Contribution to Automatic Design of a Hierarchical Fuzzy Rule Classifier

Cristhian Molina[3,2], Vincent Bombardier[1,2] and Patrick Charpentier[1,2]
[1]*Université de Lorraine, CRAN, UMR 7039, Campus Sciences, BP 70239, Vandoeuvre-lès-Nancy Cedex, 54506, France*
[2]*CNRS, CRAN, UMR 7039, Vandoeuvre-lès-Nancy Cedex, France*
[3] *Universidad de Santiago de Chile, Santiago, Chile*

Keywords: Feature Selection, Fuzzy Associative Rules, Pattern Recognition, Fuzzy Rule Classifier.

Abstract: In this paper, two ways for automatically designing a hierarchical classifier is checked. This study deals with a specific context where is necessary to work with a few number of training samples (and often unbalanced), to manage the subjectivity of the different output classes and to take into account an imprecision degree in the input data. The aim is also to create an interpretable classification system by reducing its dimensionality with the use of Feature Selection and Fuzzy Association Rules generation. The obtained results over an industrial wood datasets prove their efficacy to select input feature and they are used to make some conclusions about their performance. Finally, an original methodology to automatically build a hierarchical classifier is proposed by merging the both previous methods. Each node of the hierarchical structure corresponds to a Fuzzy Rules Classifier with selected inputs and macro classes for output. The leaves are the outputs of the classification system.

1 INTRODUCTION

In this paper, we propose a contribution to the automatic design of a fuzzy hierarchical classifier, working in a specific industrial wood context.

As presented in many classification reviews (Gordon, 1987), classification in image processing is done based on the characteristics of a bunch of features, arranged as a vector (input features) extracted from the image of a product in order to classify it in a certain group (output classes).

In classification systems, many factors are considered to evaluate their results. In most of cases, systems are considered efficient because they have a certain level of accuracy, even if they have some restrictions like their computational complexity, or if they present difficulties in order to be applied. This implies that their profitability will depend on the application context. Thus, as says the "ugly duckling theorem", a classification system depends on the application context.

In our wood industrial domain (Bombardier 2010), these constraints to be considered principally are:

- The leak of training data, often unbalanced regarding the output classes.
- The subjectivity of the classification, which is done by an operator.
- The fuzziness of the output classes, which are often not disjointed.

We will also improve the interpretability of the model given by the classification algorithms. However, many of them work as a black box, hiding the actual process from the user, or are so complex that their comprehensibility is out of limits.

In (Bombardier 2010), is shown that the Fuzzy Sets Theory seems to be one of the best techniques to deal with the subjectivity of the system and its non-disjointed output classes. Additionally, the recently application of fuzzy rule-based systems in pattern classification tasks (Nakashima, 2007) and their ability to work with few learning data sets (Wang 2008) appears like the best way to deal with almost all the limitations presented above. Finally, the desired interpretability of the system can be achieved either with the creation of a hierarchical structure of fuzzy rule classifiers (Bombardier 2007).

In (Horng 2009) it has been mentioned that hierarchical methods are computationally intensive when both, the size of the data set and the number of classes are large.

Molina, C., Bombardier, V. and Charpentier, P..
Contribution to Automatic Design of a Hierarchical Fuzzy Rule Classifier.
In *Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI 2015) - Volume 2: FCTA*, pages 150-155
ISBN: 978-989-758-157-1

There are in literature, principally three ways to reduce the complexity of a classification system, as defined before:

- to reduce the number of features, which naturally leads to the reduction of the number of rules in the system;
- to reduce the number of features per rule, keeping the most "interesting" ones under certain criterions;
- to create a hierarchical classification system, in order to simplify each decision level with the creation of different macro classes.

So, in this paper, we propose to check the first two ways in order to contribute to the third one.

In section 2, we will provide their definitions, with advantages and disadvantages, discussing their theoretical efficiency. In section 3 we will show some experiments and their results, in order to present the actual efficacy of feature selection processes and draw some conclusions about their behavior.

2 REDUCING THE COMPLEXITY OF A SYSTEM

2.1 Feature Selection

Feature Selection step is a pre-process that chooses a subset of the initial features. There exist many potential benefits in feature selection mentioned in (Guyon 2003). Among others, it facilitates the visualization and comprehension of the data, reduces the training and utilization time of the classification method, and challenges the dimensionality in order to improve the accuracy of the classification.

Langley divided the features selection methods, taking into account the presence or absence of a classification algorithm in the process. These two categories are known as "filter" methods and "wrapper" methods (Langley 1994).

"Filter" methods select a subset of features, from the dataset in a classification process. In consequence, their computational cost is low, which facilitate their application. This type of methods is independent of the classification algorithm and hence, the chosen subset can be used in different classifiers, without influencing the classification rate. As showed in (Liu 2014), the most used criterions in this kind of methods, take into account the data structure and the information that the spatial distribution contains. (Ferreira 2012), (Guyon 2003) use the dependence between the features and output classes (or correlation). (Liu 2014), (Zhang 2002) and (Zhao 2013) consider both, the interclass distance and the homogeneity of the elements in the same class in order to preserve the internal structure of data. This kind of criterions are well adapted for treating high dimensional problems, and considering that it is not our primary objective, we will not focus our attention on this kind of methods.

"Wrapper" methods use a classification algorithm to measure the efficacy of the selected subset, generating a combinatorial problem (NP-Hard) with a big computational cost (Ferreira 2012). These characteristics make Wrapper methods less efficient for working with high dimensional problems. The utilization of a classification algorithm in the selecting process creates a subset with a high discriminative power, but this power can only be guaranteed for the used classifier (Guyon 2003). The most used criterion in this kind of methods is the misclassification rate, generating a big number of tests of every subset in order to achieve an optimal classification, as in (Li 2004) (De Lannoy 2011), where SVMs are used as the trained classifier.

(Pudil 1994) introduces the Sequential Floating Search Methods (SFSM), probably one of the most known feature selection techniques, and our primary reference in feature selection processes. Within SFSM we can distinguish forward methods (SFFS) and backward methods (SFBS), as the most known and used Wrapper methods.

(Kira 1992) presents another reference method, called Relief. It works in a statistical way, searching the features which are statistically relevant. This method acts dividing the entire group of instances into positive instances and negative instances taking into account the neighbourhood of each training sample. After this, it upgrades the weight of each feature, calculating the relevance based on the information given by each feature in the before mentioned process. Finally, a feature will be considered as important, if its relevance is over a threshold τ.

(Chen 2012) also defines an interesting feature selection method, combined with a fuzzy rule extraction for classification. In this, the author creates a modulator, modifying the membership function of every feature, measuring their influence in the creation of distance based fuzzy rules. With this, it attempts to create a fuzzy clustering method, using only in those features that have the most relevant influence to the classification rate, forcing the created modulator to work as a "gate", which

remains closed for every feature until they prove that their influence is remarkable.

(Schmitt 2008) propose a really interesting feature selection method which deals with our context. The Choquet integral is used to select the more suitable features, according to their capacity with respect to it. Also, in this method, they measure the importance of the different decision criterions according to Shapley indexes, which measure the contribution of a decision criterion to the final decision, and Murofushi indexes, which measure the interaction power between two indexes, discarding negative feedback. This method works as an iterative algorithm, discarding weaker features in order to keep good recognition rates.

In (Grandvalet 2003) an automatic relevance determination of features in kernelized SVM is presented. Here, relevance is measured by scale factors defining the input space metric, and the features are selected by assigning zero weights to irrelevant features.

2.2 Reducing the Number of Features per Rule

Fuzzy association rules can be considered as an indirect way of reducing the dimensionality of a problem, showing the existing relationships between different elements present in a dataset (Han 2006). Basically, given a number of features, is possible to create a defined number of rules, combining all features, and measure the efficacy of each rule according to certain criterions such as Support and Confidence Zhang 2002).

FARC-HD is a classification method which uses fuzzy association rules in its process (Alcala 2011). Each rule is built in a hierarchical way, combining all possible features and their fuzzification terms, considering a restriction in the premise part. After this process, the Support and the Confidence of each rule are calculated in order to keep the most relevant set of rules for the classification process.

FURIA is an algorithm which is used to create a set of fuzzy rules for classification processes (Huhn 2009). This algorithm has a learning phase, where it creates the initial set of rules for each class, using a One Against All (OAA) strategy, in order to learn how to separate the current class from all the others.
These methods work in a similar way, creating a bunch of rules in order to cover all the possibilities, and afterwards they apply a selection process, to keep the most important ones. As the result, they create a suitable set of rules for classification problems.

The above mentioned property is the principal difference with a classical fuzzy rule set algorithm such as (Ishibuchi, 1992), where all the possibilities are covered and the lacking of a selection process implies the creation of a set of rules that becomes not understandable.

The structure of a set of Fuzzy Rules can be described as follow. Given N training patterns, xp= (x1, x2, …, xm) p= (1, 2, …, n) belonging to S output classes, where Xpi represents the i-th feature i= (1, 2, …, m) of the p-th training pattern, each fuzzy rule will have the following structure:

$$R_j: If\ x_1\ is\ A_{j1}\ and\ x_m\ is\ A_{jm}\ then\ Class = C_j \quad (1)$$

Where *Rj* is the label of the j-th rule, *x*= (x1, …, xm) is a feature vector with a dimension m, A_{ji} is a fuzzy label, *Cj* is a class label j= (1, 2, …, s).

In a classical approach the system will cover all the possible combination of rules, in order to have no "empty space" in the classification phase. Considering that N is the number of features and Card(Tv) the number of terms in which the feature V is divided, the number of rules is:

$$\text{Number of Rules} = \prod_{v=1}^{N} Card(T_v) \quad (2)$$

In fuzzy associative rules for classification, we appreciate the same structure as any classification rule but their efficacy is measured as follows:

$$Support\ (A \rightarrow C_j) = \frac{\sum_{x_p \in Class C_j} \mu_A(x_p)}{|N|} \quad (3)$$

$$Confiance\ (A \rightarrow C_j) = \frac{\sum_{x_p \in Class C_j} \mu_A(x_p)}{\sum_{x_p \in T} \mu_A(x_p)} \quad (4)$$

Where |N| is the number of transactions in the transactions set T, μA(xp) is the matching degree of the pattern xp with the antecedent part of the fuzzy rule. With these measures, the methods perform the mentioned selection process.

3 EXPERIMENTAL ANALYSIS

In this section a series of experiments are proposed, testing the contribution of feature selection processes applied in a similar context to which our work is immersed, to improve the interpretability of a system. For the classification process, FARC-HD will be used, which includes the creation of fuzzy association rules, allowing us to extrapolate the obtained results, merging the classic feature

selection methods with the dimensionality reduction propositions given by fuzzy association rules.

3.1 Dataset

It is important to clarify that despite of the explanations given for the dataset, specifying the classification purpose and the context in which the dataset is immersed, for privacy issues, the considered features will remain with a generic name, without explaining their significance for the industrial environment.

The dataset used is from a company which takes place in the industrial environment of wood. The associated process to this dataset, named Wood, is in charge of finding, within a set of defined features, those that represent in the best possible way the output classes, which correspond to wood singularities. The process is performed considering the aspects of the context, such as non-balanced data and the complexity of the field, represented by the imprecision of wood recognition approaches, depending on the specificity and characterization of the measures. The aim is to obtain a good accuracy level using as little features as possible, because of real time constraints. The obtained model is wished to be interpretable in order to create a knowledge model of the system.

Within the technical specifications, the Wood Database has 20 input features, 9 output classes and around 250 training patterns. These previously provided specifications make the set of rules created by the classic classification systems (SVM, KNN, Neuronal Networks, etc.) too large, and that in general the rules to be non-interpretable.

3.2 The Used Classifier FARC-HD

FARC-HD is a fuzzy association rule-based classification method, based on three stages to obtain an accurate and compact fuzzy rule set (Alcala, 2011). First, it limits the order of the associations in the association rule extraction process, performed by a basic hierarchical decision tree. Secondly, it considers a subgroup discovery process, based in a weighted relative accuracy measure, used to select the most interesting rules before a genetic postprocessing process for rule selection and parameter tuning is performed by a genetic algorithm.

As we mentioned before, the aim of these experiments, is to merge the effects of traditional feature selection processes with the ability of creating a compact fuzzy association rule set, proposed by FARC-HD.

3.3 Experimental Methodology

The methodology will be divided in two parts, selection of Features and classification using those parameters combined with FARC-HD. Within the used feature selection methods (see section 2.1), we will find SFFS, SBFS (Pudil, 1994) and RelieF (Kira, 1992), also we will use our own Fuzzy Rule Iterative Feature Selection method (FRIFS) (Schmitt, 2008), the Modulator Gate Method (MGM) (Chen, 2012) and SVM method (Grandvalet, 2003).

Feature selection processes performed by each method were applied separately to the dataset. This way, each method provides a set of the most “important” features according to their own criterion.

As the aim of the entire process lies in to prove the efficiency of feature selection in the creation of an interpretable descriptive model, we have focused our experiments in providing a general idea of the most useful features in the dataset. So, the results of feature selection processes have been analyzed considering the frequency of appearance of each feature within the different subsets.

For the second part of the experiments, the application of FARC-HD method leads to different subsets of preselected features, used separately. It includes them directly into the classifier and in combination, to perform the data-crossing process mentioned before, considering then only the most frequent, or “important” ones in the classification process. Within the iterations of the classifier, we have employed the “leave one out” strategy, in order to determine empirically the relevance degree of each selected feature, measuring the obtained accuracy in each step.

Additionally as a new experiment, we have tested the efficacy of a possible “feature selection process” within the rules selection mechanism performed by FARC-HD when keeping the most interesting association rules. In order to filter the most interesting features, we select them regarding the utilization patterns of each feature in the different rules created by FARC-HD. We assume that the most used features are the most important ones. This way, we will be able to prove or disprove the following hypothesis: to keep the most important rules, implies to keep the most important features in them.

3.4 Experimental Results

In Table 1, the different sets of features selected by the different feature selection methods applied on Wood are presented.

In Table II the performance and the analysis of the classifier FARC-HD in the dataset Wood is shown, where we have the following parameters:

1. #Features stands for the total number of features used in the process.
2. #UF stands for the number of used features in the created rules.
3. #R stands for the average number of rules.
4. #C stands for the average number of conditions (or features) in the premises of the rules.
5. TRA stands for the average classification percentage obtained over the training data.
6. TST stands for the average classification percentage obtained over the test data.

The best global result for each one is stressed in boldface in each case.

Table 2 clearly shows that the efficiency of the system, in terms of accuracy, increases using less features than the 20 original features. It also shows that using "the most important rules" created by the fuzzy association rules system, some features are discarded, not being used in the classification process (it uses only 16 out of 20). Using less features also improves the number of rules in the system, from 25 using all features until 15 or 16, depending on the case, using less features. As we mentioned before, the reduction of the number of rules increases the interpretability of the system, but here we have to consider also the number of conditions in each rule, to achieve the understandability of the rule. In fact, Table 2 shows that using in average, 2.5 conditions per rule, we can achieve the best accuracy of the system, which is an acceptable number of conditions in order to draw some conclusions about the behavior of the system.

In the other hand, with the data-crossing performed between the different subsets of features, and the utilization rates given by FARC-HD, we can also show that the reduction of dimensionality performed by the feature selection methods, and the reduction of dimensionality performed by fuzzy associative rules are from different nature. This means that if we apply a reduction of the number of rules via the application of fuzzy association rules, it does not necessarily mean that the most important features, according to the state of art feature selection methods, are going to be considered. Likewise, the application of feature selection processes on a dataset, does not assure that the rules created by the system are going to be interpretable, given the nature of regular fuzzy rules creation processes.

Table 1: Selected Features by different Feature Selection Methods applied to Wood Database.

FRAC-HD	FRIFS	FRIFS-HS1	FRIFS-HS2	SBFS	SFFS	SVM
C4	SM-Axis	SM-Axis	Area	LR_RE	LR_RE	SM-Axis
C3	DX/DY	DX/DY	DX/DY	SM_Axis	SM_Axis	Area
C1+C3	LR	LR	LR	Area	Area	DX/DY
SM_Axis	C1	C1	C1	DX/Dy	DX/DY	C1
LR	C4	C1+C3	C4	C1	C1	C4
Area_Rate	C1+C3		C1+C3	C3	C3	C1+C3
Orient				C1+C3	C1+C3	

Table 2: Results obtained by using different Feature Selection Methods.

Dataset	**#Feature**	**#UF**	**#R**	**#C**	**TRA**	**TST**
Wood	20	16	25	2.68	0.984	0.729
FRIFS	6	6	16	2.25	0.968	0.756
FRIFS HS1	5	5	16	2.25	0.896	0.742
FRIFS HS2	6	6	15	2.6	0.96	0.77
SBFS	8	8	16	2.625	0.968	0.772
SFFS	8	8	16	2.625	0.968	0.772
SVM	6	6	20	2.45	0.964	0.772
MGM	4	4	15	2.533	0.876	0.718

4 CONCLUDING REMARKS

In this paper, we have tested two ways to contribute to the automatic creation of a hierarchical classification system: reducing the number of input variables with feature selection methods and reducing the number of rules with the use of fuzzy associative rules. With the execution of some experiments, we have noticed the power of the dimensionality reduction in order to improve the interpretability of a system.

That is why, we think that both ways for reducing the dimensionality need to be merged or included simultaneously in a classifier, increasing the benefits provided in the separated scenario. The proposed methodology is based on feature selection process to reduce dimensionality, and fuzzy association rules creation to have a hierarchical structure in order to be able to divide the process in sub processes with different macro classes.

REFERENCES

Alcala-Fernandez, J., Alcala, R., and Herrera, F. (2011). A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning, IEEE Transactions on Fuzzy Systems, 19 (5), 857–872.

Bombardier, V., and Schmitt, E. Measure (2010). Fuzzy rule classifier: Capability for generalization in wood color recognition, Engineering Applications of Artificial Intelligence, 23 (6), 978–988.

Bombardier V., Mazaud C., Lhoste P. Vogrig R. (2007) Contribution of Fuzzy Reasoning Method to knowledge Integration in a wood defect Recognition System. Computers in Industry Journal 58:355–366

Chen, Y. C., Pal, N. R., and Chung, I.F. (2012). An Integrated Mechanism for Feature Selection and Fuzzy Rule Extraction for Classification, IEEE Transactions on Fuzzy Systems, 20 (4), 683–698.

De Lannoy, G., François, D., and Verleysen, M. (2011). Class-Specific Feature Selection for One-Against-All Multiclass SVMs, European Symposium on Artificial Neu. Net., Computacional Intel. and Mach. Learn.

Ferreira, A. J, and Figueiredo, M. A. (2012). Efficient feature selection filters for high-dimensional data, Pattern Recognition

Gordon, A. D. (1987). A review of hierarchical Classification. Journal of the Royal Society. Series A, 150 (2), 119-137.

Grandvalet, Y., and Canu, S. (2003). Adaptive scaling for feature selection in SVMs, in Neural Information Processing System. Cambridge, MA: MIT Press.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection, J. Mach. Learn. Res., 3, 1157-1182.

Han, J., Kamber, M., and Pei, J. (2006). Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann.

Horng, S.C., and Hsiao, Y.L. (2009). Fuzzy clustering decision tree for classifying working wafers of ion implanter, IEEE International Conference on Industrial Engineering and Engineering Management, 703–707.

Hühn, J., and Hüllermeier, E. (2009). FURIA: an algorithm for unordered fuzzy rule induction, Data Mining and Knowledge Discovery, 19 (3), 293–319.

Ishibuchi, H., Nozaki, K., Tanaka, H., (1992). Distributed representation of fuzzy rules and its application to pattern classification. Fuzzy Sets and Systems 52, 21–32.

Kira, K., and Rendell, L.A. (1992). The feature selection problem: Traditional methods and a new algorithm, Proceedings of Ninth National Conference on Artificial Intelligence, 129-134.

Langley, P. (1994). Selection of relevant features in machine learning, Proceedings of the AAAI Fall Symposium on Relevance, 1–5.

Li, G.Z., Yang, J., Liu, G.P., and Xue, L. (2004). Feature selection for multi-class problems using support vector machines, Lect. Notes in comp. science, 3157, 292-300.

Liu, Wang, L., Zhang, J., Yin, J., and Liu, H. (2014). Global and Local structure Preservation for Feature Selection, IEEE trans. Neu. net. and learn. Sys., 25 (6).

Nakashima, T., Schaefer, G., Yokota, Y., and Ishibuchi, H. (2007). A weighted fuzzy classifier and its application to image processing tasks, Fuzzy Sets and Systems, 158, 284–294.

Pudil, P., Novovicova, J., and Kittler, J. (1994).Floating search methods in feature selection, Pattern recognition letters 15, 1119-1125.

Schmitt, E., Bombardier, V., and Wendling, L. (2008). Improving Fuzzy Rule Classifier by Extracting Suitable Features From Capacities With Respect to the Choquet Integral, IEEE trans. On Systems, mand and Cybernetics-Part B: Cybernetics, 38 (5), October.

Wang, F., Man, L., Wang, B., Xiao, Y., Pan, W., Lu, X. (2008) Fuzzy-based algorithm for color recognition of license plates, Pattern Recognition Letters 29, 1007–1020.

Zhang, C., and Zhang, S. (2002). Association Rule Mining: Models and Algorithms. Berlin, Heidelberg: Springer-Verlag.

Zhao, Z., Wang, L., Liu, H., and Ye, J. (2013). On Similarity preserving Feature Selection, IEEE Trans. Knowledge and Data engineering, 25 (3).