A First Study on the Use of Fuzzy Rule Based Classification Systems for Problems with Imbalanced Data Sets [⋆]

María José del Jesus[1], Alberto Fernández[2], Salvador García[2], and Francisco Herrera[2]

[1] Dept. of Computer Science, University of Jaén, Spain,
mjjesus@ujaen.es
WWW home page: http://wwwdi.ujaen.es
[2] Dept. of Computer Science and A.I., University of Granada,
alfh@ugr.es,{salvagl,herrera}@decsai.ugr.es
WWW home page: http://sci2s.ugr.es

Abstract. In this work a preliminary study on the use of classification systems based on fuzzy reasoning in classification problems with non-balanced classes is carried out. The objective of this study is to evaluate the cooperation with pre-processing mechanisms of instances and the use of different granularity levels (5 and 7 labels) in the fuzzy partition considered. To do so, we will use simple fuzzy rule based models obtained with the Chi (and co-authors') method that extends the well-known Wang and Mendel method to classification problems.
The results obtained show that the previous step of instance selection and/or over sampling is needed. We have observed that a high over-fitting exists when we use 7 labels per variable. We will analyze this fact and we will discuss some proposals on the subject.

Key words: Fuzzy Rule Based Classification Systems, Instance Selection, Over-sampling, Imbalanced Data-sets.

1 Introduction

The design of a classification system, from the point of view of supervised learning, consists in the establishment of a decision rule that enables to determine the class of a new example in a set of known classes. When this knowledge extraction process uses as a representation tool fuzzy rules, the classification system obtained is called fuzzy rule-based classification system (FRBCS) [7].

In the classification problem field, we often encounter the presence of classes with a very different percentage of patterns between them: classes with a high pattern percentage and classes with a low pattern percentage. These problems receive the name of "classification problems with imbalanced data sets" and recently they are being studied in the machine learning field [5].

[⋆] Supported by the Spanish Project TIN-2005-08386-C05-01 and TIC-2005- 08386-C05-03.

Learning systems can have difficulties in the learning of the concept related to the minority class, so in the specialized literature it is common to use pre-processing techniques to adjust the databases to a more balanced format [4].

Studying specialized literature, we have found only a few works [10,11,12] that study the use of fuzzy classifiers for this problem, and all of them from the point of view of approximate fuzzy systems, not from the descriptive fuzzy systems ones that are the ones used in this work.

In this work our aim is to analyze the behaviour of descriptive FRBCSs applied to data-bases with non-balanced classes. We want to evaluate the pre-processing mechanism of instances that are commonly used in the field in co-operation with the FRBCS, and to study the importance of the granularity of fuzzy partitions in these problems.

To do that, this paper is organized as follows. In Section 2 we introduce the components of an FRBCS and the inductive learning algorithm used. Section 3 presents the pre-processing techniques considered in this work . In Section 4 we introduce the way to evaluate the classification systems in domains with imbalanced data-sets. Section 5 shows the experimental study carried out with seven different data-sets. Finally, in Section 6 we present some conclusions about the study done.

2 Fuzzy rule based classification systems

An FRBCS is composed of a Knowledge Base (KB) and a Fuzzy Reasoning Method (FRM) that, using the information of the KB, it determines the class for any pattern of data admissible that comes to the system.

The power of the approximate reasoning consists in the possibility to obtain a result (a classification) even when we have not an exact compatibility (with degree 1) between the example and the antecedent of the rules.

2.1 Knowledge base

In the KB two different components are distinguised:

- The *Data Base* (DB), that contains the definition of the fuzzy sets associated to the linguistic terms used in the Rule Base.
- The *Rule Base* (RB), composed of a set of classification rules

$$R = \{R_1, ..., R_L\} \tag{1}$$

  There are different types of fuzzy rules in the specialized literature but in our case we will use the following one:

  - Fuzzy rules with a class and a certainty degree associated to the classification for this class in the consequent

$$\begin{array}{ll} R_k : & \text{If } X_1 \text{ is } A_1^k \text{ and } \ldots \text{ and } X_N \text{ is } A_N^k \\ & \text{then } Y \text{ is } C_j \text{ with degree } r_k \end{array} \tag{2}$$

where $X_1, \ldots, X_N$ are features considered in the problem, $A_1^k, \ldots, A_N^k$ are linguistic labels employed to represent the values of the variables and r_k is the certainty degree associated to the classification of the class C_j for the examples that belong to the fuzzy subspace delimited by the antecedent of the rule.

2.2 Fuzzy reasoning method

The FRM is an inference procedure that uses the information of the KB to predict a class from an unclassified example. Usually, in the specialized literature [8] the FRM of the maximum has been used, also named classic FRM or the winning rule, that considers the class indicated by only one rule having account the association degree of the consequent of the rule over the example. Other FRMs combine the information contributed for all the rules that represent the knowledge of the area of which the example belongs [8]. In this work we will use, besides the classic FRM, the FRM of additive combination among rules classification degree per class.

Next we present the general model of fuzzy reasoning that combines the information given by the fuzzy rules compatibles with the example.

In the classification process of the example $e = (e_1, \ldots, e_N)$, the steps of the general model of a FRM are the following:

1. Computing the compatibility degree of the example with the antecedent of the rules.
2. Computing the association degree of the example to the consequent class of each rules by means of an aggregation function between the compatibility degree and the certainty degree of the rule with the class associated.
3. Setting the association degree of the example with the different classes.
4. Classification. Applying a decision function F over the association degree of the example with the classes which will determine, on base to the criterion of the maximum, the label of the class v with the greatest value.

At point (3) we distinguish the two methods used in this study, that is, using the function of the maximum to select the rule with the greatest association degree for each class, and using the additive function over the association degrees of the rules associated with each class.

2.3 Chi et al. Algorithm

For our experimentation we will use simple rule base models obtained with the method proposed in [7] that extends the well-known Wang and Mendel method [13] to classification problems. This FRBCS desing method establishes the relationship between the variables of the problem and sets an association between the space of the features and the space of the classes by means of the following steps:

1. *Establishment of the linguistic partitions.* Once determined the domain of variation of each feature X_i, the fuzzy partitions are computed.
2. *Generation of a fuzzy rule for each example* $e^h = (e^h_1, \ldots, e^h_N, C_h)$. To do this is necessary:
   2.1 To compute the matching degree of the example e^h to the different fuzzy regions.
   2.2 To assign the example e^h to the fuzzy region with the greatest membership degree.
   2.3 To generate a rule for the example, which antecedent is determined by the selected fuzzy region and with the label of class of the example in the consequent.
   2.4 To compute the certainty degree. In order to do that the ratio S_j/S is determined, where S_j is the sum of the matching degree for the class C_j patterns belonging to this fuzzy region delimited by the antecedent, and S the sum of the matching degrees for all the patterns belonging to this fuzzy subspace, regardless its associated class.

3 Preprocessing imbalanced datasets.

In this work we evaluate different instance selection and oversampling techniques to adjust the class distribution in training data. We have chosen the following ones[4]:

- Undersampling methods:
  - **Condensed Nearest Neighbor Rule (CNN)**. This technique is used to find a consistent subset of examples. A subset $\subseteq$ E is consistent with E if using a 1-nearest neighbor, correctly classifies the examples in E.
  - **Tomek links** This method works as follows: given two examples e_i and e_j belonging to different classes, the distance between e_i y e_j (d(e_i, e_j)) is determined. A (e_i,e_j) pair is called a Tomek link if there is not an example e_l, such that d(e_i,e_l) ¡ d(e_i,e_j) or d(e_j ,e_l) ¡ d(e_i,e_j). If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline.
  - **One-sided selection (OSS)** is an under-sampling method resulting from the application of Tomek links followed by the application of CNN. Tomek links are used as an under-sampling method and removes noisy and borderline majority class examples. CNN aims to remove examples from the majority class that are distant from the decision border.
  - **CNN + Tomek links** It is similar to the one-sided selection, but the method to find the consistent subset is applied before the Tomek links.
  - **Neighborhood Cleaning Rule (NCL)** uses the Wilson‘s Edited Nearest Neighbor Rule (ENN) [15] to remove majority class examples. ENN removes any example whose class label differs from the class of at least two of its three nearest neighbors. NCL modifies the ENN in order to increase the data cleaning.

  - **Random under-sampling** is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples.
- Oversampling methods:
  - **Random over-sampling** is a non-heuristic method that aims to balance class distribution through the random replication of minority class examples.
  - **Smote Synthetic Minority Over-sampling Technique (Smote)**[6] is an over-sampling method which form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.
- Hybrid methods: Oversampling + Undersampling:
  - **Smote + Tomek links**. In order to create better-defined class clusters, it could be applied Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.
  - **Smote + ENN**. The motivation behind this method is similar to Smote + Tomek links. ENN tends to remove more examples than the Tomek links does, so it is expected that it will provide a more in depth data cleaning.

4 Evaluation of FRBCS for imbalanced data sets

In this section we introduce our experimentation framework. First of all we present the metric we will use to compare the different methods considered. Then we will describe the data sets we have chosen for this work and all the parameters used.

4.1 Measuring error: geometric mean on positive and negative examples

Weiss and Hirsh [14] show that the error rate of the classification of the rules of the minority class is 2 or 3 time greater than the rules that identify the examples of the majority class and that the examples of the minority class are less probable to be predict than the examples of the majority one.

The most straightforward way to evaluate the performance of classifiers is based on the confusion matrix analysis. From a confusion matrix for a two class problem it is possible to extract a number of widely used metrics for measuring the performance of learning systems, such as Error Rate, defined as $Err = \frac{FP+FN}{TP+FN+FP+TN}$ and Accuracy, defined as $Acc = \frac{TP+TN}{TP+FN+FP+TN} = 1 - Err$.

Instead of using the error rate (or accuracy), in the ambit of imbalanced problems more correct metrics are considered. Specifically, it is possible to derive four performance metrics that directly measure the classification performance on positive and negative classes independently:

False negative rate $FN_{rate} = \frac{FN}{TP+FN}$ is the percentage of positive cases misclassified as belonging to the negative class;
False positive rate $FP_{rate} = \frac{FP}{FP+TN}$ is the percentage of negative cases misclassified as belonging to the positive class;.
True negative rate $TN_{rate} = \frac{TN}{FP+TN}$ is the percentage of negative cases correctly classified as belonging to the negative class;
True positive rate $TP_{rate} = \frac{TP}{TP+FN}$ is the percentage of positive cases correctly classified as belonging to the positive class.

These four performance measures have the advantage of being independent of class costs and prior probabilities. The aim of a classifier is to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates.

The metric used in this work is the geometric mean [3], which can be defined as $g = \sqrt{a^+ \cdot a^-}$, where a^+ means the accuracy in the positive examples (TP_{rate}) and a^- is the accuracy in the negative examples (TN_{rate}). This metric tries to maximize the accuracy of each one of the two classes with a good balance. It is a performance metric that links both objectives.

4.2 Data sets and parameters

In this study we have considered seven data sets from UCI which have different degrees of imbalance. Table 1 summarizes the data employed in this study and shows, for each data set the number of examples (#Examples), number of attributes (#Attributes), class name of each class (majority and minority) and class attribute distribution. All attributes are qualitative.

Table 1. Data sets summary descriptions.

Data set	**#Examples**	**#Attributes**	**Class (min., maj.)**	**%Class(min.,maj.)**
Glass	214	9	(Ve-win-float-proc, remainder)	(7'94,92'06)
Pima	768	8	(1,0)	(34'77,66'23)
Yeast	1486	8	(mit,remainder)	(16'49,83'51)
Ecoli	336	7	(iMU, remainder)	(10'42,89'58)
Haberman	306	3	(Die, Survive)	(26'47,73'53)
New-thyroid	215	5	(hypo,remainder)	(16'28,83'72)
Vehicle	846	18	(van,remainder)	(23'52,76'48)

In order to realize a comparative study, we use a ten folder cross validation approach We consider the following parameters and functions:

- Number of labels per fuzzy partition: 5 and 7 labels.
- Computation of the compatibility degree: Min t-norm.
- Combination of the compatibility degree and the certain rule degree: Min t-norm.

- Inference method: Classic method (winning rule) and additive combination among rules classification degree per class (addition) [8].

In table 2 we show the percentages of examples for each class after balancing.

Table 2. Average of class percentage after balancing.

Balance Method	% Positives (minority class)	% Negatives (majority class)
CNN_TomekLinks	63.23	36.77
CNNRb	81.29	18.71
NCL	25.52	74.48
OSS	34.56	65.44
RandomOS	50.00	50.00
RandomUS	50.00	50.00
SMOTE	50.00	50.00
SMOTE_ENN	52.85	47.15
SMOTE_TomekLinks	54.35	45.65
TomekLinks	23.84	76.16

5 Analysis of experiments

We have divided our study into three parts: the analysis of the use of preprocessing for imbalanced problems, the study of the effect of the FRM and finally the analysis of the influence of the granularity applied to the linguistic partitions together with the inference method.

Tables 3 and 4 show the global results (in training and test sets) for all the data-sets used in the experimental study, showing the behaviour of the FRBCSs. Each column represents the following:

- the FRM used (WR for the Winning Rule and AC for Additive Combination) and the number of labels employed (5-7),
- the balancing method employed, where "none" means that the original data set is maintained for training,
- the accuracy per class (a^- y a^+) where the subindex indicates if it refers to training (tr) or test (tst). It also shows the geometric mean (GM) for training (TR) and test (TST).

1. **The effect of the preprocessing methods**: Our results show that in all the cases pre-processing is a necessity to improve the behaviour of the learning algorithms.
   Specifically it is noticed that the over-sampling methods provide very good results in practice. We found a kind of mechanism (the SMOTE pre-process family) that are very good as pre-process technique, both individually and

Table 3. Global Results WMWR.

Classifier	**Balancing Method**	a^-_{tr}	a^+_{tr}	GM_{TR}	a^-_{tst}	a^+_{tst}	GM_{TST}
FRBCS-WR5	CNN_TomekLinks	23.86	*98.59*	45.04	22.49	**91.14**	40.9
FRBCS-WR5	CNNRb	70.15	73.84	68.64	65.84	63.41	60.01
FRBCS-WR5	NCL	90.87	67.23	74.54	87.26	56.13	64.26
FRBCS-WR5	None	*98.68*	52.74	68.61	*94.51*	39.78	55.01
FRBCS-WR5	OSS	86.28	62.01	71.46	83.82	52.45	63.54
FRBCS-WR5	RandomOS	82.33	88.31	*84.77*	76.9	72.88	74.46
FRBCS-WR5	RandomUS	72.28	87.53	78.11	68.06	77.59	70.61
FRBCS-WR5	SMOTE	81.19	88.32	84.13	75.91	74.86	**75.11**
FRBCS-WR5	SMOTE_ENN	74.41	90.7	81.56	70.01	80.06	74.29
FRBCS-WR5	SMOTE_TomekLinks	71.94	94.22	81.69	67.94	83.51	74.8
FRBCS-WR5	TomekLinks	93.88	63.62	73.79	90.2	51.29	62.35
FRBCS-WR7	CNN_TomekLinks	30.21	**99.1**	52.31	26.85	*80.1*	43.81
FRBCS-WR7	CNNRb	65.04	80.25	70.08	58.17	53.87	51.77
FRBCS-WR7	NCL	89.13	80.81	83.82	79.02	55.34	60.89
FRBCS-WR7	None	**99.02**	66.8	79.22	*87.13*	42.9	55.68
FRBCS-WR7	OSS	74.83	65.48	69.69	68.91	46.38	55.11
FRBCS-WR7	RandomOS	89.54	91.19	**90.23**	76.54	63.36	69.33
FRBCS-WR7	RandomUS	67.23	92.14	77.38	59.51	69.5	63.08
FRBCS-WR7	SMOTE	86.7	92.19	89.23	74.04	66.64	69.95
FRBCS-WR7	SMOTE_ENN	80.68	92.02	85.95	70.46	70.3	70.04
FRBCS-WR7	SMOTE_TomekLinks	78.94	94.99	86.35	68.7	73.47	*70.87*
FRBCS-WR7	TomekLinks	93.16	75.46	82.46	83.17	50.88	59.73

the hybrid ones. In this way, for FRBCSs we have highly competitive models. Nevertheless, this over-sampling can introduce an additional computation cost if the dataset is relatively large.

Also we may stress that the results in the case of no preprocess method is employed are very high for the negative class (majority) but quite low for the positive one (minority); hence the clear necessity of the preprocess methods.

2. **The reasoning method**: Analyzing the tables we find that there are no great differences between the type of FRM.
3. **Granularity analysis**: It is empirically shown that a big number of labels produces over-fitting, the training results are significantly better than the test ones when 7 labels per variable are used. This situation is evident in table 5. Besides, we must note that we are using relatively small databases and with few attributes, which stresses more this undesirable behaviour.

6 Concluding remarks.

In this work we analyze the behaviour of the FRBCSs applied to classification problems with imbalanced data sets and the cooperation with pre-processing methods of instances.

Table 4. Global Results FRBCS-AC.

Classifier	**Balancing Method**	a^{-}_{tr}	a^{+}_{tr}	GM_{TR}	a^{-}_{tst}	a^{+}_{tst}	GM_{TST}
FRBCS-AC5	CNN_TomekLinks	25.81	*96.88*	44.7	24.9	*89.71*	41.26
FRBCS-AC5	CNNRb	69.47	71.54	66.12	66.04	62.05	59.15
FRBCS-AC5	NCL	90.85	63.55	72.02	87.09	54.41	62.86
FRBCS-AC5	None	*98.42*	46.18	63.7	**94.65**	36.04	52.2
FRBCS-AC5	OSS	86.74	57.6	68.52	84.98	50.74	62.47
FRBCS-AC5	RandomOS	90.74	73.81	81.03	86.23	61.85	71.39
FRBCS-AC5	RandomUS	70.79	88.23	77.49	67.17	81.36	71.83
FRBCS-AC5	SMOTE	87.35	78.84	82.34	83.08	66.62	72.57
FRBCS-AC5	SMOTE_ENN	80.28	85.84	*82.55*	76.71	73.47	*74.24*
FRBCS-AC5	SMOTE_TomekLinks	77.33	88.56	81.9	73.53	75.28	72.66
FRBCS-AC5	TomekLinks	93.99	58.55	70.48	90.92	49.03	60.94
FRBCS-AC7	CNN_TomekLinks	29.15	*98.04*	50.63	26.58	*80.03*	43.12
FRBCS-AC7	CNNRb	64.77	77.46	67.73	58.76	55.38	50.37
FRBCS-AC7	NCL	89.54	77.48	81.72	79.45	54.16	60.34
FRBCS-AC7	None	*98.82*	62.14	75.74	*87.38*	40.91	54.36
FRBCS-AC7	OSS	75.92	62.24	68.17	70.17	43.17	50.91
FRBCS-AC7	RandomOS	94.06	78.71	85.3	81.54	53.9	63.62
FRBCS-AC7	RandomUS	67.33	91.2	77.28	60.46	69.28	63.79
FRBCS-AC7	SMOTE	90.66	84.94	*87.5*	78.79	58.31	65.2
FRBCS-AC7	SMOTE_ENN	84.38	87.81	85.91	74.94	63.49	*68.24*
FRBCS-AC7	SMOTE_TomekLinks	82.3	91.13	86.23	72.71	65.57	67.62
FRBCS-AC7	TomekLinks	93.06	72.83	80.42	83.7	50.34	59.53

Table 5. FRBCS with 5 labels opposite 7 labels.

FRM	**Balancing Method**	GM_{TR} 5	GM_{TR} 7	GM_{TST} 5	GM_{TST} 7
Winning Rule	RandomOS	84.77	90.23	74.46	69.33
Winning Rule	SMOTE	84.13	89.23	75.11	69.95
Winning Rule	SMOTE_TL	81.69	86.35	74.8	70.87
Additive Comb.	SMOTE	82.34	87.5	72.57	65.2
Additive Comb.	SMOTE_ENN	82.55	85.91	74.24	68.24
Additive Comb.	SMOTE_TL	81.9	86.23	72.66	67.62

The main conclusions of our analysis are: the necessity of using pre-processing instances methods to improve the balance between classes before the use of the FRBCS method, the similar behaviour of the two fuzzy reasoning methods analyzed, and the over-fitting produced when we use a high number of labels per variable.

We must point out that FRBCSs with 5 labels do not reach high classification percentages in training. It seems that classes with very few examples may need labels with a low support that enables to obtain the information associated to the class, but without including examples from the other class. It seems interesting to post-process the rule base by means of tuning methods and/or the integration of labels in a different granularity level to gather all the possible information.

Following this idea, our future work will deal with this problem. We want to use a post-processing 2-tuples and 3-tuples tuning, two methods that have shown a good behaviour adjusting the support of the membership functions for regression problems [1,2].

References

1. R. Alcalá, J. Alcalá-Fdez. and F. Herrera. A proposal for the Genetic Lateral Tuning of Linguistic Fuzzy Systems and its Interaction with Rule Selection. *IEEE Transactions on Fuzzy Systems*, 2006, to appears.
2. R. Alcalá, J. Alcalá-Fdez. M.J. Gacto and F. Herrera. Rule Base Reduction and Genetic Tuning of Fuzzy Systems based on the Linguistic 3-Tuples Representation. *Soft Computing*, 2006, to appears.
3. R. Barandela, J.S. Sánchez, V. García, E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36:3 849–851, 2003.
4. G.Batista, R.C.Prati, M.C.Monard. A study of the behaviour of several methods for balancing machine learning training data. *SIGKDD Explorations* 6:1 20–29, 2004.
5. N.V. Chawla, N. Japkowicz, A. Kolcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6:1 1–6, 2004.
6. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *JAIR* 16, 321–357, 2002.
7. Z. Chi, H. Yan, and T. Pham. Fuzzy algorithms with applications to image processing and pattern recognition. *World Scientific*, 1996.
8. O. Cordón, M.J. del Jesus, and F. Herrera. A proposal on reasoning methods in fuzzy rule-based classification systems. *International Journal of Approximate Reasoning*, 20:1 21–45, 1999.
9. J. Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution. *Tech. Rep. A-2001-2*, University of Tampere, 2001.
10. S. Visa, and A. Ralescu. Learning Imbalanced and Overlapping Clases using Fuzzy Sets *Workshop on Learning from Imbalanced Datasets II*, Washington DC., 2003
11. S. Visa, and A. Ralescu. Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size. *IPMU*, Perugia (Italy), 393–400, 2004.
12. S. Visa, and A. Ralescu. The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers - Experimental Study. *IEEE International Conference on Fuzzy Systems*, 749–754, 2005
13. L.X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:2 353–361, 1992.
14. G. M. Weiss, and H. Hirsh. A quantitative study of small disjuncts. *Seventeenth National Conference on Artificial Inteligence*, AAAI Press 665–670, 2000.
15. D.L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Communications* 2:3 408–421, 1972.