



Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets.

Alberto Fernández Hilario

In Collaboration with

S. García, F. Herrera and M.J. del Jesus

Research Group:

"Soft Computing and Intelligent Information Systems"

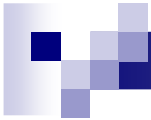
Dept. Computer Science and A.I.

University of Granada

alberto@decsai.ugr.es <http://sci2s.ugr.es>

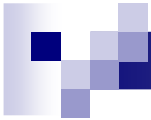


DECSAI
Universidad de Granada



Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions

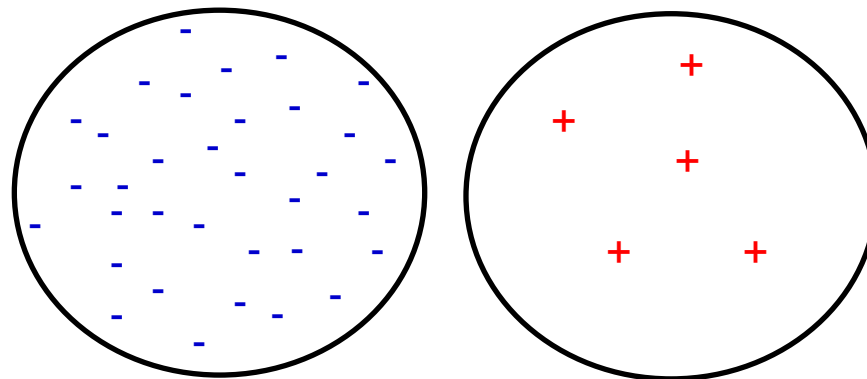


Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions

Introduction to the Problem of Imbalanced Data-sets

In many real application areas, the data used are highly skewed and the number of instances for some classes are much higher than that of the other classes.



Solving a classification task using such an imbalanced data-set is difficult due to the bias of the training towards the majority classes.



Introduction to the Problem of Imbalanced Data-sets

- The problem of imbalanced data-sets is extremely significant because it is implicit in most real-world applications:
 - **Fraud detection** (T. Fawcett and F.J. Provost 1997)
 - **Risk management** (Y.M. Huang et al. 2006)
 - Specially in **medical diagnosis** (J.W. Grzymala-Busse et al. 2005, M.A. Mazurowski et al. 2008, X. Peng and I. King 2008)
- Regarding FRBCSs, there are only a few works in the specialized literature that study their use in imbalanced data-sets.
 - *Approximate fuzzy systems* without linguistic rules (S. Visa and A. Ralescu 2003-2005).
 - *Fuzzy decision tree classifiers* (K. Crockett et al. 2006).
 - Extraction of fuzzy rules using *fuzzy graphs and genetic algorithms* (Soler et al. 2006).
 - *Enumeration algorithm*, called the E-Algorithm (Xu et al 2007).



Introduction to the Problem of Imbalanced Data-sets

■ Some References:

T. Fawcett and F. J. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291-316, 1997

M. Mazurowski, P. Habas, J. Zurada, J. Lo, J. Baker, and G. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3):427-436, 2008.

S. Visa and A. Ralescu. The effect of imbalanced data class distribution on fuzzy classifiers - experimental study. In *IEEE International Conference on Fuzzy Systems*, pages 749-754, 2005.

K. Crockett, Z. Bandar, and J. O'Shea. On producing balanced fuzzy decision tree classifiers. In *IEEE International Conference on Fuzzy Systems*, pages 1756-1762, 2006.

V. Soler, J. Cerquides, J. Sabria, J. Roig, and M. Prim. Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms. In *IEEE International Conference on Data Mining - Workshops*, pages 330-336, 2006

L. Xu, M. Y. Chow, and L. S. Taylor. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm. *IEEE Transactions on Power Systems*, 22(1):164-171, 2007.

Introduction to the Problem of Imbalanced Data-sets

Ways to evaluate the performance in this domain:

- The use of common metrics like accuracy may lead to erroneous conclusions:
 - Examples from the majority class well-classified.
 - Examples from the minority class misclassified.
- Confusion Matrix for a 2-class problem:

		Prediction	
		Pos. Class	Neg. Class
Real	Positive Class	TP	FN
	Negative Class	FP	TN

True Diagonal

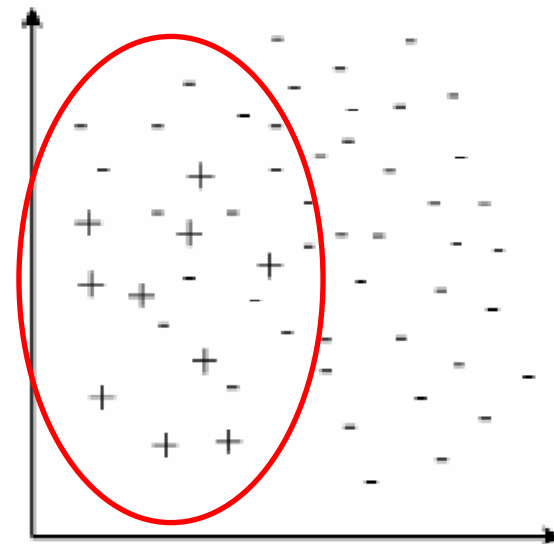
- Evaluation based on the geometric mean:
 - True Positive Rate: $a+ = TP / (TP + FN)$
 - True Negative Rate: $a- = TN / (FP + TN)$
 - Evaluation Function: **Geometric Mean of the True Rates**
$$g = \sqrt{(a+ \cdot a-)}$$

R. Barandela, J.S. Sánchez, V. García, E. Rangel. Strategies for learning in class imbalance problems. Pattern Recognition 36:3 (2003) 849-851

Introduction to the Problem of Imbalanced Data-sets

- Why is so difficult to learn from imbalanced data?

- The imbalance between classes is not the only problem for the decreasing in performance in the learning algorithm.
- **The overlapping between classes**, which is another feature of this type of problem, have also an influence in the behaviour of the algorithm.





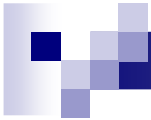
Introduction to the Problem of Imbalanced Data-sets

Solutions to deal with the Imbalanced Data-set problem:

We may distinguish between two types of solutions:

1. **Solutions at the Data Level: “Sampling”:**
 1. Random Over-Sampling.
 2. Random Under-Sampling.
 3. Generation of new synthetic examples.
 4. Combinations of all these techniques.
2. **Solutions at the Algorithm Level:**
 1. Modifying the costs per class.
 2. Adjusting the probability estimation in the leaves of a decision tree (establishing a bias towards the positive class)
 - Learning from just one class (“recognition based learning”) instead of learning from two classes (“discrimination based learning”)...

Chawla, N.V., Japkowicz, N., Kolcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations, 6:1 (2004) 1–6.



Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

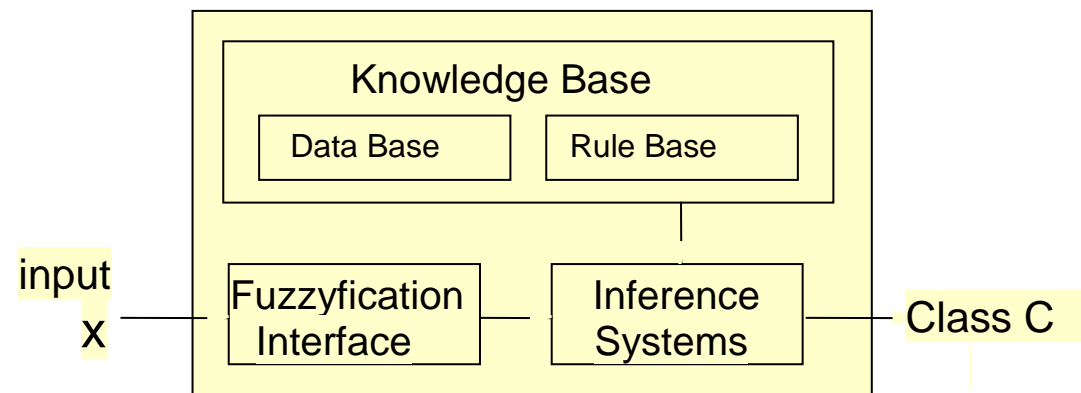
- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions

Fuzzy Rule Based Classification Systems

■ Weighted Rules:

- IF X_1 is A_1 and ... and X_n is A_n THEN Y is C
with RW

**Chi et al. Algorithm
(rule extraction)**



Fuzzy Rule Based Classification Systems

■ Rule Weights:

- They are used in order to improve the performance of the FRBCSs.
- There are some heuristics that can be used:

■ Certainty Factor:

$$CF_j = \frac{\sum_{x_p \in \text{Class } C_j} \mu_{A_j}(x_p)}{\sum_{p=1}^m \mu_{A_j}(x_p)}$$

■ Penalized Certainty Factor:

$$P-CF_j = CF_j - \frac{\sum_{x_p \notin \text{Class } C_j} \mu_{A_j}(x_p)}{\sum_{p=1}^m \mu_{A_j}(x_p)}$$

Ishibuchi, H., Yamamoto, T.: Rule Weight Specification in Fuzzy Rule-Based Classification Systems. IEEE Trans. on Fuzzy Systems 13, 428–435 (2005)

Fuzzy Rule Based Classification Systems

■ Rule Weights:

□ Mansoori Rule Weight System:

$$M-CF = \begin{cases} \mu_{A_j}(x_p) \cdot RW_j & \text{if } \mu_{A_j}(x_p) < n_j \\ \left(\frac{p_j - n_j \cdot RW_j}{m_j - n_j} \right) \cdot \mu_{A_j}(x_p) - \left(\frac{p_j - m_j \cdot RW_j}{m_j - n_j} \right) \cdot n_j & \text{if } n_j \leq \mu_{A_j}(x_p) < m_j \\ RW_j \cdot \mu_{A_j}(x_p) - RW_j \cdot m_j + p_j & \text{if } \mu_{A_j}(x_p) \geq m_j \end{cases}$$

$$n_j = t_j \sqrt{\frac{2}{1 + RW_j^2}}$$

$$m_j = \{t_j \cdot (RW_j + 1) - (RW_j - 1)\} / \sqrt{2RW_j^2 + 2}$$

$$p_j = \{t_j \cdot (RW_j - 1) - (RW_j + 1)\} / \sqrt{2RW_j^2 + 2}$$

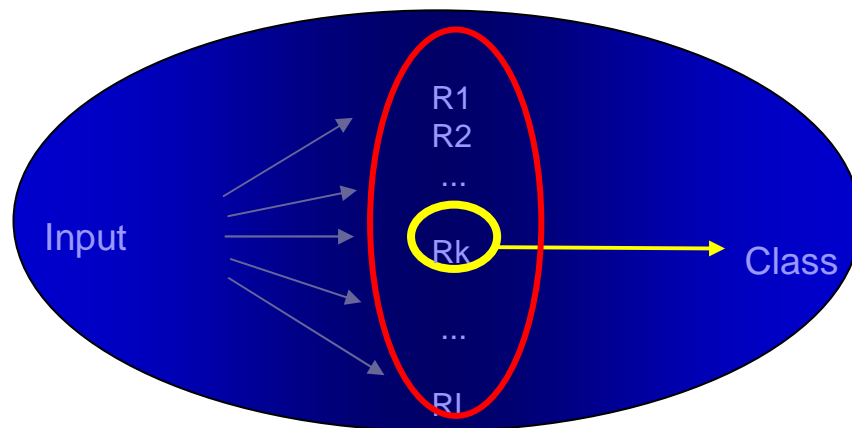
Mansoori, E.G., Zolghadri, M.J., Katebi, S.D.: AWeigthing Function for Improving Fuzzy Classification Systems Performance. Fuzzy Sets and Systems 158(5), 583–591 (2007)

Fuzzy Rule Based Classification Systems

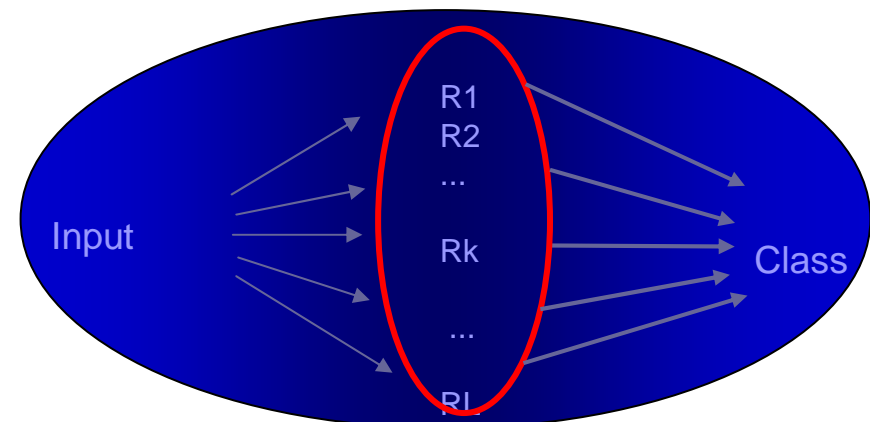
■ Fuzzy Reasoning Method

□ Two Inference Models for Classification:

- Winning Rule
- Additive Combination

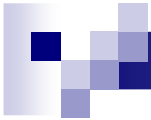


Winning Rule



Additive Combination

Cordón, O., del Jesús, M.J., Herrera, F.: A proposal on reasoning methods in fuzzy rule-based classification systems. International Journal of Approximate Reasoning, 20:1 (1999) 21-45.



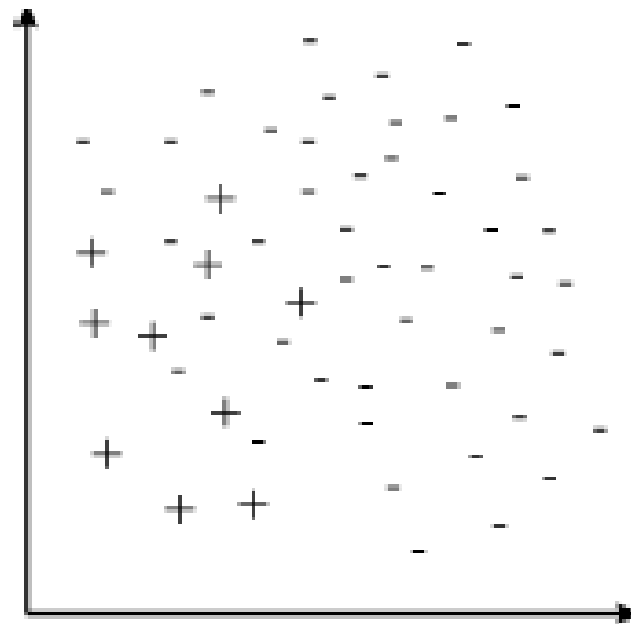
Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions

Data Preparation: Preprocessing Techniques

- We may find 4 types of data:

- ☐ Noisy Data
- ☐ Boundary Data
- ☐ Redundant Data
- ☐ Safe Data





Data Preparation: Preprocessing Techniques

■ **Motivation:**

- To balance the training data.
- To remove noisy examples that may lead to a misclassification.



Data Preparation: Preprocessing Techniques

■ **Under-sampling Methods:**

- ☐ Condensed Nearest Neighbour Rule (CNN)
- ☐ Tomek links
- ☐ One-sided selection" (OSS)
- ☐ CNN + Tomek links
- ☐ Neighbourhood Cleaning Rule" (NCL)
- ☐ Random under-sampling

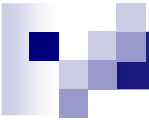
■ **Over-sampling Methods:**

- Random over-sampling
- Smote

■ **Hybrid approaches:**

- Smote + Tomek links
- Smote + ENN

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations, 6:1 (2004) 20-29.



Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions



Experimental Study

■ Objectives:

- To analyze the necessity of applying a preprocessing step to deal with the problem of imbalanced data-sets.
- For the components of the FRBCS, we are interested in:
 - The granularity of the fuzzy partitions.
 - The configuration of the FRBCS:
 - The use of distinct conjunction operators.
 - The application of some approaches for the rule weights.
 - The use of different Fuzzy Reasoning Methods.
- We will employ a strong statistical study for the comparison of our results: Non parametric Tests.

Experimental Study

□ Summary Description for the Imbalanced Data-Sets:

Data set	#Ex.	#Atts.	Class (min., maj.)	%Class(min.,maj.)	IR
<i>Low Imbalanced Datasets (1.5 to 3 IR)</i>					
Glass2	214	9	(build-window-non float-proc, remainder)	(35.51, 64.49)	1.82
EcoliCP-IM	220	7	(im, cp)	(35.00, 65.00)	1.86
Wisconsin	683	9	(malignant, benign)	(35.00, 65.00)	1.86
Pima	768	8	(tested-positive, tested-negative)	(34.84, 66.16)	1.9
Iris1	150	4	(Iris-Setosa, remainder)	(33.33, 66.67)	2
Glass1	214	9	(build-window-float-proc, remainder)	(32.71, 67.29)	2.06
Yeast2	1484	8	(NUC, remainder)	(28.91, 71.09)	2.46
Vehicle2	846	18	(Saab, remainder)	(28.37, 71.63)	2.52
Vehicle3	846	18	(bus,remainder)	(28.37, 71.63)	2.52
Vehicle4	846	18	(Opel, remainder)	(28.37, 71.63)	2.52
Haberman	306	3	(Die, Survive)	(27.42, 73.58)	2.68
<i>Medium Imbalanced Datasets (3 to 9 IR)</i>					
GlassNW	214	9	(non-window glass, remainder)	(23.83, 76.17)	3.19
Vehicle1	846	18	(van,remainder)	(23.64, 76.36)	3.23
Ecoli2	336	7	(im, remainder)	(22.92, 77.08)	3.36
New-thyroid3	215	5	(hypo,remainder)	(16.89, 83.11)	4.92
New-thyroid2	215	5	(hyper,remainder)	(16.28, 83.72)	5.14
Ecoli3	336	7	(pp, remainder)	(15.48, 84.52)	5.46
Segment1	2308	19	(brickface, remainder)	(14.26, 85.74)	6.01
Glass7	214	9	(headlamps, remainder)	(13.55, 86.45)	6.38
Yeast4	1484	8	(ME3, remainder)	(10.98, 89.02)	8.11
Ecoli4	336	7	(iMU, remainder)	(10.88, 89.12)	8.19
Page-blocks	5472	10	(remainder, text)	(10.23, 89.77)	8.77
<i>High Imbalanced Datasets (higher than 9 IR)</i>					
Vowel0	988	13	(hid, remainder)	(9.01, 90.99)	10.1
Glass3	214	9	(Ve-win-float-proc, remainder)	(8.78, 91.22)	10.39
Ecoli5	336	7	(om, remainder)	(6.74, 93.26)	13.84
Glass5	214	9	(containers, remainder)	(6.07, 93.93)	15.47
Abalone9-18	731	8	(18, 9)	(5.65, 94.25)	16.68
Glass6	214	9	(tableware, remainder)	(4.2, 95.8)	22.81
YeastCYT-POX	482	8	(POX,CYT)	(4.15, 95.85)	23.1
Yeast5	1484	8	(ME2, remainder)	(3.43, 96.57)	28.41
Yeast6	1484	8	(ME1, remainder)	(2.96, 97.04)	32.78
Yeast7	1484	8	(EXC, remainder)	(2.49, 97.51)	39.16
Abalone19	4174	8	(19, remainder)	(0.77, 99.23)	128.87



Experimental Study. Analysis of the Behaviour of the FRBCSs: Cooperation with Preprocessing Techniques and Study of the Components

■ Configuration y initial parameters:

- **5 folder cross validation.**
- **Membership Function**: Linear triangular membership function.
- **Number of labels per fuzzy partition**: 5 labels.
- **Computation of the compatibility degree**: Product T-norm.
- **Combination of compatibility degree and rule weight**: Product T-norm.
- **Rule Weight**: Penalized Certainty Factor
- **Fuzzy Reasoning Method**: Winning Rule

Experimental Study. Analysis of the Behaviour of the FRBCSs: Cooperation with Preprocessing Techniques

- The results show in almost all cases **preprocessing is necessary** to improve the behaviour of the FRBCSs.
- Better performance with the over-sampling techniques: **SMOTE family**.

Balance Method	GM_{Tr}	\hat{GM}_{Tst}
None	75.81 \pm 26.23	61.94 \pm 28.52
CNNRb	72.27 \pm 20.27	61.54 \pm 23.09
TomekLinks	79.83 \pm 24.34	67.00 \pm 26.25
OSS	68.70 \pm 20.41	59.81 \pm 23.18
CNN-TomekLinks	57.10 \pm 23.64	50.41 \pm 22.95
NCL	80.17 \pm 23.63	67.70 \pm 26.20
RandomUnderSampling	84.71 \pm 11.36	75.16 \pm 15.46
RandomOverSampling	90.67 \pm 9.69	78.36 \pm 15.45
SMOTE	90.24 \pm 9.96	79.57 \pm 14.74
SMOTE-TomekLinks	88.76 \pm 11.27	79.03 \pm 15.08
SMOTE-ENN	88.79 \pm 10.77	78.97 \pm 15.08

Comparison	R^+	R^-	Hypothesis for $\alpha = 0.1$
SMOTE vs. None	545.5	15.5	Rejected

Test de Wilcoxon para los métodos de Preprocesamiento

Experimental Study. Analysis of the Granularity of the Fuzzy Partitions

- We analyze the performance when 5 and 7 labels are used.
- It is empirically shown that a high number of labels produces **overfitting**.

Number of Labels	GM_{Tr}	GM_{Tst}
5 Labels	90.24 ± 9.96	79.57 ± 14.74
7 Labels	93.26 ± 7.64	73.54 ± 17.55

*Average results for FRBCSs varying the number of fuzzy labels.
SMOTE method is used as preprocessing mechanism*

Comparison	R^+	R^-	Hypothesis for $\alpha = 0.1$
5 Labels vs. 7 Labels	505	56	Rejected

Wilcoxon's Test for the granularity of the fuzzy partitions

Experimental Study. Conjunction Operators, Fuzzy Reasoning Methods and Rule Weights

- We will now study the effect of the conjunction operators (minimum and product T-norm) rule weights and FRMs, fixing SMOTE as the preprocessing mechanism and the number of fuzzy subspaces as 5 labels per variable..

Weight	Conjunction operator	Winning Rule GM_{Tr}	Winning Rule GM_{Tst}	Additive Comb. GM_{Tr}	Additive Comb. GM_{Tst}
CF	Minimum	89.46 ± 10.34	77.90 ± 15.49	88.20 ± 10.76	76.62 ± 17.87
CF	Product	90.83 ± 9.68	78.90 ± 14.87	90.77 ± 9.75	78.32 ± 17.00
P-CF	Minimum	90.02 ± 9.76	78.71 ± 15.15	89.22 ± 9.63	77.82 ± 15.37
P-CF	Product	90.24 ± 9.96	79.57 ± 14.74	90.80 ± 9.72	78.96 ± 15.75
M-CF	Minimum	88.75 ± 10.99	76.63 ± 17.57	83.91 ± 15.40	73.59 ± 17.46
M-CF	Product	90.58 ± 10.83	78.08 ± 17.00	85.03 ± 16.29	72.75 ± 18.98
Total	—	89.98 ± 10.16	<u>78.30 ± 15.67</u>	87.99 ± 12.39	76.34 ± 17.06

Comparison of the average results for FRBCSs with different T-norms, Rule Weights and FRMs. SMOTE method is used as preprocessing mechanism.

Experimental Study. Conjunction Operators, Fuzzy Reasoning Methods and Rule Weights

■ Rankings:

T-norm+Rule Weight	Ranking
Product+P-CF	2.7727
Product+CF	3.04545
Product+M-CF	3.2273
Minimum+P-CF	3.4091
Minimum+CF	4.0606
Minimum+M-CF	4.4848

Rankings obtained through Friedman's test for FRBCS configuration. **FRM of the winning rule**

<i>i</i>	algorithm	<i>z</i>	<i>p</i>	α/i	Hypothesis
5	Minimum+M-CF	3.71742	2.01262E-4	0.0125	R for Product+P-CF
4	Minimum+CF	2.79629	0.00517	0.01667	R for Product+P-CF
3	Minimum+P-CF	1.38170	0.16706	0.025	A
2	Product+M-CF	0.98693	0.32368	0.05	A
1	Product+CF	0.59216	0.55375	0.1	A

Holm's Table for the configuration of the FRBCS.
FRM of the winning rule (FRBCS with product T-norm and P-CF for the rule weight is the control method)

T-norm+Rule Weight	Ranking
Product+CF	2.8485
Product+P-CF	2.8939
Product+M-CF	3.5606
Minimum+CF	3.5909
Minimum+P-CF	3.8030
Minimum+M-CF	4.3030

Rankings obtained through Friedman's test for FRBCS configuration. **FRM of additive combination**

<i>i</i>	algorithm	<i>z</i>	<i>p</i>	α/i	Hypothesis
5	Minimum+M-CF	3.15817	0.00159	0.0125	R for Product+CF
4	Minimum+P-CF	2.07255	0.03821	0.01667	R for Product+CF
3	Minimum+CF	1.61198	0.10697	0.025	A
2	Product+M-CF	1.54619	0.12206	0.05	A
1	Product+P-CF	0.09869	0.92138	0.1	A

Holm's Table for the configuration of the FRBCS.
FRM of Additive Combintation (FRBCS with product T-norm and P-CF for the rule weight is the control method)




Experimental Study. Conjunction Operators, Fuzzy Reasoning Methods and Rule Weights

■ Conclusions obtained from the ranking:

- Regarding the conjunction operator, we can conclude that very good performance is achieved when using the **product T-norm** rather than the minimum T-norm, independently of the rule weight and FRM.
- For the rule weight we may emphasize as good configurations the **P-CF in the case of the FRM of the winning rule and the CF in the case of the FRM with additive combination**. They have a higher ranking, although statistically they are similar to the remaining configurations.
- When comparing both FRMs we observe that they are equal statistically. We select as a good model the one that uses the **winning rule**.

Comparison	R^+	R^-	Hypothesis
WR vs. AC	286	275	Accepted

Test de Wilcoxon para la comparativa del MRD en SCBRDs



Experimental Study. Analysis of the Behaviour of the FRBCSs according to the imbalance degree

- Selected Model: Product T-norm, P-CF for the rule weight and FRM of the winning rule.
- Imbalance Ratio (IR):
 - ☐ Low: Positive instances between the 25 and 40% of the total (from 1.5 to 3)
 - ☐ Medium: Positive instances between the 10 and 25% of the total (from 3 to 9)
 - ☐ High: Positive instances less than the 10% of the total (higher than 9)
- Comparison with:
 - ☐ Ishibuchi Learning Method + SMOTE
 - ☐ E-Algorithm
 - ☐ C4.5+SMOTE

A. Orriols-Puig, E. Bernadó-Mansilla, Evolutionary rule-based systems for imbalanced datasets, Soft Computing In Press (2008)

H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems 13 (2005) 428-435

L. Xu, M.Y. Chow, and L.S. Taylor. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm. *IEEE Transactions on Power Systems*, 22(1):164–171, 2007.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations, 6:1 (2004) 20-29.

Experimental Study. Analysis of the Behaviour of the FRBCSs according to the imbalance degree

■ Data-sets with low imbalance

	FRBCS-Chi SMOTE Pre.		FRBCS-Ish SMOTE Pre.		the E-Algorithm No Preprocessing		C4.5 SMOTE Pre.	
Data-set	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
EcoliCP-IM	98.19	95.56	97	96.7	95.16	95.25	99.26	97.95
Haberman	70.86	60.4	64.36	62.65	8.47	4.94	74	61.32
Iris1	100	98.97	100	100	100	100	100	98.97
Pima	85.53	66.78	71.31	71.1	55.86	55.01	83.88	71.26
Vehicle3	96.36	87.19	66.28	67.82	46.24	43.83	98.95	94.85
Wisconsin	99.72	43.58	96.17	95.78	96.04	96.01	98.31	95.44
Yeast2	72.75	69.66	51.83	51.41	0	0	80.34	70.86
Glass1	74.44	63.69	72.22	69.39	0	0	94.23	78.14
Glass2	77.3	64.91	65.33	59.29	10.24	0	89.74	75.11
Vehicle2	91.18	71.88	64.83	64.89	5.93	3.09	95.5	69.28
Vehicle4	90.22	63.13	63.21	63.12	0	0	94.88	74.34
Average	86.96	71.43	73.87	72.92	37.99	36.19	91.74	80.68
Std. Dev	11.35	16.38	16.21	16.67	42.27	43.43	8.72	13.49

- C4.5 is better than the FRBCS and the E-Algorithm, (Holm Test and Wilcoxon Test*).
- Low classification rates for the FRBCS could be due to the characteristics of the data-sets, not only to the imbalance and overlapping between classes.

* i	algorithm	z	p	α/i	Hypothesis					
3	E-Algorithm	3.46804	0.00052	0.03333	Rejected for C4.5	Comparison	R^+	R^-	Hypothesis	p-value
2	FRBCS-Chi	1.81659	0.06928	0.05	Accepted	FRBCS-Ish vs. C4.5	10	56	Rejected for C4.5	0.041
1	FRBCS-Ishibuchi	1.32116	0.18645	0.1	Accepted					

Sets

29

Experimental Study. Analysis of the Behaviour of the FRBCSs according to the imbalance degree

■ Data-sets with medium imbalance.

	FRBCS-Chi SMOTE Pre.		FRBCS-Ish SMOTE Pre.		the E-Algorithm No Preprocessing		C4.5 SMOTE Pre.	
Data-set	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Ecoli2	93.78	86.05	85.45	85.71	75.34	77.81	96.28	76.1
GlassNW	98.48	85.94	85.68	88.56	82.08	82.09	99.07	90.13
New-Thyroid2	99.58	95.38	90.97	89.02	88.92	88.52	99.21	97.98
New-Thyroid3	99.58	96.34	94.34	94.21	88.94	88.57	99.57	96.51
Page-Blocks	88.64	87.25	32.41	32.16	64.65	64.51	98.46	94.84
Segment	98.19	95.88	42.61	42.47	95.64	95.33	99.85	99.26
Vehicle1	96.26	84.93	76.54	75.94	44.68	39.07	98.97	91.1
Ecoli3	92.9	87.64	87.23	87	71.98	70.35	95.11	91.6
Yeast4	92.01	89.33	79.97	77.06	82.09	81.99	95.64	88.5
Glass7	94.75	91.61	85.78	85.39	80.21	78.54	98.14	88.77
Ecoli4	98.06	78.13	86.42	86.27	90.84	90.23	99.59	83
Average	95.66	88.95	77.04	76.71	78.67	77.91	98.17	90.71
Std. Dev	3.55	5.54	20.23	20.27	14.42	15.69	1.70	6.78

- There is a smaller difference between our model between our selected model for FRBCS (Chi et al.) and C4.5 (Holm Test and Wilcoxon Test*). Thus, the behaviour is improved in this imbalance scenario.

*

i	algorithm	z	p	α/i	Hypothesis
3	E-Algorithm	3.13775	0.00170	0.03333	Rejected for C4.5
2	FRBCS-Ishibuchi	2.47717	0.01324	0.05	Rejected for C4.5
1	FRBCS-Chi	0.66058	0.50888	0.1	Accepted

Comparison	R^+	R^-	Hypothesis	p-value
FRBCS-Chi vs. C4.5	17	49	Accepted	0.155

Experimental Study. Analysis of the Behaviour of the FRBCSs according to the imbalance degree

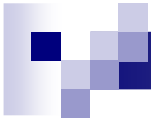
■ Data-sets with High Imbalance.

	FRBCS-Chi SMOTE Pre.		FRBCS-Ish SMOTE Pre.		the E-Algorithm No Preprocessing		C4.5 SMOTE Pre.	
Data-set	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Abalone9-18	71.07	66.47	66.42	65.78	39.67	32.29	95.2	53.19
Abalone19	75.99	66.71	66.93	66.09	0	0	84.31	15.58
Ecoli5	98.12	92.11	89.21	86.92	92.8	92.43	97.67	81.28
Glass3	71.39	49.24	45.25	43.55	27.03	9.87	95.68	33.86
Yeast5	87.94	83.07	75.8	71.36	38.31	32.16	90.76	65
Vowel0	99.64	97.87	89.99	89.03	89.84	89.63	99.67	94.74
YeastCYT-POX	82.35	78.76	74.01	72.83	74.01	72.83	90.93	78.23
Glass5	98.87	81.75	87.03	78.27	84.82	83.38	98.42	83.71
Glass6	98.77	64.33	89.88	89.96	80.6	50.61	99.76	86.7
Yeast5	95.4	93.64	94.93	94.94	88.66	88.17	97.75	92.04
Yeast7	89.57	87.73	88.48	88.42	53.82	51.72	92.15	80.38
Average	88.10	78.34	78.90	77.01	60.87	54.83	94.75	69.52
Std. Dev	11.26	14.99	14.93	15.09	31.02	33.09	4.77	25.38

- Good behaviour of the FRBCS in highly imbalanced data-sets (Holm and Wilcoxon*).
- We observe that the FRBCSs improve their results in comparison with C4.5 when the IR increases. Of course, both methods decrease the geometric mean of true rates when using data-sets with a higher IR.

*

i	algorithm	z	p	α/i	Hypothesis	Comparison	R^+	R^-	Hypothesis	p-value
3	E-Algorithm	3.22032	0.00128	0.03333	Rejected for FRBCS-Chi	FRBCS-Chi vs. C4.5	53	13	Rejected for FRBCS-Chi	0.075
2	C4.5	1.81659	0.06928	0.05	Accepted	FRBCS-Ish vs. C4.5	53	13	Rejected for FRBCS-Ish	0.075
1	FRBCS-Ishibuchi	1.23858	0.21549	0.1	Accepted					



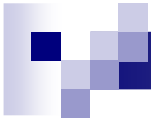
Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions



Lessons Learned and Future Work

- We may emphasize five important lessons learned:
 - The cooperation with pre-processing methods of instances is very positive. We have empirically shown that balancing the classes before the use of the linguistic FRBCS method clearly improves the classification performance. We have found a type of mechanism (**SMOTE**) that provides very good results as a preprocessing technique for FRBCSs. It helps fuzzy methods to become a very competitive model in high imbalanced domains.
 - We have also compared the use of a simple FRBCS obtained with the Chi et al. approach and with the Ishibuchi et al. approach, using a preprocessing step to balance the training set, against an existing ad-hoc fuzzy algorithm for imbalanced data-sets, the E-Algorithm. The first two approaches perform better than the last, showing the necessity of a preprocessing step when dealing with imbalanced data-sets.



Lessons Learned and Future Work

- The analysis of the granularity partitions demonstrates that when increasing the number of fuzzy labels per variable the FRBCSs tend to overfit on the training data.
- We have studied the differences in the application of different conjunction operators, concluding that the **product T-norm** is a good choice for computing the matching degree between the antecedent of the rule and the example
- Comparing the performance of FRBCSs in contrast with the well-known algorithm C4.5, the latter obtains good results when the IR is low or medium, but when this ratio increases then the **FRBCSs are more robust to the class imbalance problem** and in data-sets with high imbalance our approach outperforms C4.5



Lessons Learned and Future Work

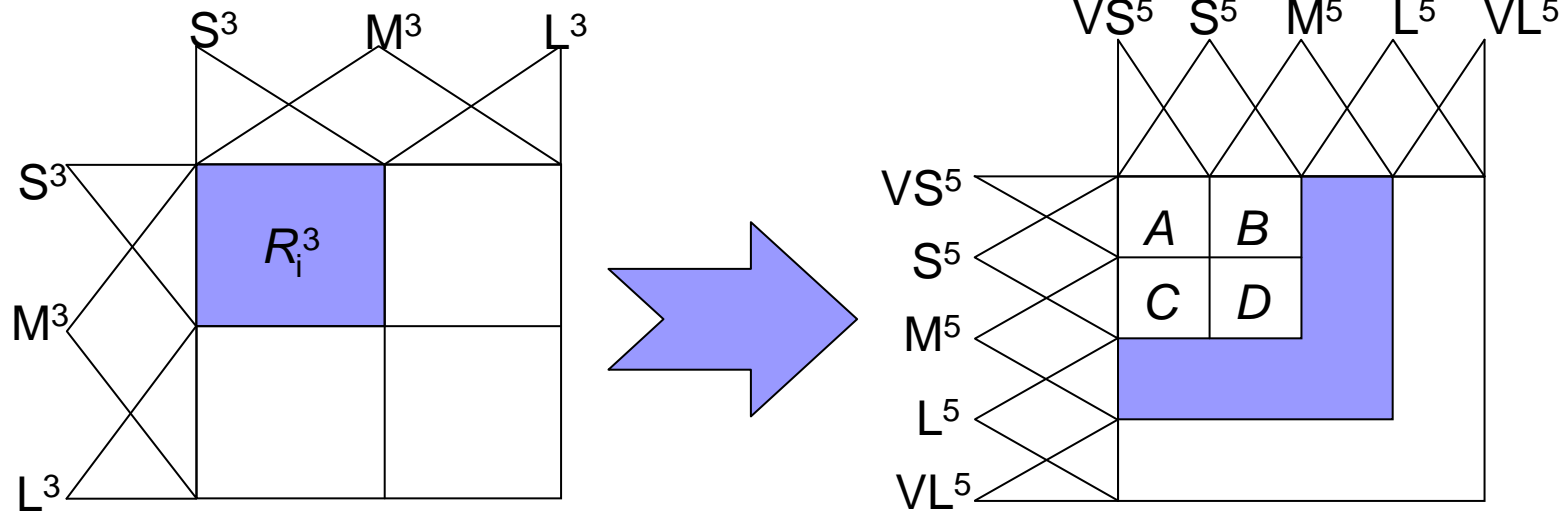
■ Future Work:

- The combination of fuzzy partitions with different granularity can be very useful for learning in FRBCSs.
- Use of an hierarchical system of linguistic rules:
 - It has achieved very good results on regression tasks.

O. Cordon, F. Herrera, I. Zwir: Linguistic Modeling by Hierarchical Systems of Linguistic Rules. IEEE Transactions on Fuzzy Systems 10:1 (2002) 2-20.

Lessons Learned and Future Work

■ Hierarchical System of Linguistic Rules:



$$R_i^3 = \text{IF } x_1 \text{ is } S^3 \text{ AND } x_2 \text{ is } S^3 \text{ THEN } \text{Class} = C \text{ with } RW_i$$

$$R_{i1}^5 = \text{IF } x_1 \text{ is } VS^5 \text{ AND } x_2 \text{ is } VS^5 \text{ THEN } \text{Class} = C \text{ with } RW_{i1}$$

$$R_{i2}^5 = \text{IF } x_1 \text{ is } VS^5 \text{ AND } x_2 \text{ is } S^5 \text{ THEN } \text{Class} = C \text{ with } RW_{i2}$$

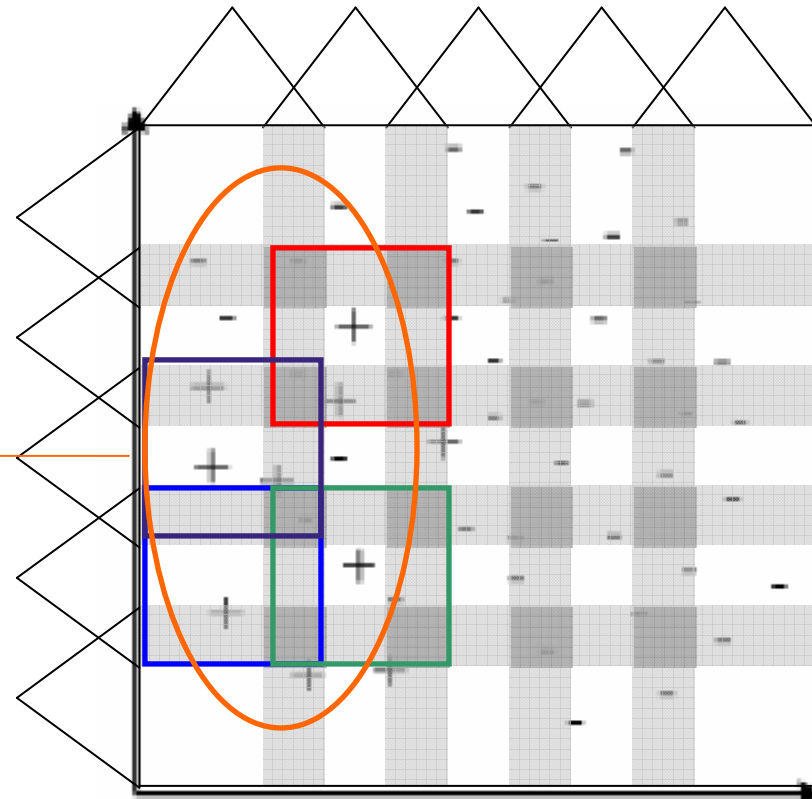
$$R_{i3}^5 = \text{IF } x_1 \text{ is } S^5 \text{ AND } x_2 \text{ is } VS^5 \text{ THEN } \text{Class} = C \text{ with } RW_{i3}$$

$$R_{i4}^5 = \text{IF } x_1 \text{ is } S^5 \text{ AND } x_2 \text{ is } S^5 \text{ THEN } \text{Class} = C \text{ with } RW_{i4}$$

Lessons Learned and Future Work

■ Future Work:

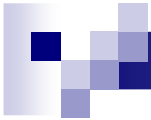
A tuning of the labels can improve the behaviour of the FRBCS



R. Alcalá, J. Alcalá-Fdez, F. Herrera, J. Otero , Genetic Learning of Accurate and Compact Fuzzy Rule Based Systems Based on the 2-Tuples Linguistic Representation. International Journal of Approximate Reasoning 44:1 (2007) 45-64.



Diagram illustrating the construction of a piecewise linear approximation to a function f . The x-axis is marked with points s_0, s_1, s_2, s_3, s_4 . The function f is shown as a solid black line, and its approximation is shown as a blue shaded area. The approximation is defined by the points $(s_0, 0), (s_1, 0.5), (s_2, -0.3), (s_3, 0.5), (s_4, 0)$. The value of the function at s_2 is -0.3 , which is highlighted in a box. The diagram also shows the intervals $[-0.5, 0.5]$ and $[0.5, 1.0]$ on the x-axis.



Some results on the use of Fuzzy Rule Based Systems for Imbalanced Data-sets

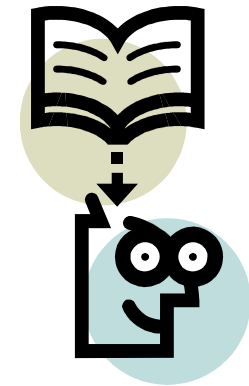
- Introduction to the Problem of Imbalanced Data-sets
- Fuzzy Rule Based Classification Systems
- Data Preparation: Preprocessing Techniques
- Experimental Study
- Lessons Learned and Future Work
- Final Conclusions



Final Conclusions

- In this work we have considered the problem of imbalanced data-sets in classification using linguistic FRBCSs.
- Our results have shown **the necessity of using pre-processing methods of instances** to improve the balance between classes before the use of the FRBCS method.
- We suggest as good components the following ones:
 - Product T-norm as conjunction operator.
 - As rule weight the P-CF heuristic (Penalized Certainty Factor).
- Regarding the FRM there are few differences, and we have chosen the winning rule approach; nevertheless, in the design of a learning method both approaches must be analyzed
- Finally, we have found that the linguistic FRBCSs perform well against the C4.5 decision tree in the framework of highly imbalanced data-sets.

¡Thank you!



¿Any Question?



FRBCSs with Imbalanced Data-Sets



DECSAI
Universidad de Granada