World Applied Sciences Journal 28 (9): 1331-1335, 2013
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2013.28.09.1502

Fuzzy Rule Based Classification for Heart Dataset using Fuzzy Decision Tree Algorithm based on Fuzzy RDBMS

[1]*Idris Mala,* [2]*Pervez Akhtar,* [3]*Tariq Javid Ali and* [4]*Syed Saood Zia*

[1]Usman Institute of Technology, Hamdard University, Karachi, Pakistan
[2]PNEC, National University of Science and Technology, Karachi, Pakistan
[3]HITEC University, Taxila Cantt., Pakistan
[4]Sir Syed University of Engineering and Technology, Karachi, Pakistan

Abstract: A fuzzy rule-based system design concentrates on accuracy and interpretability of the system. Fuzzy decision tree method is proposed based on fuzzy RDBMS and rule generation based on C4.5 algorithm known as fuzzy rule generation system (FRGS) algorithm. A fuzzy decision tree is developed by first converting a medical application of heart relational database to fuzzy heart relational database and then developing the decision tree using C4.5 algorithm. In the next step rule generation is performed from the C4.5 algorithm. A pruning process is performed to prevent overfitting. Different pruning rates are analyzed to show the variation of interpretability in the generated model. Results show that good tradeoff between accuracy and interpretability can be made by varying pruning.

Key words: Fuzzy decision tree · FRGS algorithm · C4.5 algorithm · Decision tree pruning

INTRODUCTION

Data mining popular tasks are classifications, clustering and predictions. Classification process comprises of supervised learning where class level or classification is known. Well known techniques for classifications are Decision Tree, Fuzzy Logic, Bayesian, Neural Network, Rough Set Theory and Genetic Algorithms [1]. Decision trees are widely used due to its understandability of structure to create rules, resulting in decision making simple and easy. The interpretability of generated models is improved due to its selection of features by using most relevant ones [2].

A well-known decision tree proposed by Quinlan [3] is C4.5, where the selection and partitioning of feature is based on information gain and entropy measures. Pruning process of the decision tree of C4.5 algorithm is pruned once it is completely induced and is based on estimation of true error [4]. Different approaches to decision tree have been proposed. Concept of adding fuzzy to C4.5 algorithm was proposed [5]. The addition of fuzzy to C4.5 algorithm resulted in better results in terms of accuracy and interpretability.

In this paper, we propose a new method to construct fuzzy decision trees from relational database system and then generate fuzzy rules from the fuzzy decision tree for knowledge base called the fuzzy rule generation system FRGS algorithm. This algorithm generates knowledge from relational database system. Once developing fuzzy decision tree, using FRGS algorithm, it has been shown that varying pruning in the tree results in the variation in the accuracy and interpretability of fuzzy generated model. This FRGS algorithm keeps a good tradeoff between accuracy and interpretability of the system.

Section 2 of the paper briefly describes the decision tree C4.5 algorithm and PART used in WEKA open source software. Section 3 presents our proposed algorithm of Fuzzy Decision Tree method based on FRGS algorithm, section 4 presents conversion of relational database to fuzzy relational database, section 5 presents experimental results and the last section 6 make an evaluation of the work and suggest some future direction.

DECISION TREES

Decision tree algorithm was developed by J. Ross Quinlan, in late 1970s and early 1980s and was called ID3

Corresponding Author: Idris Mala, Usman Institute of Technology, Hamdard University, Karachi, Pakistan.

(Iterative Dichotomiser). Quinlan later presented C4.5 (a successor of ID3), a supervised learning algorithm. A generation of binary decision tree was published in book of Classification and Regression Trees (CART) in 1984 by Breiman, Friedman, Olshen and Stone.

Here we present C4.5 decision tree method as well as PART (an algorithm to generate rules from C4.5).

a) C4.5 Decision Tree: C4.5 algorithm is a well know decision tree induction, of the basic method of its predecessor, which is ID3 algorithm. Basic decision tree algorithm is a greedy algorithm constructing decision trees in a top-down recursive divide-and-conquer manner. The algorithm generates the tree structure through partitioning until whole decision space is completely divided. Another important characteristic of decision trees besides generating the tree structure is the selection of features for the partitioning process. Only relevant features are used in the tree, which improves the time of classification as well as interpretability of the model.

Overfitting can be avoided by preventing some subsets of training examples from being subdivided. It can also be done by removal of some part of the structure of decision tree after being generated. Author of C4.5 prefers latter method as it allows potential interactions among attributed explored in order to make decision about decision of results [6].

The algorithm calculates information gain and entropy measures on importance of features, making it possible to select features. C4.5 uses post-pruning, which is used to assess error rates of the tree and its components directly on the set of training examples after the tree is completed [6].

Process of pruning is understandable by assuming N training cases covered by a leaf, E of them incorrectly. Error rate of leaf is defined by E=N. If N training cases are considered as sample, it gives the probability of error over the entire population of examples covered by leaf. The probability of distribution is usually calculated by pair of confidence limits.

The default confidence limit used in C4.5 is 25%. A chance of pruning is higher for smaller confidence limit and a chance of pruning is smaller for higher confidence limit. Setting the confidence limit to 100%, is that the predicted error, obtained with the examples at hand, is equal to the real error and no pruning will be performed. The idea conflicts with the natural intuition one might have that a 25% confidence limit will produce less pruning than an 80% confidence limit. This way, one should not associate the default 25% confidence limits of C4.5 with a 25% pruning rate [2].

b) Part: PART relates to algorithm based on partial decision tree [7]. It is a rule-induction procedure that adopts the separate-and-conquer strategy. It creates rules recursively by building rules and then removing instances until none is left. It uses pruned tree to extract rules, using separate-and-conquer rule learner. To accelerate process PART develops partial decision tree instead of full decision tree. A single rule is selected from the subtree, once the tree induction ceases [8]. PART is used to develop rules from C4.5 used in software package WEKA. WEKA comprise of different classification problems in data mining, where C4.5 decision tree is developed by J48 classifier and PART classifier is used to develop rule generation from C4.5 algorithm.

FUZZY DECISION TREES

We propose a new method to construct fuzzy decision trees from relational database system and then generate fuzzy rules from the fuzzy decision tree for knowledge base. The method is named as fuzzy rule generation system (FRGS) algorithm which generates knowledge from relational database system.

Our proposed fuzzy decision tree based on FRGS algorithm is as follows:

i) Define the fuzzy data base from relational database using appropriate mapping (FRDBMS process).
ii) Develop the decision tree using C4.5 algorithm.
iii) Generate rules of C4.5 algorithm using PART at default 25 % confidence limit.
iv) Apply pruning decision tree to optimize the computation efficiency i.e. tradeoff between accuracy and interpretability of the system.

The algorithm is based on fuzzy relational database which is generated in first step, by converting the relational database into fuzzy relational database by mapping with the membership function of fuzzy sets that are defined. In the next step, C4.5 algorithm is applied by calculating entropy measures and information gain in developing the fuzzy decision tree.

Rules are generated from the developed fuzzy decision tree based on C4.5 algorithm at 25 % confidence limit which is taken as default and in the last stage pruning at different levels are applied to optimize results of getting a better tradeoff between accuracy and interpretability of the fuzzy generated model.

FUZZY DATABASE FROM RELATIONAL DATABASE

Fuzzy rule based classification systems, are based on the fuzzy set and fuzzy logic theories proposed by Loft A.

Zadeh [9]. The relational database model is very popular data model of database systems used in commercial applications as it can be very easily understood and implemented. The combination of fuzzy logic and relational database model, results in fuzzy relational database management system, was proposed in [10], called Fuzzy DB System model.

The Fuzzy DB System model, defined in [10] was a fuzzy software layer for conversion used for converting fuzzy query to standard query which was then executed by standard relational database management system (RDBMS). The results from the RDBMS where again converted back to fuzzy linguistic variable term results more human understandable manner.

The module required in case fuzzy decision tree is for conversion of relational database to fuzzy database. This conversion only take place by help of some fuzzy metadata definition defining the fuzzy term or fuzzy membership functions. Mapping from numerical values to fuzzy linguistic variable then take place based on the fuzzy metadata definition. Implementation work has shown the algorithm for conversion of this methodology.

RESULTS AND DISCUSSION

For the implementation of this model in medical informatics, heart dataset is being used of UCI Repository [11]. Practical implementation of the model was developed in java using standard RDBMS. Some basic results were also shown in the previous work in our research paper [10]. The algorithm has been modified in this work to convert any relational database into fuzzy relational database.

The above shown is the base heart database of UCI Repository in figure 1. This table is converted to fuzzy relational database by using a reference table based on fuzzy membership definition of each field, as shown above, in which there are four input fields and one output or decision field.

For Meta knowledge base of UCI repository of heart data, we have considered membership function definition taken up by Ali at el. [12] for heart ailment database where different membership functions are defined for the four attributes selected on the basis of medical practitioner recommendation.

A Meta knowledge Base is defined which is called the fuzzy linguistic data definition and is implemented using a table named as fuzzysql table, which comprises of all the attributes necessary to define a catalog or dictionary to extend RDBMS which stores are necessary information to describe and manipulate fuzzy RDBMS query.

fuzzyheart | heart

age	bp	ch	hr	heart_disea
70	130	322	109	
67	115	564	160	1
57	124	261	141	2
64	128	263	105	1
74	120	269	121	1
65	120	177	140	1
56	130	256	142	2
59	110	239	142	2
60	140	293	170	2
63	150	407	154	2
59	135	234	161	1
53	142	226	111	1
44	140	235	180	1
61	134	234	145	2

Fig. 1: Heart database of UCI Repository

fuzzyheart | heart | fuzzysql

OBJECT	Field	FLV	LOW	HIGH
heart	bp	Low	0	133
heart	bp	Medium	127	153
heart	bp	High	142	172
heart	bp	VeryHigh	154	999
heart	hr	Low	0	141
heart	hr	Medium	111	194
heart	hr	High	152	999
heart	ch	Low	0	197
heart	ch	Medium	188	250
heart	ch	High	217	307
heart	ch	VeryHigh	281	999
heart	age	Young	0	38
heart	age	Mild	33	45
heart	age	Old	40	58
heart	age	VeryOld	52	999

Fig. 2: Fuzzy Metadata definition

In the first step of the FRGS algorithm, the heart dataset as shown in fig. 1 is converted to fuzzy heart dataset by using the fuzzy metadata definition as shown in figure 2.

The conversion algorithm of step 1 of FRGS algorithm is as follows:

INPUT : Heart Dataset
Fuzzy Metadata Definition
OUTPUT : Fuzzy Heart Dataset
Begin Load driver of jdbc: odbc
Make multiple connections
Execute query to get data of heart dataset
Get result in a result set 1
Execute query to get fuzzy metadata definition
Get result in a result set 2
Clear all records from fuzzy heart dataset
Loop until end of the result set 1
Get numeric data into variables
Get each variable
Run query of each variable to get its fuzzy linguistic variable
Store in a variable (range of values)

Store in as new record in fuzzy heart dataset
Next line
End loop
Close all connections
Close database
End

Running the algorithm in java, converts the relational heart dataset into fuzzy heart dataset using the mapping fuzzy metadata definition. The results of heart dataset are shown in figure 3.

The table generated by the algorithm in java, is then used to generate the decision tree using C4.5 algorithm in Weka by J48 classification. Pruning is used in J48 classification and the results are shown in figure 4.

The resultant fuzzy decision tree from the FRGS algorithm is shown in figure 5.

Next we applied the PART in Weka to generate the rules for the fuzzy heart dataset. At the default value of 25% of confidence limits, the results of fuzzy rule generation are shown in figure 6.

The results generated 10 rules with 25 % default confidence limit. The same algorithm was checked for heart dataset before converting it to fuzzy heart dataset and it was found that the rule generated by the same algorithm (PART classification) is 4 rules, so the accuracy is increased by 2.5 times. The results of fuzzy heart dataset was also checked by varying confidence limit. A reduction of confidence limit to 0.1 resulted in only 5 rules being generated i.e. decrese in accuracy with decreasing the confidence limit and if the confidence limit is increased to 0.5, then the rules generated are 14. This results in increase in accuracy and complexity of the system with increase in the number of rules.

fuzzyheart

age	bp	ch	hr	hd
Old	Low	Low	Medium	No
VeryOld	Low	Medium	Low	Yes
Old	Medium	Medium	Medium	No
Old	Low	Low	Low	Yes
Mild	Low	Low	Medium	Yes
VeryOld	High	Medium	Low	Yes
VeryOld	Medium	Medium	Low	Yes
Mild	Medium	VeryHigh	Low	Yes
Old	Low	Medium	Medium	No
VeryOld	Medium	VeryHigh	Medium	No
VeryOld	Medium	Low	Medium	No
Mild	Low	Medium	Medium	No
Mild	Low	Medium	Medium	No
VeryOld	Medium	Low	Low	Yes
VeryOld	Low	Low	Low	No
VeryOld	VeryHigh	Medium	Medium	Yes
VeryOld	Low	High	Low	Yes
Old	Medium	Low	Medium	No
Old	Low	High	Medium	Yes
Mild	Low	Medium	Medium	No
VeryOld	Medium	Medium	Medium	No
VeryOld	Medium	Medium	Medium	No
Old	Medium	High	Low	Yes

Fig. 3: Resultant Fuzzy Heart Dataset

```
J48 pruned tree
------------------

hr = Low: Yes (84.0/24.0)
hr = Medium
|   age = VeryOld
|   |   ch = VeryHigh: No (9.0/3.0)
|   |   ch = High: Yes (23.0/8.0)
|   |   ch = Low: Yes (4.0/1.0)
|   |   ch = Medium: No (18.0/7.0)
|   age = Old: No (84.0/24.0)
|   age = Mild: No (39.0/5.0)
|   age = Young: No (7.0/2.0)
hr = High: Yes (2.0/1.0)

Number of Leaves  :     9

Size of the tree :      12


Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         180               66.6667 %
Incorrectly Classified Instances        90               33.3333 %
```

Fig. 4: Decision tree of C4.5 algorithm using J48 Classification

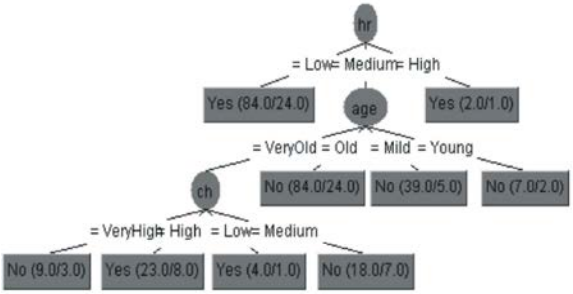

Fig. 5: Fuzzy decision tree from the FRGS algorithm

```
hr = Low AND
ch = High: Yes (30.0/6.0)

hr = Medium AND
age = Old: No (84.0/24.0)

hr = Low AND
ch = Medium: Yes (30.0/11.0)

age = Mild AND
hr = Medium: No (39.0/5.0)

age = VeryOld AND
ch = High: Yes (23.0/8.0)

age = VeryOld AND
bp = Medium: No (18.0/7.0)

age = Old: Yes (9.0/1.0)

hr = Low AND
ch = VeryHigh: Yes (6.0)

ch = Medium AND
bp = Low: No (13.0/5.0)

: No (18.0/7.0)

Number of Rules  :       10
```

Fig. 6: Fuzzy rule generation from C4.5 algorithm

Therefore increasing the confidence in training data increases the tree size, which increases the accuracy of the system but decreases the interpretability of the system. Reversing the process by decreasing the confidence, decreases the tree size, reduces the accuracy and increases the interpretability of the system. The interpretability of fuzzy rules is an important consideration for fuzzy model generation. Here, we have used number of rules as indicator to ascertain the accuracy versus the interpretability of fuzzy systems. We have shown that increase in fuzzy rules is based on decreasing the pruning (by increasing confidence limit) which results increases accuracy and decreases interpretability. Similarly, decrease in fuzzy rules is based on increasing the pruning (by decreasing the confidence limit) which results in decreasing the accuracy and increasing the interpretability.

CONCLUSION

A new method of fuzzy decision tree and fuzzy rule generation called the FRGS algorithm has been proposed and its application in medical domain has been shown in heart dataset. A complete process of the FRGS algorithm has been shown related to heart dataset. A fuzzy decision tree was developed by first converting the heart relational database to fuzzy heart database and then developing the tree using C4.5 algorithm (J48 classification in Weka). In the next step fuzzy rules were generated using PART classification by the same C4.5 algorithm and pruning decision tree was used to show how accuracy and interpretability varies with increasing and decreasing of confidence limit. The experimental results also verified that number of rules in fuzzy rules based system can be specified by user based on confidence limits which can result in a good tradeoff between accuracy and interpretability. As future work we intend to experiment and investigate more medical datasets and other methodologies for comparison of multi-objective problems of accuracy and interpretability.

REFERENCES

1. H. Jantan, A. R. Hamdan and Z. A. Othman, 2010. "Human Talent Prediction in HRM using C4.5Classification Algorithm", International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 08: 2526-2534.
2. Cintra, M.E., M.C. Monard and H.A. Camargo, 2010. "Evaluation of the Pruning Impact on Fuzzy C4.5", Anais Congresso Brasileiro de Sistemas Fuzzy (CBSF).
3. Quinlan, J.R., 1988. "C4.5 Programs for Machine Learning". Morgan Kaufmann, CA.
4. Kohavi, R. and R. Quinlan, 1999. "Decision tree discovery". In Handbook of Data Mining and Knowledge Discovery, University Press, pp: 267-276.
5. Kazunor, H., U. motohide, S. Hiroshi and U. Yuushi, 1999. "Fuzzy C4.5 for generating fuzzy decision trees and its improvement". Faji Shisutemu Shinpojiumu Koen Ronbunshu, 15: 515-518.
6. Quinlan, J.R., 1993. "C4.5: Programs for Machine Learning". Morgan Kaufmann Series in Machine Learning, 1st ed. Morgan Kaufmann.
7. Frank, E. and I.H. Witten, 1998. "Generating accurate rule sets without global optimization," in ICML '98: Proceedings of the 15th Int. Conf. on Machine Learning. Morgan Kaufmann, pp: 144-151.
8. Cintra, M.E., M.C. Monard and H.A. Camargo, 2011. "Fuzzy and Classic Rule Learning Methods: a Comparative Analysis", Proceedings of the World Conference on Soft Computing (WCSC 2011) San Francisco State University, pp: 182-190.
9. Zadeh, L., 1965. "Fuzzy sets", Information and Control, 8: 338-353.
10. Mala, I., P. Akhtar, S. Zia and S. Mirza, 2011. "Application of Fuzzy Relational Databases in Medical Informatics", Proceedings of the 14th IEEE Multi-topic conference (INMIC), pp: 41-44.
11. Robert Detrano and M.D. and PhD, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Available: www.archive.ics.uci.edu/ml/dataset/Heart+ Disease.
12. Ali, A. and M. Mehdi, 2010. "A Fuzzy Expert System for Heart Disease Diagnosis", Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS.