

A fuzzy decision tree approach to start a genetic algorithm for data classification

R. P. Espíndola & N. F. F. Ebecken

COPPE/Federal University of Rio de Janeiro, Brazil

Abstract

This paper introduces a fuzzy decision tree to initiate the first population of a genetic algorithm to perform data classification. On large datasets, the evolutive process tends to waste computational resources until some good individual is found. It is expected that the use of a fuzzy decision tree can significantly reduce this feature. The genetic algorithm aims to obtain small fuzzy classifiers by means of optimization of fuzzy rules bases. It is shown how a fuzzy rules base is generated from a numerical database and how its best subset is found by the genetic algorithm. The classifiers are evaluated in terms of accuracy, cardinality and number of features employed. The results obtained are compared with a known study in the literature and with an academic decision tree tool. The method was able to produce small fuzzy classifiers with very good performance.

Keywords: classification, feature selection, fuzzy systems, genetic algorithms, fuzzy decision tree.

1 Introduction

One of the major drawbacks of a genetic algorithm is the high computational costs on performing its search. When dealing with large datasets, this feature is a key aspect to be considered. In this work, feature selection [1] and classification [2] are performed by a fuzzy genetic system. Fuzzy rules are generated automatically from the datasets and a genetic algorithm is applied to find the shortest and most accurate subset of rules. As each rule employs only one feature, the final subset possibly uses few features. In this model of rules, the classification is done along with a value which estimates the relationship between the condition and the class, defined by the concept of fuzzy subsethood.



According to experimental results, the best fuzzy classifiers found by the genetic algorithm were formed by less than 10% of the rule bases sizes. Based on this fact it is intended to speed up the evolutive process by generating the initial population with good candidate solutions – short subsets of rules with good accuracy – by means of fuzzy decision trees [3]. This subject has already been addressed quite successfully by reducing the inclusion probability of a rule [4]. The focus of this study is to minimize the random dependence of the genetic algorithm initiation and verify if this strategy brings some improvement on the system performance.

This fuzzy genetic system was inspired by the work of Ishibuchi *et al* [5] which employed a genetic algorithm to obtain a fuzzy classification system using Mamdani's model of rules. This method suffers from combinatorial explosion of rules when applied to other than very simple problems. Espíndola & Ebecken [6] applied the same genetic algorithm to optimize decomposed zero-order Takagi-Sugeno-Kang (TSK) fuzzy rules bases. This kind of rule uses only one feature on the antecedent part and concludes about the class of a new pattern. The gains in processing time, simplicity of implementation and comprehension are notable. Espíndola & Ebecken [7] improved the fuzzy genetic system in order to perform feature selection as well.

The strategy of induction of fuzzy decision trees employed was based on the reduction of classification ambiguity with fuzzy evidence and was presented by Yuan & Shaw [8]. Given a problem, the induction process is not performed to generate the best possible fuzzy tree but a suitable one, small and with good accuracy. It was expected that the conversion of the tree into TSK rules yields a good candidate solution for the genetic algorithm. So this individual is used to generate the remaining ones by randomly mutating some alleles.

To assess this methodology some datasets from UCI Machine Learning Repository were studied along with a large dataset of the fog around the International Airport of Rio de Janeiro. The results are compared to others obtained from a decision tree tool applied to the same problems.

In the next section, the fuzzy genetic system is shown. In section 3, the use of the fuzzy decision tree is detailed. In section 4, the experiments realized are presented and commented. In the last section, final comments and future researches are exposed.

2 The fuzzy genetic system

2.1 Rule base generation

The process of rule generation was presented in Evsukoff *et al* [9] which applied the decomposition scheme proposed by Kosko [10] to zero-order TSK fuzzy rules. In this work, each feature space was normalized and divided in five partitions defined by triangular membership functions associated to the linguistic labels small, medium small, medium, medium large and large.



Their definitions are:

$$\mu_{\text{small}}(x) = \max\{0, 1 - 4 \cdot |x|\} \quad (1)$$

$$\mu_{\text{medium_small}}(x) = \max\{0, 1 - 4 \cdot |x - 0.25|\} \quad (2)$$

$$\mu_{\text{medium}}(x) = \max\{0, 1 - 4 \cdot |x - 0.5|\} \quad (3)$$

$$\mu_{\text{medium_large}}(x) = \max\{0, 1 - 4 \cdot |x - 0.75|\} \quad (4)$$

$$\mu_{\text{large}}(x) = \max\{0, 1 - 4 \cdot |x - 1|\} \quad (5)$$

The rules are constructed in such a way that they inform, besides the class, an output value defined by the fuzzy subethood. Given training patterns $x^m = (x_1^m, \dots, x_n^m)$ with classes $y^m = (y_1^m, \dots, y_K^m)$, the degree to which the set of the antecedent $X_{i,j}$ is subset of the set of class k is:

$$\phi_{kj}^i = \frac{\sum_{m=1}^M \mu_{X_{ij}}(x_i^m) \cdot y_k^m}{\sum_{m=1}^M \mu_{X_{ij}}(x_i^m)} \quad (6)$$

Thus, the rules have the following structure:

$$\text{Rule } R_{k,j}^i: \text{ If } x_i \text{ is } X_{i,j} \text{ then class} = k \text{ with output value } \pi_{k,j}^i = \phi_{k,j}^i \quad (7)$$

in which $k = 1, \dots, K$ and $j = 1, \dots, 5$.

Considering a database with n attributes and K possible classes, a rule base generated in this way has $5 \cdot K \cdot n$ elements.

2.2 Classification of new patterns

Given a pattern $(x_1, \dots, x_n)$, to determine its class requires the execution of the following steps:

1. for each feature i , combine the outputs of those rules related to the same class k

$$\pi_k^i = \frac{\sum_{j=1}^5 \phi_{kj}^i \cdot \mu_{X_{i,j}}(x_i)}{\sum_{j=1}^5 \mu_{X_{i,j}}(x_i)} = \sum_{j=1}^5 \phi_{kj}^i \cdot \mu_{X_{i,j}}(x_i), \text{ in which } k = 1, \dots, K. \quad (8)$$

2. for each class, aggregate the previous combined non-zero outputs by the following function:

$$\pi_k = \min\{\pi_k^i / i = 1, \dots, n\} \quad (9)$$

3. the class with the highest value π_k is associated to the pattern.



2.3 The genetic algorithm

In many problems, to use the entire rule base on the classification of a new pattern can not be appropriate due to, among other reasons, an excessive delay to the pattern evaluation and the rule base may not produce the correct class due to the negative influence of rules with low confidences. The aim of this genetic algorithm is to select a small subset of rules with the greatest accuracy and which employs possibly few features.

Each rule subset is a candidate solution and is represented as a chromosome using the binary alphabet $\{0,1\}$. Each rule is represented by a gene in the chromosome. If the rule is present on the subset, its correspondent gene receives the allele 1. Otherwise, it receives the allele 0.

The amount of rules of a candidate solution S acts as a penalty factor on the fitness function, since small subsets are desirable. The same occurs with the amount of features employed. The ability of correctly classifying patterns from a database by S is more important than the previous penalty factors. Thus, that characteristic must have a distinct weight on the constitution of the fitness function defined by:

$$f(S) = W_{NCP} \cdot NCP(S) - W_C \cdot C(S) - W_F \cdot F(S) \quad (10)$$

in which:

- $NCP(S)$ is the amount of patterns correctly classified by S and W_{NCP} is its weight;
- $C(S)$ is the amount of rules – cardinality – of S and W_C is its weight;
- $F(S)$ is the amount of features employed by the rules of S and W_F is its weight.

Due to the higher importance of the accuracy, $0 < W_C, W_F \ll W_{NCP}$. In the experiments, W_{NCP} was set to 1000, W_S and W_F to 1, that is, the cardinality of a rule subset is as important as the amount of features employed.

As genetic operators, it was applied the uniform recombination with probability of 0.5 and the selection of the best chromosome pairs without restitution. In other words, all the individuals are selected in pairs on a decreasing fashion to perform recombination.

Considering the aim of reducing the amount of rules in each individual, two strategies of mutation were defined. The first one, $P_m(1 \rightarrow 0) = 0.1$, executes the change of the allele 1 to allele 0. The other one, $P_m(0 \rightarrow 1) = 0.0075$, changes the allele 0 to allele 1. Also, in each generation, elitism was applied in order to maintain the best individual found by the search process. An eventually null individual generated – a solution without rules – was replaced by one of its parents.

It was worked with a population of 20 individuals. When the random initiation strategy was chosen, at each locus the attribution probability of allele 1 during the generation of initial population was set to 0.25. After 500 generations the algorithm was stopped.



3 The use of fuzzy decision trees

3.1 Tree induction

The branching criterion of the tree algorithm selects the attribute with the smallest classification ambiguity. The stopping criterion produces a leaf when there is no attribute that reduces this value or when the truth level of classifying the objects within the branch into one class is above a given threshold. In the former case a null class is associated to the leaf. In the latter case, the class with the highest truth level is chosen.

The truth levels are measured by fuzzy subethood and the threshold was set to 0.5. This low value was chosen because the objective is not to produce the best fuzzy decision tree but a good one. The higher the threshold the slower is the induction process and the bigger are the trees generated. So 0.5 seemed to be a suitable value to deal with those restrictions.

3.2 Tree conversion into TSK rules

The heuristic of converting a fuzzy decision tree into TSK fuzzy rules is very simple: for each branch it produces a rule with each decision node and the class on the leaf. After all branches have been converted, the first individual of the genetic algorithm is generated by setting the genes associated to the rules with allele 1.

It is clear that this conversion is not a mathematical mapping between trees and rules systems but only a strategy of identifying some interesting relationships between attributes, its linguistics values and classes.

The remainder individuals are generated based on the first one, here called base individual: mutations with probability of occurrence of 0.05 are applied until the amount of rules of an individual surpass half of the amount of the base one. When this occurs the mutation probability is reduced to 0.0, that is, the remainder genes are copied from the base individual. The initiation strategy was done in this fashion to maintain a high similarity between the individuals of the population.

4 Experimental results and analysis

4.1 Experiments performed

In order to evaluate the performance of the fuzzy genetic system, seven datasets obtained in www.ics.uci.edu/~mllearn/MLRepository.html were studied besides a meteorological dataset. Table 1 shows the datasets, their dimensions and the amount of rules generated.

A decision tree tool was also applied to have its results compared. This tool is part of a system called Weka which is developed at the University of Waikato (www.cs.waikato.ac.nz/ml/weka). Weka contains an implementation of the very known decision tree algorithm C4.5 revision 8.



Before working with the datasets, some changes were made to allow the analysis of the method. Repeated records or records with incomplete information were eliminated and qualitative features were converted to discrete quantitative features. The scheme of testing employed was ten-fold cross-validation.

Table 1: Summary of datasets' characteristics.

Dataset	Valid features	Classes	Valid records	Rules generated
balance scale	4	3	625	60
car evaluation	6	4	1728	120
credit card approval	15	2	653	150
ionosphere	33	2	351	330
iris plant	4	3	150	60
meteorological dataset	18	7	26482	630
pima indian diabetes	8	2	768	80
wine recognition	13	3	178	195

4.2 Results analysis

Table 2 shows the average performances from decisions trees induced by C4.5 and the fuzzy ones for the studied problems. In terms of amount of rules/leaves, it was already expected that the fuzzy trees would be the smallest due to the low induction threshold. The same reason may be used to justify the generation of less accurate fuzzy trees. It is relevant to reaffirm that the objective is not to produce the best decision trees. They are only used to initiate the genetic algorithm.

The fuzzy trees with up to 5 leaves are the ones with a unique decision node. So the conversion of them into TSK rules produced equivalent classifiers because each leaf was identical to a rule. When the amount of leaves was higher, this equivalence does not occur and Table 3 shows the conversions performed.

As it can be observed, only on the meteorological dataset the accuracy was drastically reduced. On the other problems there is no significant variation. These results suggest that the bigger the tree the more dangerous is this strategy of conversion. To verify whether these individuals were better than those randomly generated, Table 4 presents the best ones from the latter scheme.

Comparing the accuracy by discarding differences up to 2%, the random initiation was able to generate better individuals than those obtained from the fuzzy trees except on credit card and iris dataset. Even then the superiority was not so great. It is relevant to notice that this comparison is not far: the best random individuals vs. those created by the fuzzy trees conversions. As previously explained (cf. section 3.2), other individuals are generated from the latter ones and Table 5 presents the best individuals obtained from the fuzzy decision tree initiation. The little superiority of random scheme in some datasets was reduced (balance dataset), eliminated (meteorological) or inverted (wine).



Table 2: Average decision trees generated.

Dataset	Fuzzy tree		C4.5 tree	
	Leaves	Accuracy (%)	Leaves	Accuracy (%)
balance	6.5	55.0	30.9	60.0
car	5.0	70.0	54.3	96.9
credit card	2.0	86.4	24.3	86.4
indian	5.0	69.9	20.1	75.1
ionosphere	5.0	65.8	27.3	70.7
iris	5.0	91.3	4.5	94.0
meteorological	24.5	70.0	1374.9	82.1
wine	5.0	74.7	5.6	93.3

Table 3: Fuzzy rules subsets formed by the fuzzy tree conversion.

Dataset	Rules		Accuracy (%)	
	Mean	Std. Dv.	Mean	Std. Dv.
balance	7.3	4.9	54.9	2.9
car	5.0	0.0	70.0	0.2
credit card	2.0	0.0	86.4	1.5
indian	5.0	0.0	71.6	2.7
ionosphere	5.0	0.0	65.8	3.7
iris	5.0	0.0	96.0	5.6
meteorological	35.5	21.6	52.5	3.8
wine	5.0	0.0	75.3	10.7

Focusing on the amount of rules it is clear that the initiation by fuzzy trees produced the smallest individuals. It is worthy of mention that the less the amount of rules the faster is the evolutive process. On the meteorological dataset the gain in running time was notable, although it was not recorded.

Table 6 extends this idea to the entire first population by presenting the average amounts of rules present on the candidate solutions and the diversity as well. It can be observed that the initiation by fuzzy decision trees generated much less rules and this has affected the diversity of the population. In fact this consequence was already expected (cf. section 3.2).

Whether this low diversity has prevented the genetic algorithm to find good classifiers is a question answered by analyzing the information showed on Table 7. Except on car and meteorological datasets the fuzzy classifiers found were better than the crisp decision trees induced by C4.5. On those datasets, although less accurate, the fuzzy classifiers are much more compact and with a simpler structure than the rules obtained from the crisp decision trees, besides employing few attributes on most cases.



Table 4: Best fuzzy rules subsets randomly generated.

Dataset	Rules		Accuracy (%)	
	Mean	Std. Dv.	Mean	Std. Dv.
balance	18.5	4.2	62.0	3.9
car	31.6	5.4	71.2	2.4
credit card	39.3	9.3	72.1	5.1
indian	20.5	4.6	70.7	4.1
ionosphere	84.7	17.2	70.4	4.5
iris	14.3	2.9	83.8	6.9
meteorological	168.5	35.0	57.7	2.3
wine	52.0	12.2	80.7	9.0

Table 5: Best fuzzy rules subsets generated from the fuzzy tree.

Dataset	Rules		Accuracy (%)	
	Mean	Std. Dv.	Mean	Std. Dv.
balance	7.7	5.1	56.8	4.0
car	4.7	0.5	70.0	0.2
credit card	2.4	0.8	86.7	1.7
indian	5.4	1.0	72.8	2.4
ionosphere	7.0	1.2	69.9	3.6
iris	5.4	1.1	97.6	4.1
meteorological	40.8	22.6	55.7	3.2
wine	7.5	1.0	85.4	8.2

Table 6: Average amounts of rules and diversity of the first populations.

Dataset	Fuzzy trees scheme		Random scheme	
	Total	Diversity	Total	Diversity
balance	152.3	14.0	317.6	59.6
car	128.6	25.9	638.8	119.6
credit card	75.8	28.6	800.9	149.7
indian	106.9	11.0	428.0	79.3
ionosphere	144.5	37.7	1734.1	328.6
iris	134.2	29.5	317.3	59.8
meteorological	814.2	118.9	3356.7	624.5
wine	148.3	39.9	1038.1	194.8



Hence, based on the datasets studied, the objective of initiating the genetic algorithm with better candidate solutions (than those obtained randomly) by means of fuzzy decision trees was reached without compromising the system performance.

Table 7: Best fuzzy rules subsets found by the genetic algorithm.

Dataset	Fuzzy trees scheme			C4.5	
	Rules	Attributes	Accuracy (%)	Leaves	Accuracy (%)
balance	12.1	4.0	77.6	30.9	60.0
car	9.0	4.9	84.0	54.3	96.9
credit card	5.5	3.9	91.9	24.3	86.4
indian	7.6	5.1	85.5	20.1	75.1
ionosphere	15.5	11.4	95.4	27.3	70.7
iris	3.3	1.2	100.0	4.5	94.0
meteorological	40.6	15.4	73.4	1374.9	82.1
wine	4.6	3.1	100.0	5.6	93.3

5 Final considerations

As observed, the fuzzy genetic system obtained very good results in every case studied when compared to a very known decision tree algorithm. The quality of a classifier was estimated by some characteristics considered important in this work such as accuracy, comprehensibility, amount of rules and features employed. The objective of finding accurate and relatively short fuzzy classifiers and feature selectors was successful. The results obtained from the eight problems suggest that the initiation of the genetic algorithm by fuzzy decision trees is an interesting strategy because it has reduced the amount of rules of the first population without compromising its accuracy neither the system performance. Higher dimension problems will be studied to verify the robustness of this scheme. Future studies may consider other kinds of fuzzy trees and alternative ways of genetic algorithm initiation.

Acknowledgements

This research was supported by CNPQ and the Petroleum National Agency under the program PRH-ANP/MME/MCT.

References

- [1] Liu, H. & Motoda, H., *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers: Boston, 1998.



- [2] Gordon, A.D., *Classification*, Chapman and Hall: London, 1981.
- [3] Janikow, C.Z., Fuzzy Decision Trees: Issues and Methods. *IEEE Trans. on Systems, Man, and Cybernetics*, **28(1)**, pp. 1-14, 1998.
- [4] Espindola, R.P. & Ebecken, N.F.F., Data classification by a fuzzy genetic system approach. *Proceedings of Fourth International Conference on Data Mining*, eds C.A. Brebbia, N.F.F. Ebecken & A. Zanasi, WIT Press: Southampton, pp. 467-476, 2000.
- [5] Ishibuchi, H., Murata, T. & Tanaka, H., Construction of Fuzzy Classification System with Linguistic If-Then Rules Using Genetic Algorithms (Chapter 11). *Genetic Algorithms for Pattern Recognition*, eds S.K. Pal & P.P. Wang, CRC Press: New York, pp. 227-251, 1996.
- [6] Espindola, R.P. & Ebecken, N.F.F., Evolving TSK fuzzy rules for classification tasks by Genetic Algorithms. *Proceedings of Second International Conference on Data Mining*, eds N.F.F. Ebecken & C.A. Brebbia, WIT Press: Southampton, pp. 467-476, 2000.
- [7] Espindola, R.P. & Ebecken, N.F.F., Seleção de atributos e classificação por regras fuzzy TSK otimizadas por algoritmos genéticos. *Proceedings of 22nd Iberian Latin-American Congress On Computational Methods In Engineering*, São Paulo, 2001.
- [8] Yuan, Y., Shaw, M.J., Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, **69(1)**, pp. 125-139, 1995.
- [9] Evsukoff, A., Branco, A.C.S. & Gentil, S., A Knowledge Acquisition Method for Fuzzy Expert System in Diagnosis Problems. *Proceedings of Sixth IEEE International Conference on Fuzzy Systems*, Barcelona, 1997.
- [10] Kosko, B., *Neural Networks and Fuzzy Systems - A Dynamical Systems Approach to Machine Intelligence*, Prentice Hall: New Jersey, 1992.

