I WORKSHOP
KNOWLEDGE EXTRACTION BASED ON EVOLUTIONARY ALGORITHMS

15-16 May, 2008, Thursday, Friday
ETSI Informática y de Telecomunicación
UNIVERSITY OF GRANADA, GRANADA

Evolutionary prototype selection evaluated with a data complexity measure

J.-R. Cano (jrcano @ujaen.es)

Departament of Computer Science, University of Jaén

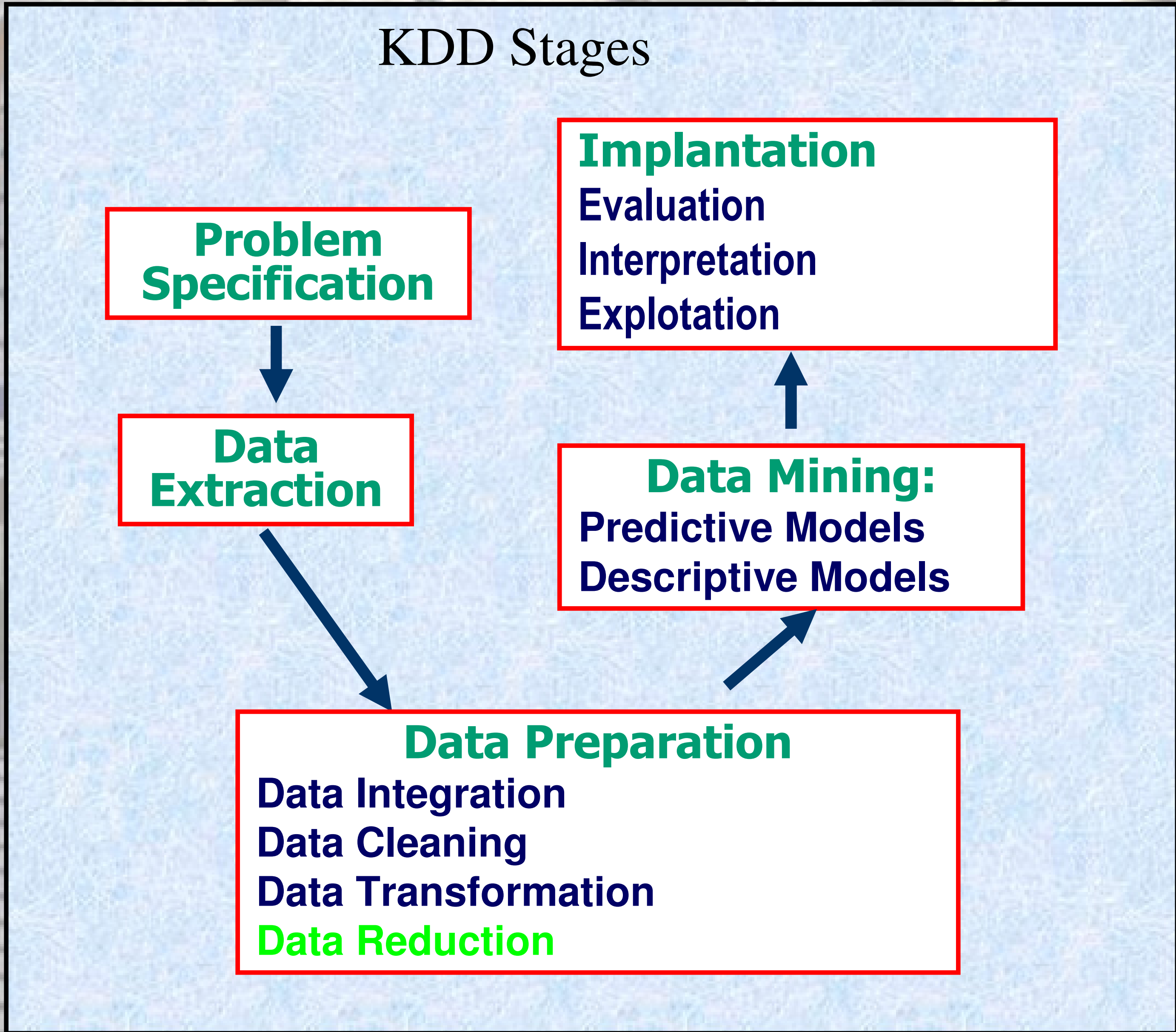
In collaboration with:

S. García and F. Herrera

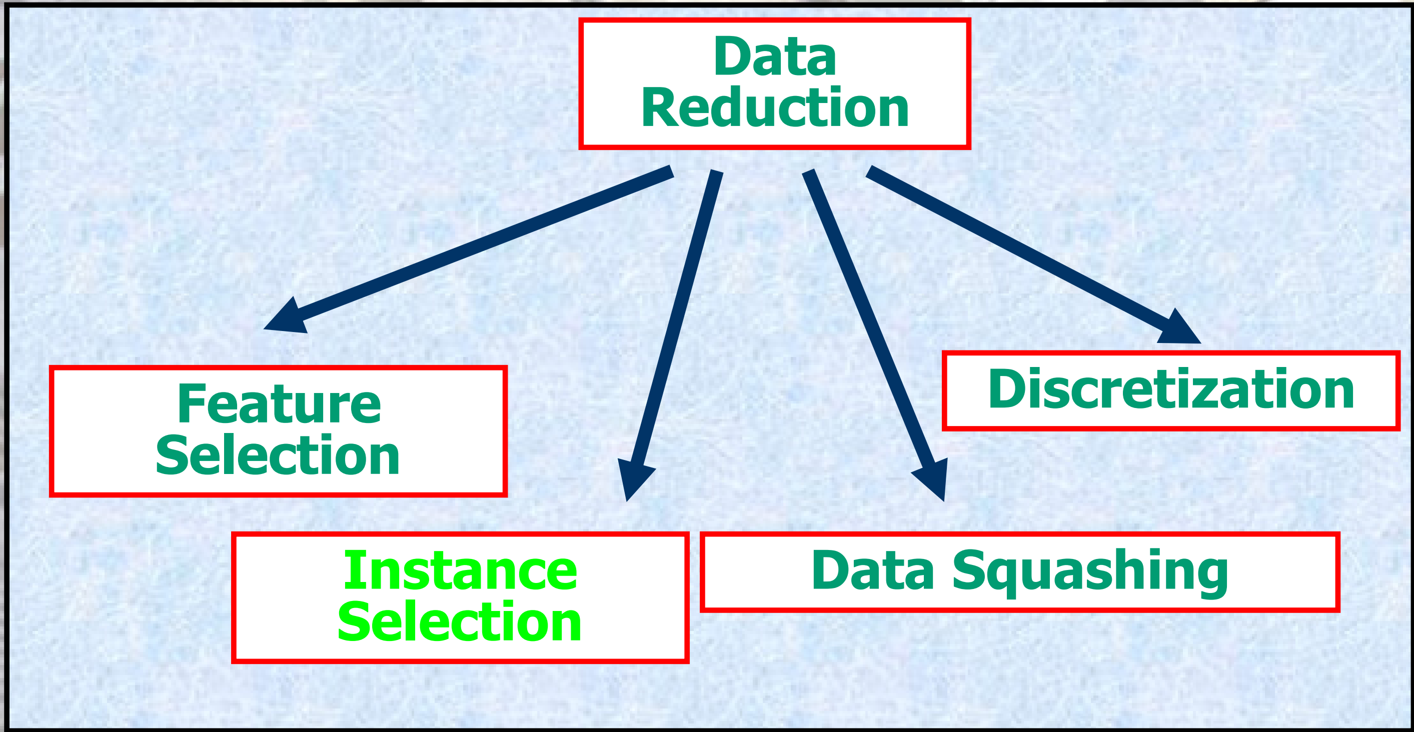
Departament of Computer Science and Artificial Intelligence, Granada

Index:

1.- Instance Selection Introduction:



1.- Instance Selection Introduction:



IS selects the relevant examples for a particular application, increasing the perfomances. AS result, the IS offers::

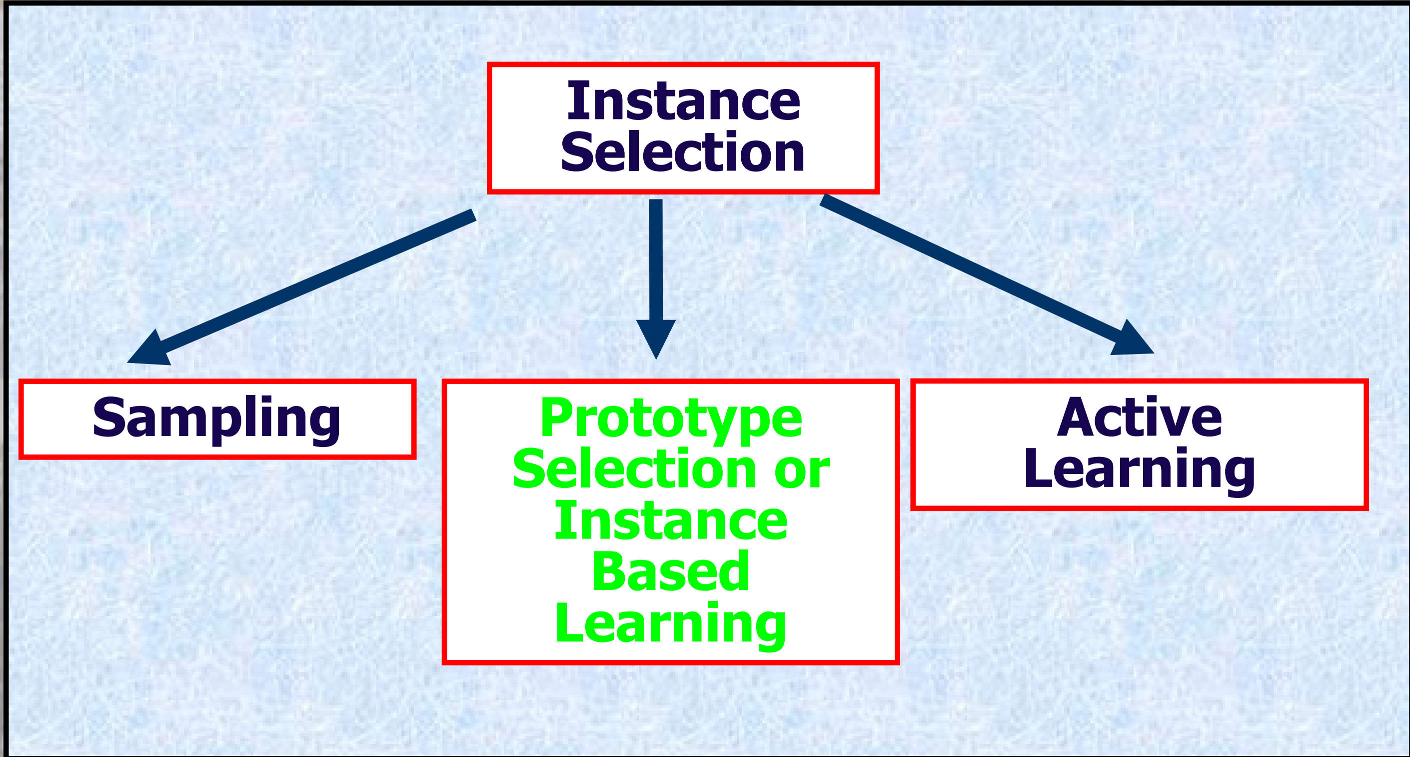
Less data → The algorithms learng faster

More accuracy → The classification algorithm improves its generalization capabilties

More Interpretable results → The results are easier to understand

[Wilson 2000, Reinartz 2002]

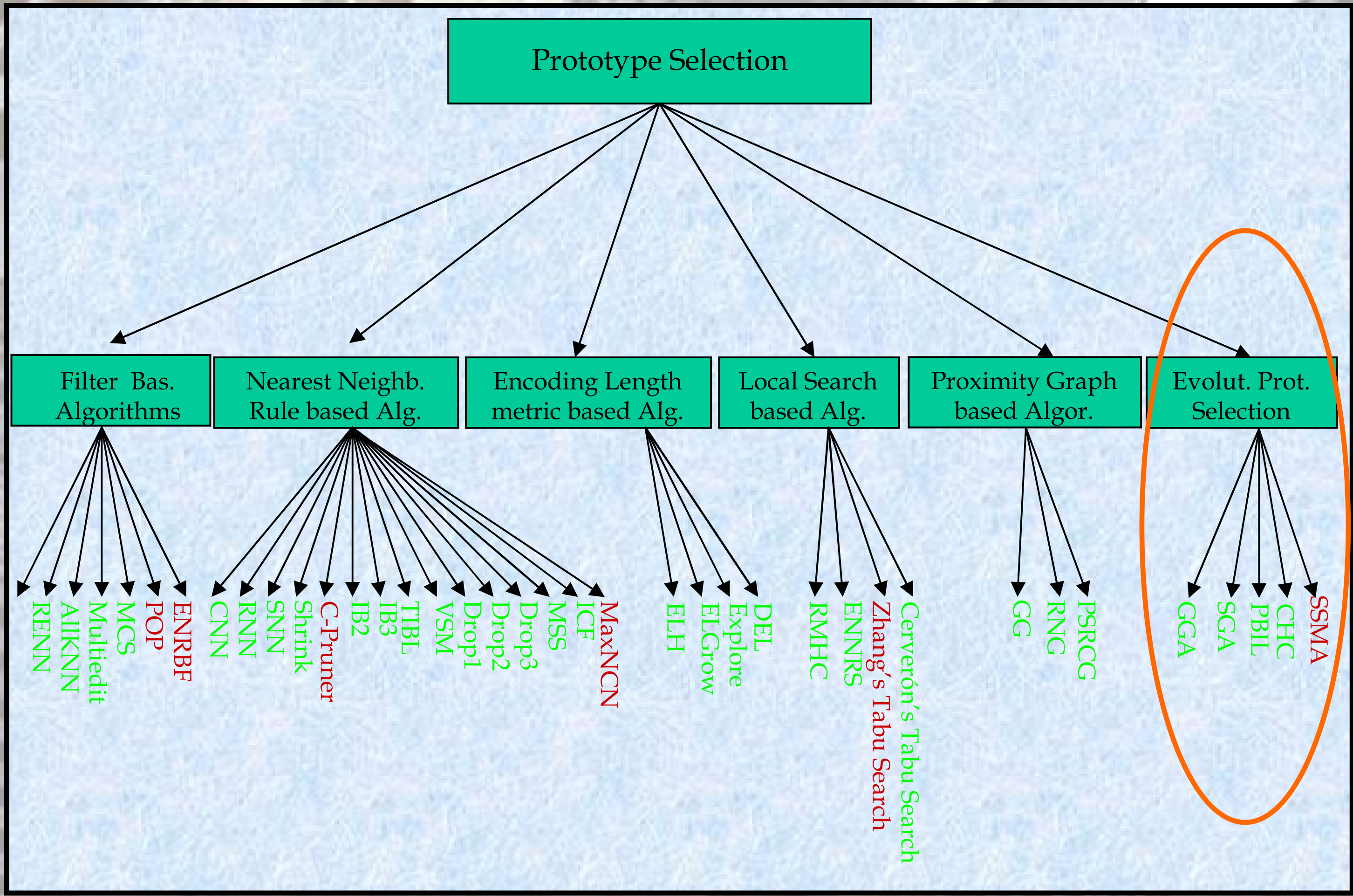
1.- Instance Selection Introduction:



Prototype Selection selects the most representative examples to improve the accuracy of the nearest neighbor classifier.:

1.- Instance Selection Introduction:

Prototype Selection Algorithms



The algorithms marked in green are available in KEEL software. The recent proposals tagged in red will be incorporated soon. (www.keel.es)

2.- Problem Description

	Execution Time(sec)	% Reduction	*1-NN*	
			%Ac Trn	%Ac Test
1-NN	46		94.19	94.18
Cnn (*)	4	97.32	79.33	80.17
Drop1 (*)	254	96.72	78.00	76.85
Drop2 (*)	215	89.57	88.62	88.62
Drop3 (*)	338	96.25	89.03	85.44
Enn	139	5.82	95.43	94.39
Ib2 (*)	2	**97.78**	75.24	75.84
Icf (*)	386	90.04	76.77	76.68
Mcs	101	2.66	**95.96**	**94.38**
Multied	1778	13.99	92.43	92.10
Renn	489	6.47	95.37	94.30
Rnn (*)	13017	96.88	81.27	81.74
Shrink	206	4.89	95.19	**94.38**
Vsm	94	1.04	94.16	94.17
Ib3 (*)	42	71.67	91.49	92.61
Rmhc (*)	34525	90.02	91.59	91.15
Ennsr (*)	37802	90.02	92.79	92.75
GGA	66157	62.53	94.74	93.85
SGA	54656	62.91	95.00	93.67
CHC (*)	8072	**99.29**	93.31	93.53
PBIL (*)	32942	73.13	**96.23**	94.13

From: Table 8, pag. 568. Res. 3 Medium Size Datasets (>5000 instanc.)

J. R. Cano, F. Herrera, M. Lozano (2003). Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: an Experimental Study, IEEE Transactions on Evolutionary Computation 7:6, 561-575.

2.- Problem Description

¿Could it be diagnose the cases when the Evolutionary Prototype Selection is **Effective** to compensate its time cost?

Data Complexity

Analysis

3.- Complexity Measures

- Bayes error-based parametric and nonparametric approaches, entropy measures, nonparametric estimation including k nearest neighbor, Parzen estimation, etc.
- Scatter matrices.
- Information-theory-based approaches.
- Nonparametric methods.
- Overlap between individual attribute values: Fisher's discriminant, volume of overlap region, feature efficiency, etc..
- Measures of separability of classes: Linear Separability, Mixture identificability, etc.
- Measures of Geometry, Topology and Density of manyfolds.
- …

T.K. Ho, M. Basu (2002). Complexity Measures of Supervised Classification Problems, IEEE Trans. on Pattern Analysis and Mach. Intell. 24:3, 289-300.

S. Singh (2003). Multiresolution Estimates of Clasification Complexity, IEEE Transactions on Pattern Analysis and Machine Intelligence 25:12, 1534-1539.

E. Bernadó-Mansilla, T.K. Ho, A. Orriols-Puig (2006). Data Complexity and Evolutionary Learning: Classifier's Behavior and Domain of Competence. In: T.K. Ho, M. Basu (Eds.) Data Complexity in Pattern Recognition, Springer, accepted

3.- Complexity Measures

The studies developed by Mollineda et al. and Sotoca et al. in:

J.M. Sotoca, R.A. Mollineda, J.S. Sánchez (2006). A meta-learning framework for pattern classification by means of data complexity measures. Revista Iberoamericana de Inteligencia Artificial 29, 31-38.

R.A. Mollineda, J.S. Sánchez, J.M. Sotoca (2005). Data Characterization for Effective Prototype Selection. IbPRIA 2005, LNCS 3523, 27-34.

Show that the complexity measures, and concretelly their proposal of multiclass generalization of Fisher's separablity discriminant, as class separablity measure is usefull to distinguish the effective use of classic instace selection algorithms.

3.- Complexity Measures

Fisher's Discriminant: Measure of the overlap between two classes considering the features of the instances.

DEFINITION:

$$f = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$$

where:

- μ_1 and μ_2 are the means of the two classes.
- δ_1^2 and δ_2^2 are the variances of the two classes.

Small values of F1 represents high overlapping.

3.- Complexity Measures

Multiclass Generalization of Fisher's Separablity Discriminant:
Measure of the separability among the classes based in nearest neighbor distance.

DEFINITION:

$$F1 = \frac{\sum_{i=1}^{C} n_i \cdot \delta(m, m_i)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \delta(x_j^i, m_i)}$$

donde:

- C is the number of classes
- n_i the number of instances of class i
- δ ( ) is the metric
- m is the global mean of all the the instances
- m_i is the mean of class i
- X^i_j is the instance j which belongs to class i

Small values of F1 indicate high separability

4.- Evolutionary Protoype Selection: Multiclass Generalization of Fisher's Separability Discriminant.

Datasets:

Table 1: Data Sets.

	Instances	Features	Classes
Bupa	345	6	2
Ecoli	336	7	2
Iris	150	4	3
Glass	214	9	6
Led24digit	200	24	10
Led7digit	500	7	10
Lymphography	148	18	4
Monks	432	6	2
Penbased	10992	16	10
Pima	768	8	2
Wine	178	13	3
Wisconsin	683	9	2
Satimage	6435	36	7
Thyroid	7200	21	3
Zoo	100	16	7

Parameters:

- CNN: Whitout parameters to fix.
- ENN: Number of Neighbors=3.
- EIS-CHC: Evaluations=10000, Population=50, α=0.5

4.- Evolutionary Protoype Selection: Multiclass Generalization of Fisher's Separability Discriminant.

Results:

Table 2: Test accuracy rate and percentage of training instances sorted by the F1 measure.

	F1	Without PS	ENN	CNN	EIS-CHC
Thyroid	0.03	0.93(1)	**0.94**(0.93)	0.88(0.14)	**0.94**(0.01)
Lymphography	0.17	0.35(1)	**0.47**(0.38)	0.27(0.71)	**0.4**(0.04)
Bupa	0.17	0.68(1)	0.66(0.64)	0.63(0.42)	**0.69**(0.03)
Pima	0.22	0.70(1)	**0.71**(0.71)	0.6(0.37)	**0.75**(0.01)
Ecoli	0.24	0.82(1)	**0.88**(0.90)	**0.85**(0.12)	**0.91**(0.01)
Monks	0.36	0.95(1)	0.91(0.91)	0.84(0.23)	**0.99**(0.01)
Led24digit	0.47	0.15(1)	**0.15**(0.42)	**0.25**(0.67)	**0.3**(0.07)
Glass	0.74	**0.57**(1)	0.52(0.68)	0.52(0.41)	0.48(0.05)
Penbased	1.16	**0.99**(1)	**0.99**(0.99)	0.98(0.04)	0.96(0.01)
Led7digit	1.34	0.58(1)	**0.66**(0.74)	0.5(0.45)	**0.64**(0.03)
Wisconsin	1.35	0.94(1)	**0.96**(0.97)	**0.97**(0.07)	**0.94**(0.01)
Zoo	1.38	**0.99**(1)	**0.99**(0.92)	0.9(0.2)	**0.99**(0.09)
Satimage	1.47	**0.90**(1)	**0.90**(0.90)	0.88(0.15)	0.86(0.01)
Wine	1.82	**0.72**(1)	**0.72**(0.96)	**0.72**(0.15)	**0.72**(0.03)
Iris	2.66	0.93(1)	0.93(0.96)	0.93(0.09)	**0.99**(0.01)

5.- Conclusions

- The evaluation of the original dataset using the measure proposed by Mollineda et al. let us to predict the effectivity of the Evolutionary Prototype Selection.

- In data sets with small separability among classes (small measure), the Evolutionary Prototype Selection improves the accuracy capability after they are used.

- When the separability is medium or high, the effectivity of the Evolutionary Prototype Selection appears in the reduction rate but not clearly in the accuracy.

References:

[García et al. 2007] S. García, J.R. Cano, F. Herrera (2007). Un algoritmo memético para la selección de prototipos: Una propuesta eficiente para problemas de tamaño medio. Proceedings del Congreso Español sobre Metaheuristicas, Algoritmos Evolutivos y Bioinspirados, 563-570.

[Grochowski et al. 2004] M. Grochowski, N. Jankowski (2004). Comparison of instance selection algoritms II: Resultsand Comments. Proceeding of ICAISC 2004, 580-585.

[Lozano et al. 2003] M. Lozano, J.S. Sánchez, F. Pla (2003). Reducing training sets my ncn-based explanatory procedures, Proceedings of the First iberian conference on pattern recognition and image analysis, LNCS 2652, 453-461.

[Reinartz 2002] T. Reinartz (2002). A Unifying View on Instance Selection. Data Mining and Knowledge Discovery 6, 191-210.

[Riquelme et al. 2003] J.C. Riquelme, J.S. Aguilar, M. Toro (2003). Finding representative patterns with ordered projections, Pattern Recognition 36,1009-1018.

[Wilson et al. 2000] D.R. Wilson, T.R. Martinez (2000). Reduction tecniques for instance-based learning algorithms. Machine Learning 38 (2000) 257-268.

[Zhao et al. 2003] K.P. Zhao, S.G. Zhou, J.H. Guan, Y. Zhou (2003). C-Pruner: An improved instance prunning algorithm, Proceedings of the Second International Conference on Machine Learning and Cybernetics, 94-99.

[Zhang et al 2000] H. Zhang, G. Sun. Optimal reference subset selection for nearest neighbor classification by tabu search, Pattern Recognition 35, 1481-1490.

References:

[Devroye 1982] L. Devroye (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size, IEEE Trans Pattern Analysis and Machine Intelligence 4:2, 154–157

[Ho et al. 2002] T.K. Ho, M. Basu (2002). Complexity Measures of Supervised Classification Problems, IEEE Trans. on Pattern Analysis and Mach. Intell. 24:3, 289-300.

[Singh 2000] S. Singh (2003). Multiresolution Estimates of Clasification Complexity, IEEE Transactions on Pattern Analysis and Machine Intelligence 25:12, 1534-1539.

[Sotoca et al. 2000] J.M. Sofoca, R.A. Mollineda, J.S. Sánchez (2006). A meta-learning framework for pattern classification by means of data complexity measures. Revista Iberoamericana de Inteligencia Artificial 29, 31-38.

[Mollineda et al. 2005] R.A. Mollineda, J.S. Sánchez, J.M. Sotoca (2005). Data Characterization for Effective Prototype Selection. IbPRIA 2005, LNCS 3523, 27-34.

References:

[Bernadó et al. 2005] E. Bernadó, T.K. Ho (2005). Domain of competence of XCS classifier system in complexity measurement space. IEEE Transactions on Evolutionary Computation 9:1, 82-104.

[Ho et al. 1994] T.K. Ho, H.S. Baird (1994). Estimating the intrinsic difficulty of a recognition problem, Proceedings of the 12th International Conference on Pattern Recognition, 178-183.

[S.-Y. Sohn 1999] S.-Y. Sohn (1999). Meta Analysis of Classification Algorithms for Pattern Recognition, IEEE Trans. on Pattern Analysis and Mach. Intell. 21:11, 1137-1144.

[Michie et al.] D. Michie, D.J. Spiegelhalter, C.C. Taylor (1994). Machine Learning, Neural and Statistical Classification, Ellis Horwood.