Wrapper-based Fuzzy Inductive Reasoning model identification for imbalance data classification

Solmaz Bagherpour Àngela Nebot Francisco Mugica
Soft Computing Group
Universitat Politècnica de Catalunya - BarcelonaTech (UPC)
Jordi Girona Salgado 1-3, Barcelona, Spain
sbagherpour@cs.upc.edu, angela@cs.upc.edu, fmugica@cs.upc.edu

***Abstract*— Fuzzy Inductive Reasoning (FIR) is a qualitative inductive modeling and simulation methodology for dealing with complex dynamical systems. FIR has proven to be a powerful tool for qualitative model identification and prediction of future behavior of different kinds of system domains including biology, medicine, ecology, etc. FIR has been mainly applied to regression problems, but recently we are interested in studying the feasibility of FIR as a classifier. The main objective of this study is to analyze and revise the model selection process in FIR methodology from the perspective of a classifier when dealing with imbalance data. In this research we propose a wrapper technique for fuzzy model identification in the context of FIR. We demonstrate that this new approach exhibits a significant improvement comparing to classical FIR model selection when applied to imbalanced data classification. In this paper we also compare FIR Classifier with wrapper model selection to similar genre of classic rule-based and instance-based classifiers, i.e. RISE, kNN, C4.5, CN2, PART, RIPPER and Modlem, when applied to a set of classification benchmarks.**

Keywords—Fuzzy inductive reasoning; imbalance data; classification; wraper-based models

I. INTRODUCTION

Fuzzy inductive reasoning (FIR) is a qualitative inductive modeling and simulation methodology for dealing with dynamical systems. It has been proved that FIR can be a powerful tool for qualitative model identification and prediction of future behavior of various kinds of dynamical systems especially in soft sciences, such as biology, bio-medicine and ecology [1]. While FIR has been mainly applied to regression problems, recent studies demonstrate that it can be also useful for classification problems and its performance is comparable to other well know similar classifier [2, 3]. However, previous attempts to develop FIR classifiers were based on the classical model identification process of FIR methodology, which was originally proposed to model dynamical systems. These previous approaches are referred in this paper as the Basic FIR classifier. Therefore, the main objective of this study is to analyze and revise the model identification/selection process of FIR methodology from the perspective of a classifier when dealing with imbalance data.

Class imbalance where one class (a minority class) is underrepresented in comparison to the remaining majority classes is a very common problem in real world classification. Most rule-based classifiers are biased towards the majority classes and they have difficulty with correct recognition of minority classes. Class imbalances have been also observed in many different application problems including medical problems where the number of patients requiring special attention is much smaller than the number of patients who do not need it.
A number of different approaches have been proposed for dealing with class imbalance in classifiers. They are mainly divided to either data level preprocessing methods or methods that deal with the algorithm itself. Although several specialized methods already exist, the identification of conditions for the efficient use of a particular method is still an open research problem. The main underlying factors that contribute to the difficulty of this problem are nature of the imbalanced data, key properties of its underlying distribution and nature of each particular algorithm and their consequences [4, 5].

As mentioned before the main objective of this research is to analyze and revise the model identification/selection process of FIR methodology from the perspective of a classifier. In this paper we empirically show that when FIR is applied to classification problems with imbalance data, the selection of the model structure performed by the traditional FIR model identification process might not be the best choice if we want to give importance to minority and rare cases. To solve this problem we propose a new Wrapper-based approach to obtain FIR classification models.

Section II introduces the main characteristics of FIR as a classifier. The Basic FIR classifier is reviewed and the new Wrapper-based FIR approach is proposed and described. In section III the main metrics to measure the performance of imbalance data classification are presented. The results of the classification experiments performed in this research are described in section IV. Finally, the main conclusions and future research are presented in section V.

II. FIR AS A CLASSIFIER

There are different computational approaches to deal with classification problems. The discriminative approaches, that are the simplest ones, construct a discriminating function that

 DOI 10.1109/FUZZ-IEEE.2018.8491622

directly assigns each vector x to a specific class. More powerful approaches separate inference and decision by using conditional probability and they are referred as generative approaches [6]. The reason the second group is considered more powerful is because the linear decision boundaries of the first group arise from a simple assumption about the distribution of data, which is usually not the case in real world problems. FIR as a classifier is similar to the second group, it´s model selection or variable selection is done through an inference process and its decision making is done through a *k*NN approach. According to previous studies, FIR's variable selection analysis turns out to work well even in those applications where standard statistical variable selection analysis does not provide any useful information [1].

A. Basic FIR Classifier

The FIR methodology has four main processes, i.e. fuzzification, qualitative model identification, fuzzy classification and defuzzification [1], as described in Fig. 1.

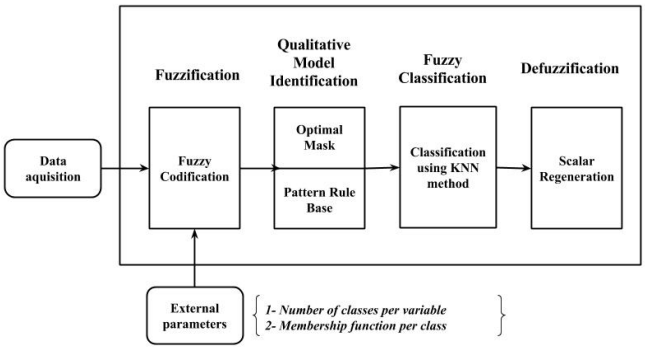

Fig. 1. Main processes of FIR methodology

The methodology starts with Fuzzification function which converts each quantitative (real-valued) data point into a qualitative triple. Conversion of quantitative values into qualitative triples requires external parameters which include providing the number of classes into which the space is going to be discretized for each variable, as well as the membership function or discretization algorithm that should be used per class. The qualitative model identification is responsible for finding causal relations between variables, i.e. a mask in the FIR terminology, and selecting the most optimal mask among them. A mask is a matrix representation of the spatial and temporal relations between the selected variables, which from machine learning terminology can be viewed as a form of feature selection process. The optimal mask is the one that maximizes the forecasting power of the qualitative model.

Each mask is evaluated by a quality measure that is mainly based on the Shannon entropy. The obtained mask is used to form the behavior matrix of the system, which can be interpreted as a special kind of fuzzy finite state machine relating the mask inputs to the mask output (fuzzy rules). For a more detailed description of the qualitative model identification process and of the mask concept the reader is refereed to [1].

Once the mask and the behavior matrix are available, FIR classification algorithm applies a specialization of the *k*-nearest neighbors technique, commonly used in pattern recognition, to obtain the class of the output. Basic FIR classification algorithm is brought next for reference.

P = Apply_Mask (Input_Instances);
applies the selected Mask (the one that has the highest quality taking into account all the complexities) to the instance to be classified and obtains the input pattern P

SP = Find_Patterns (P,BM);
finds all the instances of Pattern P in the Behavior Matrix (BM) of the selected Mask (*SP* means Same Patterns)

SPD = Same_Pattern_Distances (P, SP);
finds distances between P and all the same pattern instances found in the BM

if (all SPD = 0)

Class_Output = More_Represented_Class (SPD)
predicts as output the class that is more represented in the patterns with distance 0

else

Neighbours = FindNeighbours (K,SPD);
finds *k*NN (k-nearest neighbors) among SPDs

Class_Output = More_Represented_Class (Neighbors);
predicts as output the class that is more represented in the *Neighbors*

B. Wrapper-based FIR classifier

There are two ways we can go around the problem of improving FIR for dealing with minority class, one approach is by improving the learning process of finding the optimal mask and, another approach is by improving the classifier itself.

In this paper we are focusing on the first approach considering mask selection as a feature selection and by getting inspiration from the literature on feature selection for imbalanced data sets. Feature selection algorithms for flat features (where features are assumed to be independent), are usually divided into three groups: filter models, wrapper models, and embedded models. Relying on the characteristics of data, filter models evaluate features without utilizing any classification algorithm. Filter models select features independent of any specific classifier. Therefore, the major disadvantage of the filter approach is that it totally ignores the effects of the selected feature subset on the performance of the induction algorithm [7]. The optimal feature subset should depend on the specific biases and heuristics of the induction algorithm. Based on this assumption, wrapper models utilize a specific classifier to evaluate the quality of selected features, and offer a simple and powerful way to address the problem of feature selection, regardless of the chosen learning machine [8].

Basic FIR classifier can be viewed as a form of a filter model. It provides an optimal mask for each complexity starting from only one feature (complexity 1) to n features (complexity n), being n the number of variables.

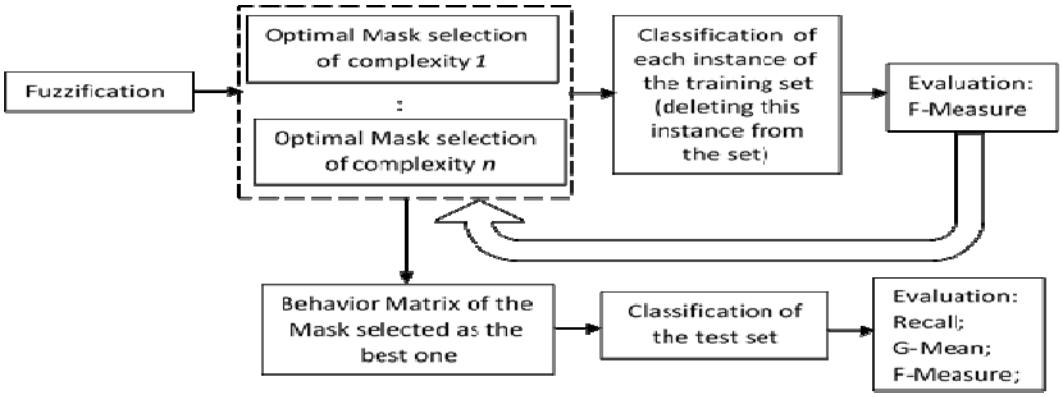

Fig. 2. Schema of the Wrapper-based FIR approach

The optimal mask of each complexity is the mask with the highest quality and generalizability among all other masks of the same complexity. The problem here is how to chose the best complexity for both majority and minority classes. Our proposal is to follow the wrapper models and evaluate each mask with regard to their classification ability using a proper metric for dealing with minority classes (F-Measure). Fig. 2 presents an schema of the Wrapper-based FIR approach.

In a wrapper approach, in order to be able to evaluate the classification performance of each mask for each complexity, a validation set is usually needed in order to avoid overfitting. However, when working with binary imbalanced datasets this technique can be very costly since the datasets may have a reduced number of instances on the minority class, reducing even more the number of data available for training and test. Therefore, in this research we have decided to choose another approach. Since the core of the FIR classification process is based on a *k*NN algorithm, we can bypass the validation set and evaluate the classification accuracy of each model using the initial training set. In order to do that, we perform a classification for each single instance of the training set, but the instance to be classified is eliminated from the behavior matrix. In this way we maximize the use of the available data preventing classifying instances that belong to the training set i.e. instances that were used to obtain the model, achieving a realistic evaluation of the FIR model (mask). We have made an extensive validation of this model evaluation technique, comparing the results with the ones obtained when a validation set is used, giving comparable results. Notice that this approach is inspired in the leave-one-out technique. However, it is not the same since the behavior matrix is obtained using all the data instances available, and only when the classification process take place the test instance to be classified is eliminated from the matrix.

III. Evaluation Metrics for Imbalanced Data Classification

Evaluation metrics play an important role in evaluating and guiding the learning algorithms.

Empirical evaluation remains the most used approach for the algorithm assessment, although machine learning algorithms can be evaluated through empirical assessment or theory or both. Empirical comparison is most often done by applying algorithms on various data sets and then evaluating the performance of the classifiers that the algorithms have produced; accuracy is being the most often used measure [9].

The choice of metrics depend on different factors mainly summarized as: the type of classifier (Deterministic, Scoring or probabilistic), data distribution (balanced or imbalanced class distribution), the type of classification task in hand (Binary, Multi-Class, Multi-labelled or Hierarchical) and problem domain area (depending on the domain of the problem in hand for example some measures might be more wide spreadly used in Medical domain while different ones might be applied to Information retrieval domain) [10].

When there is imbalance in data distribution, if the choice of metrics doesn't value the minority class then the imbalance problem is not issued well. The most commonly used metric which is the overall classification rate (Accuracy) is not a suitable metric for imbalanced datasets. The minority class has less affect to accuracy comparing to the majority class so accuracy is biased towards majority class. There are other metrics that can be used and are less affected by imbalance as precision, recall, F-measure, Sensitivity, Specificity, geometric mean, ROC curve, AUC, and precision-recall curve.

What is essential is using different measures not only one, so combination of different metrics can reflect different aspects of the learning algorithm. In choice of this combination of measures, it is important to pay attention to the fact that each measure focuses on different aspect of learning, sometimes pairs of measures can be so different in nature that good results on one yields to bad results on the other one. In such cases global measures might help but then even pairs of those global measures also might disagree because they themselves also focus on different aspects of learning as composite measures.

In this research we have decided to use three metrics to evaluate the performance of the classifiers. These metrics are Sensitivity or True Positive Rate (also known as Recall), F-

TABLE I. DESCRIPTION OF THE DATASETS. OBTAINED FROM UCI MACHINE LEARNING REPOSITORY [12] AND THE WORK PRESENTED IN [11]

Dataset	Category	# Variables	Variable Type	# Instances	% of instances of the minority class
New-thyroid	Safe	5	Int.(1), Real(4)	215	16,28
Vehicle	Safe	18	Int.(18)	846	23,52
Breast	Safe_Borderline	9	Real(9)	699	34,47
Ionosphere	Safe_Borderline	34	Real(34)	351	35,89
Ecoli	Borderline	7	Real(7)	336	10,42
Pima	Borderline	8	Int.(2), Real(6)	768	34,89
Haberman	Borderline_rare	3	Int.(3)	306	26,47
CMC	Rare	9	Cat.(7), Int.(2)	1473	22,61
Abalone	Rare_Outlier	8	Int.(1), Real(7)	4177	7,94
Transfusion	Outlier	4	Real(4)	748	23,8
Yeast	Outlier	8	Real(8)	1484	3,44

Measure and G-Mean. All of them computed from the confusion matrix.

- *Sensitivity = TP/(TP + FN)*
- *Specificity = TN/(TN + FP)*
- *Precision = TP/(TP + FP)*
- *F-Measure = 2* Precision *Sensitivity/(Precision + Sensitivity)*
- *G-Mean = SQRT(Sensitivity*Specificity)*

As the improvement of recognizing the minority class is usually associated with the decrease of recognizing the majority classes, aggregated measures (G-Mean and F-Measure) are considered in this work to characterize the performance of the classifiers.

IV. CLASSIFICATION EXPERIMENTS

In order to compare the performances of the Basic FIR and the Wrapper-based FIR classifiers to other rule-base and instance-base classifiers, we selected eleven datasets from a very complete research, where different rule-base classification techniques were presented and discussed [11]. The eleven benchmarks are publicly available at [12], and are binary classification problems with different degrees of imbalance. In [11] the authors, using a 5NN approach, categorized these datasets based on the distribution and quality of data presented in each one. This categorization also relates to the difficulty of the classification problem at hand. The categories chosen are: *Safe* (the easiest type to classify), *Safe_Borderline*, *Borderline*, *Borderline_rare*, *Rare*, *Rare_Outlier* and *Outlier*. We tried to have at least one example of each dataset type. Some of the datasets that [11] used are multi-class by nature, but they provided a very detailed road map of how they prepared them as binary classes. We followed their guidelines and, therefore, our data sets are identical to theirs.

Table I presents the description of the eleven datasets selected. The first column contains their name. The second has the category of the data defined by [11]. The third column holds the number of variables that the dataset has. The fourth, describes the type of the variables. Next column lists the number of instances available and, the last column determines the percentage of data that has the minority class.

A. Basic FIR vs. Wrapper-based FIR Classifiers

We are first interested in comparing the new wrapper approach with the basic FIR approach to see the performance of the last one when dealing with imbalanced datasets.

In this first experiment a five times repeated 10-fold cross validation has been used in both classifiers for the eleven datasets of Table I. The Basic FIR classifier, as explained previously, selects the mask that has the highest quality. Therefore, the features selected are those obtained directly from the best mask. Contrarily, the Wrapper-based FIR classifier computes several masks of different complexities (from 2 to 9), and using a wrapper approach selects the one that performs better for the classification task. Table II presents the mask complexities selected from both methods for each dataset.

As can be seen in Table II, the complexities of the masks for each approach vary considerably in some of the datasets, as for example Vehicle, Ionosphere, CMC, Abalone or Yeast.

Let us take a look to the masks obtained by both approaches (Basic FIR and Wrapper-based FIR) for two of the datasets, i.e. Abalone and Yeast.

The mask identified by the Basic FIR classifier for the Abalone dataset, contains “sex” and “height” features. This means that only this input variables are the ones used in the classification process. On the other hand, the mask identified by the Wrapper-based FIR approach for the same dataset contains the features “sex”, “length”, “whole weight”, “shucked weight” and “viscera weight”.

TABLE II. COMPLEXITY OF HE MASK (NUMBER OF FEATURES SELECTED BY THE QUALITATIVE MODEL IDENTIFICATION PROCESS OF FIR) FOR BASIC AND WRAPPER-BASED FIR CLASSIFIERS FOR EACH DATASET

Dataset	Basic FIR Classifier	Wrapper-based FIR Classifier
New-thyroid	2	2
Vehicle	4,5	7
Breast	2,3	3,4
Ionosphere	3,4	2
Ecoli	3	2
Pima	4	4
Haberman	1,4	2
CMC	3	5,7
Abalone	2	6
Transfusion	3	4
Yeast	3	1

The wrapper approach has obtained a mask that takes into account much more features to perform the classification of new instances.

In the case of the Yeast dataset, the mask obtained by the Basic FIR classifier identifies as relevant features for classification variables: “mcg”, “alm” and “erl”. *mcg* is the McGeoch's method for signal sequence recognition; *alm* is the score of the ALOM membrane spanning region prediction program; *erl* represents the Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Surprisingly, Wrapper-based FIR approach selected as the best mask the one that only contains as a relevant feature the “alm” variable. It concludes that “alm” has enough information to perform a good classification. Therefore, each technique identify the best set of features (mask) for classification, the first one using a filter approach and the second one using a wrapper approach.

The results of the Basic FIR and Wrapper-based FIR models when used to classify the datasets described in Table I are shown in Table III. Table III presents the minority class Recall, G-Mean and F-Measure results for each dataset and for each FIR classifier.

As can be seen from this Table, the classification measures are higher when using the Wrapper-based FIR classifier for almost all datasets. The shaded boxes in Table III indicates a better performance of the Wrapper-based classifier with respect the Basic classifier.

In some of the datasets the metrics enhancement is small, like in the case of Vehicle, Breast, Pima, Transfusion, Yeast and Haberman. Contrarily, in Ionosphere, CMC and Abalone datasets the enhancement in all three measures is considerable. There are 2 datasets, i.e. New-thyroid and Ecoli that the classification accuracy remains the same for both FIR approaches. However, it is important to notice that the Wrapper-based approach does not obtain worse results than the Basic one in any of the studied datasets.

To prove statistically that the Wrapper-based FIR classifier performs better than the Basic FIR classifier, we apply the Wilcoxon signed-rank statistical test. To this end, it is tested the null hypothesis that the vector $x - y$ comes from a distribution whose median is zero at the 5% significance level, being x and y the Basic FIR and Wrapper-based FIR classifier metrics, respectively. At the default 5% significance level, the value h=1 indicates that the test rejects the null hypothesis of zero median, and h=0 indicates a failure to reject the null hypothesis. The results for each metric are shown in Table IV.

TABLE III. RESULTS OF BASIC FIR AND WRAPPER-BASED FIR CLASSIFIERES FOR THE ELEVEN DATASETS

	Recall (Sensitivity)		G-Mean		F-Measure	
Dataset	***Basic FIR Class.***	***Wrapper-based FIR Class.***	***Basic FIR Class.***	***Wrapper-based FIR Class.***	***Basic FIR Class.***	***Wrapper-based FIR Class.***
New-thyroid	0,920	0,920	0,942	0,942	0,898	0,898
Vehicle	0,866	0,890	0,906	0,924	0,855	0,883
Breast	0,922	0.926	0,932	0,936	0,909	0,915
Ionosphere	0,719	0,821	0,805	0,852	0,761	0,813
Ecoli	0,590	0,590	0,721	0,721	0,595	0,595
Pima	0,539	0,545	0,673	0,680	0,589	0,598
Haberman	0,240	0,243	0,422	0,424	0,291	0,295
CMC	0,007	0,243	0,019	0,453	0,009	0,284
Abalone	0,137	0,227	0,359	0,465	0,209	0,296
Transfusion	0,340	0,354	0,544	0,555	0,398	0,412
Yeast	0,171	0,173	0,313	0,327	0,195	0,221

TABLE IV. WILCOXON SIGNED-RANK STATISTICAL TEST FOR THE NULL HYPOTESIS THAT BOTH, BASIC FIR AND WRAPPER-BASED FIR CLASSIFIERES PERFORM EQUALLY WELL.

Evaluation Metric	*h*	*p*
Recall (Sensitivity)	1	0,0156
G-Mean	1	0,0078
F-Measure	1	0,0039

As shown in Table IV, h is equal to one for all the metrics and the p values are all lower than the significance level 0.05, therefore, the null hypothesis can be rejected, meaning that statistically the Wrapper-based FIR classifier performs better than the Basic FIR classifier.

In this experiment, we have used the best *k* that suits each dataset. Analyzing in detail those cases where the Wrapped-based FIR classifier was failing in some of the minority instances, we conclude that a small value of *k* would help to obtain better results, since instances of the majority class interfere negatively in the class inference. Therefore, we have decided to study the behavior of the Wrapper-based FIR approach when *k* is set to 1, in those datasets that have less than 20% of instances in the minority class, i.e. New-thyroid, Ecoli, Abalone and Yeast. The results are presented in Table V.

TABLE VI. RESULTS OF DIFFERENT CLASSIFIERES FOR THE ELEVEN DATASETS

Data sets	Measure	Algorithms							
		RISE	***kNN***	***C4.5***	***CN2***	***PART***	***RIPPER***	***Modlem***	***Wrapper-based FIR***
New-Thyroid	Recall	0,928	0,867	0,850	0,866	0,933	0,855	0,812	0,971
	G-Mean	0,951	0,921	0,901	0,915	0,953	0,911	0,878	0,980
	F-Measure	0,947	0,895	0,843	0,906	0,918	0,879	0,848	0,966
Vehicle	Recall	0,831	0,865	0,867	0,329	0,883	0,874	0,859	0,890
	G-Mean	0,895	0,914	0,911	0,513	0,919	0,919	0,916	0,924
	F-Measure	0,855	0,877	0,867	0,433	0,875	0,885	0,892	0,883
Breast	Recall	0,959	0,968	0,917	0,886	0,947	0,896	0,887	0,926
	G-Mean	0,963	0,969	0,929	0,929	0,950	0,928	0,926	0,936
	F-Measure	0,949	0,957	0,912	0,915	0,932	0,910	0,910	0,915
Ionosphere	Recall	0,902	0,629	0,837	0,779	0,840	0,818	0,824	0,821
	G-Mean	0,928	0,780	0,878	0,870	0,888	0,874	0,890	0,852
	F-Measure	0,913	0,747	0,847	0,850	0,864	0,848	0,872	0,813
Ecoli	Recall	0,505	0,578	0,597	0,185	0,420	0,445	0,400	0,590
	G-Mean	0,638	0,701	0,717	0,284	0,554	0,587	0,568	0,721
	F-Measure	0,517	0,592	0,593	0,244	0,450	0,473	0,465	0,595
Pima	Recall	0,551	0,558	0,507	0,408	0,591	0,377	0,485	0,545
	G-Mean	0,666	0,681	0,649	0,600	0,679	0,581	0,641	0,680
	F-Measure	0,577	0,599	0,567	0,512	0,596	0,484	0,550	0,598
Haberman	Recall	0,224	0,181	0,244	0,184	0,334	0,180	0,240	0,243
	G-Mean	0,375	0,334	0,426	0,345	0,468	0,355	0,401	0,424
	F-Measure	0,240	0,214	0,300	0,235	0,349	0,233	0,262	0,295
CMC	Recall	0,293	0,308	0,404	0,096	0,377	0,071	0,256	0,243
	G-Mean	0,507	0,517	0,586	0,258	0,543	0,255	0,472	0,453
	F-Measure	0,351	0,358	0,434	0,140	0,361	0,124	0,311	0,284
Abalone	Recall	0,128	0,137	0,339	0,160	0,188	0,184	0,245	0,316
	G-Mean	0,345	0,358	0,568	0,396	0,419	0,421	0,484	0,543
	F-Measure	0,192	0,208	0,393	0,253	0,269	0,282	0,326	0,324
Transfusion	Recall	0,297	0,319	0,386	0,150	0,429	0,088	0,371	0,354
	G-Mean	0,507	0,529	0,579	0,342	0,602	0,266	0,529	0,555
	F-Measure	0,354	0,385	0,443	0,214	0,462	0,149	0,354	0,412
Yeast	Recall	0,245	0,194	0,323	0,000	0,267	0,259	0,189	0,245
	G-Mean	0,436	0,341	0,511	0,000	0,420	0,452	0,337	0,418
	F-Measure	0,311	0,243	0,352	0,000	0,287	0,286	0,245	0,286

TABLE V. RESULTS OF THE WRAPPER-BASED FIR APPROACH OBTAINED WITH *K* SET TO 1 FOR NEW-THYROID, ECOLI, ABALONE AND YEAST DATASETS (ALL OF THEM WITH < 20% OF INSTANCES IN THE MINORITY CLASS)

Dataset	Recall (Sensitivity)	G-Mean	F-Measure
New-thyroid	0,971	0,980	0,966
Ecoli	0,590	0,721	0,595
Abalone	0,316	0,543	0,324
Yeast	0,245	0,418	0,286

As in the previous experiment, this has also been carried out through a 5 times repeated 10 fold cross validation process.

Table V shows that our intuition was right, since we get better classification results for three of the datasets, i.e. New-thyroid, Abalone and Yeast. For New-thyroid, Wrapper-based FIR with k=1 approach obtains Recall, G-Mean and F-Measure values of 0,971, 0.980 and 0.966, respectively, vs. 0,920, 0,942 and 0,898 that got, for the same metrics, previously (see Table III). Abalone and Yeast have also a not insignificant increase. Abalone gets 0,316/0,543/0,324 vs. 0,227/0,465/0,209 and Yeast 0,245/0,418/0,286 vs. 0,173/0,327/0,195.

B. Comparison to other Classifiers

Having shown that FIR is useful as a classification technique and considering that wrapper technique significantly improved FIR performance applied to imbalanced datasets, we have decided to compare the Wrapper-based FIR approach with other existing general rule-based and instance-based classifiers that were already reported in [11], i.e. RISE, kNN, C4.5, CN2, PART, RIPPER and Modlem. The results of this comparison is presented in Table VI. As in the previous experiment, Recall, G-Mean and F-Measure are used to evaluate the performance of each classifier for the eleven datasets. The best results for each measure and dataset are highlighted. From Table VI it can be seen that FIR obtains the best Recall result for two datasets (New-Thyroid and Vehicle), the best G-Mean for three datasets (New-Thyroid, Vehicle and Ecoli), and the best F-Measure for two datasets (New-Thyroid and Ecoli). It has, also, the second Recall score for Ecoli and Abalone, the second G-Mean score for Pima and Abalone and the second F-Measure score for Pima.

Following the work of Napierala and Stefanowski [11], we use a statistical approach to compare the differences in performance between all classifiers. We apply a non-parametric Friedman test to globally compare the performance of the eight classifiers on the eleven datasets. The null-hypothesis in this test is that all compared classifiers perform equally well. The p values obtained for Recall, G-Mean and F-Measure are 0.000027, 0.0021 and 0.02, respectively. Therefore, the null-hypothesis can be rejected for all the three measures since the p values are lower than $\alpha = 0.05$.

Table VII presents the mean average of the Recall, G-Mean and F-Measure of all the datasets for each algorithm. Inside parenthesis is the ranking of each classifier algorithm.

As it can be seen from Table VII, C4.5 is the algorithm that performs better on average taking into account all the measures. Then, PART and Wrapper-based FIR algorithms have almost the same performance. Both algorithms have performances very close to the ones obtained by C4.5.

Considering that our goal of conducting this experiment was to see if Wrapper-based FIR is performing significantly better or worse than other algorithms, we can see that it doesn't have any significant worse results than others. Moreover, it can be concluded that Wrapper-based FIR is performing significantly better than CN2 and RIPPER in all metrics, but it is not significantly different than the rest of the classifiers, although among all three measures, it always stands in the first three best ranked algorithms along with C4.5 and PART.

V. CONCLUSIONS AND FUTURE WORK

The main objective of this study was to analyze and revise the model selection process of FIR methodology from the perspective of a classifier when dealing with imbalance data. In this paper we empirically show that when FIR is applied to classification problems with imbalance data, the quality of the mask might not be the best choice if we want to give importance to minority and rare cases. A Wrapper-based approach has been proposed for fuzzy model identification in the context of FIR to solve this problem. We have shown that the new approach exhibits a significant improvement comparing to classical FIR model selection when applied to imbalanced data classification. In this paper we also compared Wrapper-based FIR with other rule-based and instance-based classifiers when applied to a set of benchmarks.

It has been shown that Wrapper-based FIR is performing significantly better than CN2 and RIPPER algorithms in all metrics (Recall, G-Mean and F-Measure), but it is not significantly different than the rest of the classifiers, i.e. C4.5, PART, RISE, *k*NN and Modlem. However, among all the

TABLE VII. MEAN AVERAGE RESULTS OF DIFFERENT CLASSIFIERES FOR THE ELEVEN DATASETS

	RISE	***kNN***	***C4.5***	***CN2***	***PART***	***RIPPER***	***Modlem***	***Wrapper-based FIR***
Recall	0,533 (4)	0,509 (5)	0,570 (1)	0,367 (8)	0,560 (2)	0,459 (7)	0,506 (6)	0,558 (3)
G-Mean	0,655 (4)	0,641 (5)	0,696 (1)	0,496 (8)	0,672 (3)	0,595 (7)	0,640 (6)	0,680 (2)
F-Measure	0,565 (4)	0,552 (5)	0,595 (1)	0,427 (8)	0,578 (3)	0,505 (7)	0,549 (6)	0,579 (2)

values of the three metrics, it always stands in the first three best ranked algorithms along with C4.5 and PART.

In the near future we are planning to study the use of instant selection approaches together with the Wrapper-based FIR methodology for the same type of classification problems. We are interested also in comparing Wrapper-based FIR with reported specific classifiers suited for imbalanced classification, as for example BRACID [11].

ACKNOWLEDGMENT

This research was supported by the Ministerio de Economia y Competitividad, under Project DPI2015-68651-R.

REFERENCES

[1] A. Nebot and F. Mugica, "Fuzzy Inductive Reasoning: a consolidated approach to data-driven construction of complex dynamical systems," International Journal of General Systems, vol. 41(7), pp. 645-665, 2012.

[2] S. Bagherpour, F. Mugica and A. Nebot, "A Hierarchical Perspective to Fuzzy Inductive Reasoning," Proceedings IEEE International Conference on Fuzzy Systems, Istanbul, 2015.

[3] S. Bagherpour, A. Nebot and F. Mugica, "FIR as Classifier in the Presence of Imbalanced Data," Lecture Notes in Computer Science, vol. 9719, pp. 1-7,2016.

[4] J. Zhang, E. Bloedorn, L. Rosen and D. Venese, "Learning rules from highly unbalanced data sets," Proceedings IEEE International Conference on Data Mining, Brighton, 2004.

[5] K. Napierala, "Improving rule classifiers for imbalanced data," Ph.D dissertation: Pozan University of Technology, 2013.

[6] https://cw.fel.cvut.cz/wiki/_media/courses/a3b33kui/knihy/pattern_recognition_and_machine_learning_chapter_04_part.pdf?cache=nocache

[7] C. Aggrawal, "Data Classification: Algorithms and Applications," Taylor & Francis Group, LLC New York, 2015.

[8] J. Tang, S. Alelyani and H. Liu, "Feature Selection for Classification: A Review," CRC Press, 2014.

[9] M. Sokolovaa and G. Lapalmeb, "A systematic analysis of performance measures for classification tasks," Journal of Information processing and Manegement, vol. 45(4), 427–437, 2009.

[10] N. Japkowicz and S. Mohak, Evaluating Learning Algorithms: A classification Perspective, Cambridge University Press New York, 2011.

[11] K. Napierala and J. Stefanowski, "BRACID: a comprehensive approach to learning rules from imbalanced data. Journal of Intelligent," Information Systems, vol. 39(2), 335-373, 2012.

[12] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/index.php