A comparative study on classic machine learning and fuzzy approaches for classification problems

Marcos E. Cintra [1], Maria C. Monard[1], Heloisa A. Camargo[2], Trevor P. Martin [3]

[1]São Paulo University(USP) - Mathematics and Computer Science Institute
P.O. Box 668, 13561-970, São Carlos-SP, Brazil

[2]Federal University of São Carlos(UFSCar) - Computer Science Department
P.O. Box 676, 13565-905 São Carlos-SP, Brazil

[3]University of Bristol - Department of Engineering Mathematics
University Walk, Bristol - BS8 1TR, United Kingdom

{cintra,mcmonard@icmc.usp.br,heloisa@dc.ufscar.br,trevor.martin@bristol.ac.uk}

Abstract. *A large variety of machine learning algorithms for classification problems have been proposed in the literature. Fuzzy methods have also been proposed presenting good results for classification problems. However, papers presenting comparisons between the results of those two communities are rare. Thus, this paper aims at presenting and comparing a few classic machine learning approaches for classification tasks (J4.8, Multilayer Perceptron, Naive Bayes, OneRule, and ZeroRules) and two fuzzy methods (DoC-Based and Wang & Mendel). These initial experiments were carried out using 4 datasets. The methods were compared in terms of the error rates and also the number of rules (for the rule-based methods only).*

Resumo. *Um grande número de algoritmos de aprendizado de máquina voltados para tarefas de classificação tem sido propostos na literatura. Também foram propostos métodos baseados na lógica fuzzy que apresentam bons resultados para tarefas de classificação. Entretanto, é raro encontrar-se publicações contendo comparações de métodos das duas áreas. Dessa forma, este trabalho tem como objetivo apresentar e comparar alguns métodos clássicos de aprendizado de máquina para classificação (J4.8, Multilayer Perceptron, Naive Bayes, OneRule e ZeroRules) e dois métodos baseados na lógica fuzzy (DoC-Based e Wang & Mendel). Os experimentos iniciais foram realizados usando-se 4 bases de dados. Os métodos foram avaliados em termos das taxas de erro e do número de regras (apenas para os métodos baseados em regras).*

1. Introduction

The goal of machine learning is to program computers to use example data or past experience to solve a given problem [Parsons 2005]. Classification is an important area of machine learning, and many methods have been proposed for classification tasks. Some of the most well-known methods proposed by the machine learning community might include methods based on rules, decision trees, Bayesian methods and also those inspired on artificial neural networks.

Fuzzy systems, which are systems based on the fuzzy set theory and fuzzy logic [Zadeh 1965] proposed by professor Loft A. Zadeh, have also been used for classification problems. Some of the most researched fuzzy approaches include the neuro-fuzzy systems [Atsalakis and Valavanis 2009, Abraham 2001, Lin and Lee 1996] and the genetic fuzzy systems [Cordón et al. 2007, Cordon et al. 2004, Cintra and Camargo 2008].

Although both communities work on the same areas and with similar problems, the interaction between them is not as close or productive as it could be, thus it is difficult to find papers comparing methods from both communities. These comparisons could highly contribute for the validation standards of the proposed methods and promote the collaboration between both communities. In order to fill this gap, this paper presents some classic methods from the machine learning area, as well as two rule-based fuzzy methods, all for classification problems, aiming at comparing methods from these two communities. The methods were initially tested using 4 datasets and the results are analyzed and compared in terms of error rates for all methods and number of generated rules for the rule-based ones.

This paper is organized as follows: Section 2 briefly presents some classic machine learning methods for classification; Section 3 presents the DoC-Based and the Wang & Mendel methods; Section 4 presents the experimental evaluation followed by the conclusions in Section 5.

2. Machine Learning Methods for Classification

Machine learning studies how to automatically learn how to make accurate predictions based on past observations. Classification is a sub area of machine learning, related to supervised learning, in which the task is to classify examples into one of a discrete set of possible categories (classes). Several methods for classification have been proposed in the literature. Among the most well know are the rule based approaches, the decision trees, Bayesian approaches and the artificial neural networks.

In this study, the following methods, which are briefly described next, were used and compared:

- J4.8;
- ZeroRules;
- NaiveBayes;
- OneRule;
- Multilayer Perceptron.

J4.8 It is a popular implementation of Ross Quinlan's famous decision tree algorithm C4.5 [Quinlan 1988] available at Weka [Witten and Frank 2005].
Decision trees can be seen as a collection of rules. Each rule begins with the root test, aggregating each subsequent test, and using the leaf as class.

Zero Rules This algorithm [Witten and Frank 2005] simply classifies all inputs as belonging to the majority class. Depending on the distribution of the examples, this method can be somewhat useful, since its implementation is extremely simple. Of course, this is a *primitive* algorithm and is used mainly for comparative purposes. Since Zero Rules chooses the most common class all the time, it can be applied to determine if a classifier is useful or not, especially in the presence of one large dominating class.

NaiveBayes This classifier [Domingos and Pazzani 1997] is a simple probabilistic classifier based on applying Bayes' theorem with strong or naive independence assumptions. These naive assumptions come from the fact that a naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

OneRule This algorithm, also know as OneR, is a simple classification algorithm that generates a one-level decision tree [Holte 1993]. The OneR algorithm creates one rule for each attribute in the training data, by determining the most frequent class for each attribute value, then selects the rule with the smallest error rate as its *one rule*.

Since OneR selects the rule with the lowest error rate, if two or more rules have the same error rates, the rule is chosen at random. The OneR implementation of WEKA, on the other hand, selects the rule with the highest number of correct instances, not lowest error rate [Witten and Frank 2005].

Multilayer Perceptron An artificial neural network, usually called *neural network*, is a computational model that tries to simulate the structure and/or functional aspects of biological neural networks [Mitchell 1997]. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data, thus, they can be used for classification problems.

A Multilayer Perceptron (MLP) [Haykin 1998] is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron. It uses three or more layers of neurons (nodes) with nonlinear activation functions in order to be able to distinguish data that is not linearly separable, or separable by a hyperplane.

One of the problems generally attributed to the neural networks is the fact that it does not generate an interpretable structure, such as rules, for instance, although there are several papers on the extraction of rules from neural networks [Zhou 2004, Hayashi et al. 2000].

Next Section presents the fuzzy approaches for classification tasks.

3. Fuzzy Classification Methods

Fuzzy Classification Systems are rule based fuzzy systems designed to perform a classification task that requires the features domains to be granulated by means of fuzzy partitions. The linguistic variables in the antecedent of the rules represent features, and the consequent part represents a class. A typical fuzzy classification rule can be expressed by

$$R_k : \textbf{IF } X_1 \text{ is } A_{1l_1} \textbf{ AND } ... \textbf{ AND } X_m \text{ is } A_{ml_m} \textbf{THEN } \textit{Class} = c_i$$

where R_k is the rule identifier, $X_1, ..., X_m$ are the features of the example considered in the problem (represented by linguistic variables), $A_{1l_1}, ..., A_{ml_m}$ are the linguistic values used to represent the feature values, and $c_i \in C$ is the class, fuzzy or crisp, the example belongs to. In the classification process, an inference mechanism compares a given unlabeled example to each rule in the fuzzy rule base aiming at determining the class it belongs to.

In this work, the classification method used in the inference mechanism is the Classic Fuzzy Reasoning Method (CFRM), which classifies an example using the rule with the highest compatibility degree with it [Cintra 2007]. Two different fuzzy methods were used, the DoC-Based and the Wang & Mendel methods. They are briefly explained next.

The DoC-Based Method This method focuses on a particular type of genetic fuzzy system, namely the Rule Based Genetic-Fuzzy Systems (RBGFSs), which are rule based fuzzy systems equipped with genetic algorithm learning capabilities [Michalewicz 1996].

The DoC-based uses a criteria based on heuristic knowledge for the preselection of candidate rules to be considered by a genetic algorithm, so the number of rules that will form the search space can be reduced and the codification of the chromosomes simplified. The heuristic is associated to the Degree of Coverage (DoC) of the rules. Once the fuzzy partitions of the attribute domains are defined, the DoC values are calculated for all possible rules; the rules are then decreasingly ordered by their DoC values. The candidate rules are selected from this ordered list and used in the GA-based generation process. The DoC-Based method uses an auto-adaptive algorithm to adjust the fitness function in order to consider not only the correct classification rates of the rule bases, but also the number of rules in each rule base, thus, trying the balance the accuracy ***versus*** interpretability problem. A drawback of the DoC-Based method is that it requires the generation of all possible rules, as well as the calculation of their degree of compatibility. As these steps have exponential complexity, an attribute selection is essential in order to use the method effectively. A complete description of the DoC-Based method can be found in [Cintra 2007, Cintra and Camargo 2007].

Wang & Mendel Method This method [Wang 2003] for the automatic generation of fuzzy rule bases has low complexity and produces relatively small rule bases with good classification rates and no conflicting or redundant rules. However, nowadays it is possible to generate more precise fuzzy rule bases with a fewer number of rules using other approaches, such as genetic algorithms.

Assuming that the fuzzy partition for the domain in use is known, the generation of the fuzzy rule base using the Wang & Mendel method basically takes the following steps:

- a rule is created for each of the input examples using the fuzzy terms with highest compatibility with the input example values;
- an importance degree for each of the generated rules is assigned according to the aggregation of the degrees of membership of the input example values for the chosen fuzzy terms;
- redundant and conflicting rules are eliminated from the set of rules based on the assigned importance degree, forming the final rule set.

Next section presents the initial experiments and comparisons of the machine learning and fuzzy methods described previously.

4. Experiments

The experiments were conducted using 5-fold cross validation on 4 datasets from the UCI Machine Learning Repository [Asuncion and Newman 2007]. Table 1 shows a summary of the characteristics of the datasets used in the experiments presenting the number of examples in each dataset (# Examples), number of attributes (# Attribs), number of classes (# Classes), and the majority class error (MCE). All the attributes for the 4 datasets are continuous.

Table 1. Characteristics of the datasets.

Dataset	# Examples	# Attribs	# Classes	MCE
AutoMPG	209	7	3	53.57
Diabetes	769	8	2	34.98
Iris	150	4	3	66.67
Machine	392	7	3	64.37

The experiments with the DoC-Based method were carried out using 4 attributes from each dataset. These attributes were selected with the aid of an expert. The fuzzy data bases consisted of 3 triangular shaped fuzzy sets for each attribute, evenly distributed in the partitions. For the AutoMPG and Machine datasets, the classes, which are continuous, were fuzzyfied using 3 triangular shaped fuzzy sets. This fuzzification of the class was the base for the discretization used with the methods, thus, the split point of the classes was chosen as the point where the triangular shaped fuzzy sets crossed with 0.5 degree of membership. To allow further comparisons, experiments with all the attributes of the diabetes, AutoMPG and Machine datasets were also carried out.

The experiments with the Zero Rules, J4.8, NaiveBayes, OneRule, and MLP were carried out using the Experimenter tool of the opensource Weka suite for machine learning [Witten and Frank 2005], written in Java, using default parameters.

Table 2 presents the error rates and the standard deviation (in parenthesis) for the experiments carried out using 4 attributes from each dataset. The light-gray shaded cells highlight the best rates. The DoC-Based method shows the best performance for all datasets.

Table 2. Error rates and standard deviation for the experiments with 4 attributes.

	Fuzzy		Classic				
	DoC-Based	WM	Zero Rules	J4.8	NaiveBayes	OneRule	MLP
AutoMPG	13.68 (2.12)	20.70 (0.08)	53.57 (0.65)	21.43 (3.64)	24.93 (4.05)	22.44 (3.62)	21.81 (3.27)
Diabetes	0.03 (0.05)	8.90 (0.08)	34.39 (0.18)	26.38 (4.09)	24.89 (3.42)	27.35 (4.03)	26.10 (2.94)
Iris	0.29 (0.40)	1.77 (0.05)	66.67 (0.00)	4.00 (2.79)	4.53 (3.55)	6.00 (1.49)	3.60 (2.92)
Machine	4.90 (0.40)	6.30 (0.04)	64.60 (0.80)	9.09 (3.88)	10.43 (4.45)	13.87 (5.14)	8.71 (4.00)

Table 3 presents the error rates and the standard deviation (in parenthesis) for the experiments carried out using 4 attributes for the DoC-Based method and all the attributes for Wang & Mendell and the classic machine learning methods. The DoC-Based method was not tested with all attributes for the AutoMPG, Diabetes and Machine datasets due to the fact that it would be unfeasible to generate all possible fuzzy rules and calculate their degree of coverage, using all original attributes, as that is a requirement of the method. The light-gray shaded cells highlight the best rates. The DoC-Based method shows the best performance for 3 datasets, and the J4.8 method shows best performance for the Machine dataset, although its standard deviation is high.

Table 3. Error rates and standard deviation for the experiments with all attributes.

	Fuzzy		Classic				
	DoC-Based	WM	Zero Rules	J4.8	NaiveBayes	OneRule	MLP
AutoMPG	13.68 (2.12)	27.73 (0.72)	53.57 (0.60)	16.58 (3.82)	18.39 (2.89)	23.21 (2.42)	16.20 (3.27)
Diabetes	0.03 (0.05)	28.52 (0.22)	34.90 (0.21)	28.78 (3.31)	24.57 (3.76)	28.25 (2.87)	25.38 (3.01)
Iris	0.29 (0.40)	8.67 (0.32)	66.67 (0.00)	4.00 (2.79)	4.53 (3.55)	6.00 (1.49)	3.60 (2.92)
Machine	4.90 (0.40)	45.79 (0.93)	64.60 (0.80)	4.80 (4.48)	14.84 (5.11)	12.43 (1.91)	7.51 (4.01)

Since the NaiveBayes and the MLP methods do not produce rules, the models induced by these methods cannot be compared in terms of induced rules with the models of the other methods. The ZeroRule and OneRule methods, on the other hand, always use only one rule, so they are not taken into consideration for the comparisons carried out in this work.

Table 4 presents the number of rules and the standard deviation rates obtained by the rule-based models. The light-gray shaded cells highlight the best rates. J48 obtained best average number of rules (smallest) for all datasets, but Machine. The DoC-Based method shows the smallest number of rules for the Machine dataset.

Table 4. Number of rules generated and standard deviation rates.

	Fuzzy			Classic	
	DoC-Based	WM (4 attribs)	WM (all attribs)	J4.8 (4 attribs)	J4.8 (all attribs)
AutoMPG	13.00 (2.12)	20.00 (1.41)	73.40 (0.20)	7.20 (4.38)	15.40 (3.29)
Diabetes	14.20 (0.44)	24.00 (2.34)	110.90 (0.21)	4.60 (2.41)	24.20 (4.62)
Iris	7.60 (1.34)	15.00 (1.92)	14.70 (0.48)	4.40 (0.55)	4.40 (0.55)
Machine	6.20 (1.30)	13.8 (1.78)	22.30 (1.64)	10.60 (1.14)	8.60 (0.55)

It is important to notice that for the Machine dataset, the DoC-Based method shows the best number of rules and error rate comparable to the best one, obtained by J4.8, while for the remaining datasets, although J4.8 shows the smallest number of rules, its performance in terms of error rates is not comparable to the ones obtained by the DoC-Based method. In fact, we can say that the reduction of the models generated by J4.8 does not pay for their poor performance in terms of error rates. It is also worth mentioning that the number of rules generated by the Wang & Mendel method increases considerably with the number of attributes.

To test whether there is a significant difference among the methods, assessed in terms of the error rates, the Friedman test [Demšar 2006] was used with the null-hypothesis that the performance of all methods are comparable. As the null-hypothesis was rejected with a 95% confidence level, the Bonferroni-Dunn and Nemenyi post-hoc tests (to detect whether the differences among the methods are significant) were used [Demšar 2006].

Results showed that the DoC-Based method is significantly better than the NaiveBayes, OneRule, and ZeroRule methods with a 95% confidence level. Results also showed that the J4.8, MLP and Wang & Mendel methods are significantly better than the ZeroRules method with a 95% confidence level.

Table 5 shows the position and average ranks for each of the methods tested for the classic and general fuzzy reasoning methods. DoC-Based is the first one, followed by Wang & Mendel, MLP, J4.8, NaiveBayes, OneRule, and ZeroRules.

Table 5. Ranking for the Friedman's test.

	DoC-Based	WM	Zero Rules	J4.8	NaiveBayes	OneRule	MLP
Position	1	4	7	3	5	6	2
Ranking Average	1.14	4.13	7.00	3.37	4.14	5.00	3.00

Summing up, it is possible to observe the following:

- Although the DoC-Based method did not perform as well as the J4.8 method in terms of interpretability, its performance in terms of error rates is satisfactory and might justify future work on improving its interpretability;
- The MLP and NaiveBayes approaches performed well in terms of accuracy, but their results lack interpretability, which can be a fundamental issue depending on the application in focus;
- The Wang & Mendel method had reasonable accuracy, but its interpretability is affected with the increase of attributes, while OneRule and ZeroRules, although they are very simple, lack accuracy power.

5. Conclusions

Both the machine learning community and the fuzzy logic community work on classification problems and both have proposed several methods which show good performance for this particular task. Nevertheless, papers comparing results obtained with methods from these two communities are rare. These comparisons could highly contribute for the validation standards of the proposed methods from both communities and promote more collaboration between both communities.

To this end, this initial work presented and compared 4 classic machine learning methods and 2 fuzzy methods for classification problems. The experiments were executed using 4 datasets and a 5-fold cross-validation strategy.

The results show that the DoC-Based method performs well in terms of accuracy and interpretability. J4.8 performs well in terms of interpretability, although paying a high price in terms of accuracy. The MLP and NaiveBayes methods obtained good accuracy but their results are not interpretable. The Wang & Mendel method also had good accuracy, however, it generates a high number of rules when compared to the other methods. OneRule and ZeroRules methods had poor accuracy.

As future work, we intend to proceed with further comparisons of methods from both, the machine learning and fuzzy communities, using more algoritms and more datasets for classification problems.

Acknowledgements

This research was supported by CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and the São Paulo State Research Foundation - FAPESP.

References

[Abraham 2001] Abraham, A. (2001). Neuro fuzzy systems: State-of-the-art modeling techniques. *Lecture Notes in Computer Science*, 2084:269–276.

[Asuncion and Newman 2007] Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

[Atsalakis and Valavanis 2009] Atsalakis, G. S. and Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications: An International Journal*, 36(7):10696–10707.

[Cintra 2007] Cintra, M. E. (2007). Genetic generation of fuzzy rules with preselection of candidate rules. Dissertacao de mestrado, Universidade Federal de São Carlos. Programa de Pós-Graduação em Ciência da Computação.

[Cintra and Camargo 2007] Cintra, M. E. and Camargo, H. A. (2007). Fuzzy rules generation using genetic algorithms with self-adaptive selection. *IEEE International Conference on Information Reuse and Integration - IRI*, 13-15:261–266.

[Cintra and Camargo 2008] Cintra, M. E. and Camargo, H. A. (2008). Generation of fuzzy rule bases with preselection of candidate rule. *19th Brazilian Symposium on Artificial Intelligence - VI Best MSc Dissertation/PhD Thesis Contest (CTDIA 2008)*, 1:1–10.

[Cordón et al. 2007] Cordón, O., Alcalá, R., Alcalá-Fdez, J., and Rojas, I. (2007). Special section on genetic fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 15:533–592.

[Cordon et al. 2004] Cordon, O., Gomide, F. A. C., Herrera, F., Hoffmann, F., and Magdalena, L. (2004). Ten years of genetic fuzzy systems: Current framework and new trends. *Fuzzy Sets and Systems*, 141(1):5–31.

[Demšar 2006] Demšar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

[Domingos and Pazzani 1997] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130.

[Hayashi et al. 2000] Hayashi, Y., Setiono, R., and Yoshida, K. (2000). A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. *Artificial Intelligence in Medicine*, 20:205–216.

[Haykin 1998] Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Macmillan College Publishing Company, 2 edition.

[Holte 1993] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90.

[Lin and Lee 1996] Lin, C. T. and Lee, C. S. G. (1996). *Neural Fuzzy Systems*. Prentice Hall.

[Michalewicz 1996] Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer.

[Mitchell 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

[Parsons 2005] Parsons, S. (2005). *Introduction to Machine Learning*, volume 20. Cambridge University Press.

[Quinlan 1988] Quinlan, J. R. (1988). *C4.5 Programs for Machine Learning*. Morgan Kaufmann, CA.

[Wang 2003] Wang, L. (2003). The WM method completed: a flexible fuzzy system approach to data mining. *IEEE International Conference on Fuzzy Systems*, 11:768–782.

[Witten and Frank 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, second edition.

[Zadeh 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

[Zhou 2004] Zhou, Z. (2004). Rule extraction: Using neural networks or for neural networks? *Journal of Computer Science and Technology*, 19:249–253.