
If info-gain(A_{max})<MIN-INFO-GAIN then return
Set Relation of A_{max} to active
Divide D into fuzzy subsets D_1, D_2 .. Dm according to the feature A_{max}
Dj.T.MF = MF.D * $F_{max,j}$ of A in D // The membership values of tuples in Dj is the product of membership values in D and fuzzy value $F_{max,j}$ of in D
Generate N1,N2..N_m for fuzzy subsets D_1, D2 ..., Dm
Label the fuzzy sets $F_{max,j}$ to edges that connect between N_j and N.
Replace D = D_j where j=1, 2,...m and repeat from step 2 recursively until all paths are leaf node
For each relation R D that is set active
Set R to inactive
Return N

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

IV. PERFORMANCE ANALYSIS

.The proposed system is implemented with WEKA tool [14] based on JAVA. The experiments are executed on Intel Core i5 2.67 GHz with 4 GB RAM, running Windows 7. The estimation method used for this study is 10-fold cross validation method. The experiments are conducted on Mutagenesis dataset and the results are compared with the existing system CrossMine [15]. The parameters used for this analysis is shown in Table 1.

Table 1. Parameter used

Name	Value
MAX_NUM_NEGATIVE	600
MIN_FOIL_GAIN	2.50
MAX_RULE_LENGTH	6
NEG_POS_RATIO	1.0
MIN_INFO_GAIN	0.05
MIN_SUP	10
MIN-NUM-FUZZY_ SET	3

A. Data Set

Mutagenesis is a benchmark data set that consists of the structural details about 188 Regression friendly and 42 Regression unfriendly molecules [16]. These molecules are to be classified as mutagenic or not. The total number of tuples in this dataset is 15,281. The data set includes two learning problems that are named as task 1 and task2. The task1 is to categorize 188 Regression friendly molecules. Out of 188 tuples, 125 tuples are positive and 63 are negative. The task2 is to classify all the Regression molecules apart from Regression Friendly or Regression unfriendly. It contains 154 positive tuples and 76 negative tuples. The dataset also contain background relations such as atom, bond.

A. Evaluation Measures

The following measures have been chosen in order to evaluate the study

- Accuracy: The percentage of correctly classified tuples given by

$$Accuracy = \frac{No.of\ tuples\ correctly\ classified}{Total\ no.of\ tuples} \quad (1)$$

- Sensitivity: True positive rate given by

$$Sensitivity = \frac{No.of\ tuples\ that\ are\ truly\ Positives}{No.of\ True\ Positives\ t + No.of\ False\ Negatives} \quad (2)$$

- Specificity - True negative rate given by

$$Specificity = \frac{No.of\ True\ Negative\ tuples}{No.of\ True\ Negatives + No.of\ False\ Positive} \quad (3)$$

- Rule set : Number of rules generated

Runtime: Induction time of the classifier in seconds

C. Experimental Results

The first experiment was conducted to measure the accuracy of the two classifiers on mutagenesis dataset and the results are shown in the Table 2. As seen by the table, the proposed system achieves highest accuracy for both classification tasks compared to CrossMine.

Table 2.Accuracy of the Classifier

Dataset	MRFT	CrossMine
Task1	90.02%	87.2 %
Task2	92.34%	84.43%

The Table 3 shows the sensitivity and specificity values of the classifier. From this table, one can understand that MRFT is highly efficient in terms of sensitivity and specificity compared to CrossMine.

Table 3.Sensitivity and Specificity

Dataset	MRFT	CrossMine
Sensitivity		
Task1	80.78	72.2
Task 2	82.56	70.09
Specificity		
Task1	82.11	68.89
Task2	81.45	83,77

The Fig.3 shows the number of rules generated by the classifier when 10-fold cross validation is used for mutagenesis dataset. This figure indicates that the proposed system generates less number of rules than CrossMine for task1 and task2.

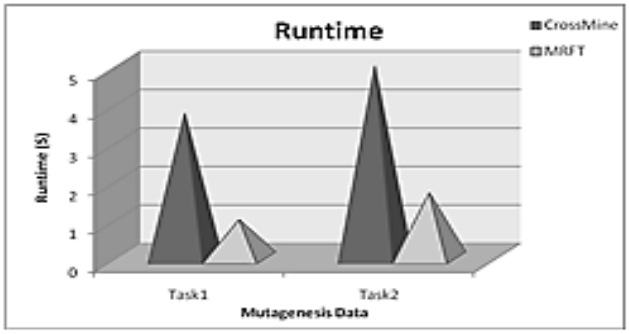
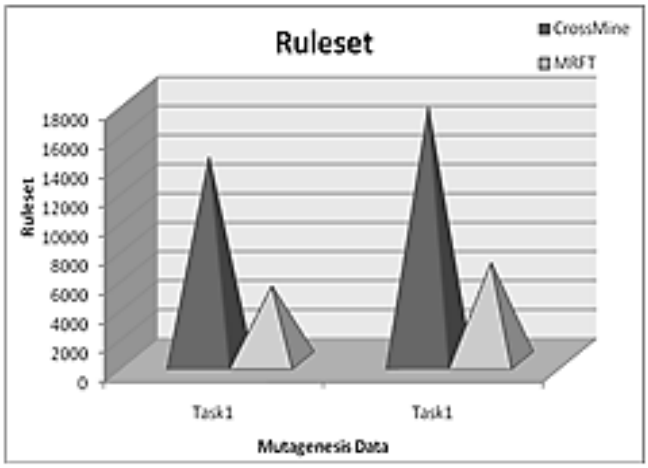


Fig 3: Rule Set generation

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

Runtime of the classifier is considered to be one of the important properties of machine learning algorithm. From the Fig. 4 it can be observed that the proposed system runs faster than the earlier system for the mutagenesis prediction.



V. CONCLUSION

This paper introduced a novel method for classifying the multi relational data using upgrading approach. The method is based on fuzzy decision tree which has been applied in various applications. In this work, fuzzy tree is used for predicting multi relational mutagenesis data. The Proposed MRFT achieved high accuracy and efficiency in term of sensitivity, specificity, run time and rule generation when compared with well-known multi-relational learning algorithm.

REFERENCES

[1] W. Emde, and D. Wettschereck ,":Relational Instance – Based Learning", In Saitta, L., editor, Proc. of 13th Intl. Conf. on Machine Learning, Morgan Kaufmann, 122-130, 1996.

[2] J. Neville, D. Jensen, L. Friedland and M. Hay, Learning Relational Probability Trees. Technical Report (02-25), Dept. of Computer Science, University of Massachusetts Amherst, 2002, Revised version February 2003.

[3] M. Thangaraj and C.R. Vijayalakshmi, "An efficient multi relational framework using fuzzy rule based classification technique", Intl. J. of Data Mining, Data Modeling and Management. 8(4):348-368, 2016. Print ISSN: 1759-1163.

[4] X. Yin, J. Han, and J. Yang,"Efficient Multi relational Classification by Tuple-id Propagation", Proc. of KDD workshop on MRDM, 2003.

[5] C.Z.Janikow, "Fuzzy decision trees: issues and methods.", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. Vol.28, 1, 1-14, 1998.

[6] M. Umano, H, Okamoto, I. Hatono, H. Tamura and J. Kinoshita,," Fuzzy Decision Trees by Fuzzy ID3 algorithm and Its Application to Diagnosis Systems" In Proc. of the third IEEE Conference on Fuzzy Systems (Orlando, FL , Jun 26-29,. IEEE, 2113 – 2118., 1994.

[7] M. U.Khan, J.P.Choi, H.Shin and M. Kim," Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008.

[8] A.Simone , L. D.Jakobovic and S. Picek, "Analyzing Gene Expression Data: Fuzzy Decision Tree Algorithm applied to the Classification of Cancer Data",

[9] A. Chida, R. Harrison and Y. Zhang, "Protein Model assessment using extended fuzzy decision tree with spatial neighborhood features", Proc. Of CIBCB, San diego, CA, USA, 2012.

[10] G. Mirceva, A. Naumoski and D. Davcev, " A novel fuzzy decision tree based method for detecting protein active sites, ICT Innovations AINSC, Vol.150,pp 51-60, 2011.

[11] He, J., Liu, H. and Hu, B. (2010) ' Selecting effective features and relations for efficient multi-relational classification', Computational Intelligence, Vol.26, 258-281.

[12] B. Zhu, C. He, and P. Liatsis,," A robust missing value imputation method for noisy data", Applied Intelligence, Vol.36 (1), 61-74., 2012. doi: 10.1007/s10489-010-0244-1.

[13] Y. Yuan, and M.J.Shaw, " Induction of fuzzy decision trees", J. Fuzzy Sets and Systems, Vol 69, 2, 125-139, 1997.

[14] H.W. Ian, and E. Frank, Data Mining: Practical

Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers,2000.

[15] X. Yin, J. Han, and P.S.Yu, " Efficient Classification across Multiple Database Relations: A CrossMine Approach", J. IEEE Transactions on Knowledge and Data Engineering, 18, 6(June. 2006), 770-78, 2006.

[16] A. Srinivasan, SH. Muggleton, MJE. Sternberg and RD.King ,'Theories for Mutagenicity: a study in first- order and feature-based induction", Artificial Intelligence, Vol.85(1-2):277-299,1996.