I WORKSHOP
KOWLEDGE EXTRACTION BASED ON EVOLUTIONARY ALGORITHMS

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

Some studies on the construction of artificial data sets for some data complexity measures

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.- Introduction.

2.- Overlap Complexity Measures Anaylisis.

  2.1. Fisher's Discriminant Ratio

  2.2. Volume of Overlap Region

  2.3. Feature Efficiency

3.- Conclusions

4.- Future works.

1.- Introduction:

The aim of the study consists on the analysis of some complexity measures by means of the generation of artificial data sets.

To address this, we normalize the measures in a [0,1] range, generating artificial data sets covering that range and studying graphically the instances distribution and some classificator behaviour.

2.- Overlap 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

2.1. Fisher's Discriminant Ratio

Measure: **Fisher's Discriminant Ratio** (F1)
Behaviour: Small values indicate High overlap
Cites: [Bernadó et al. 2005] , [Dong et al. 2003] , [Hernandez et al. 2005], [Ho et al. 2000], [Ho et al. 2002a], [Ho et al. 2002b], [Ho et al. 2006], [Mollineda et al 2005], [Sotoca et al. 2006]
Definition: For each feature, the measure f is calculated as: $$f = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$$ where μ_1, μ_2, ${\sigma_1}^2$, ${\sigma_2}^2$ are the means and variances of the two classes respectively. It is used the maximum over all the feature dimensions to describe a problem.

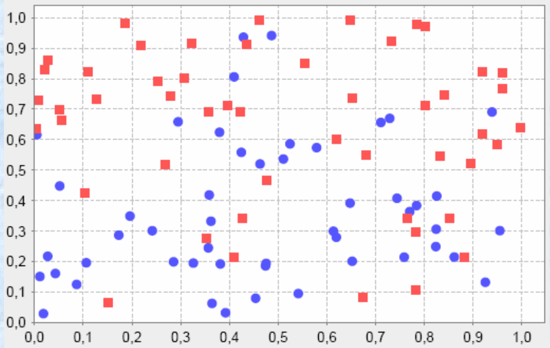
2.1. Fisher's Discriminant Ratio

Measure: **Fisher's Discriminant Ratio** (F1)

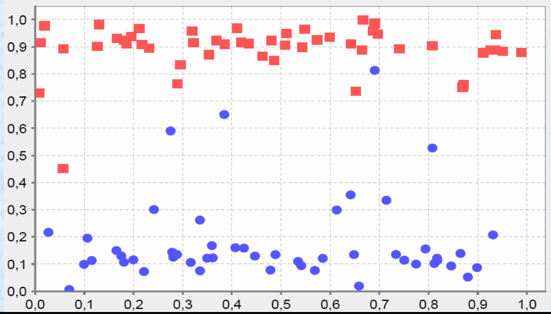
Artificial datasets:

Instances=100, Features=2, Classes=2

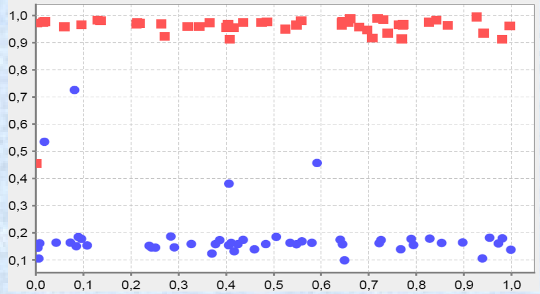
F1=0



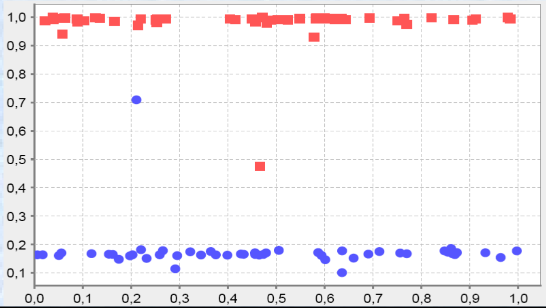
F1=0.25



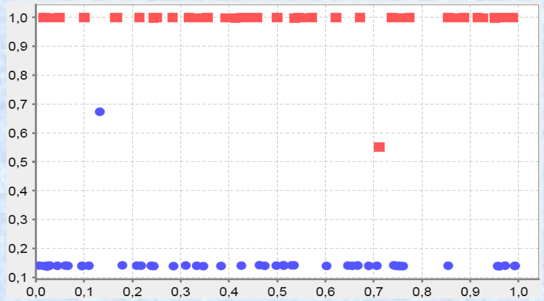
F1=0.5



F1=0.75



F1≈1.0

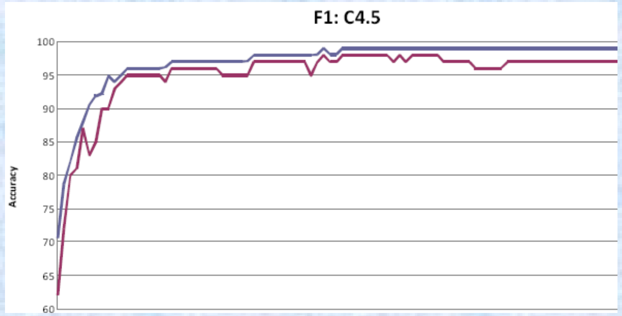


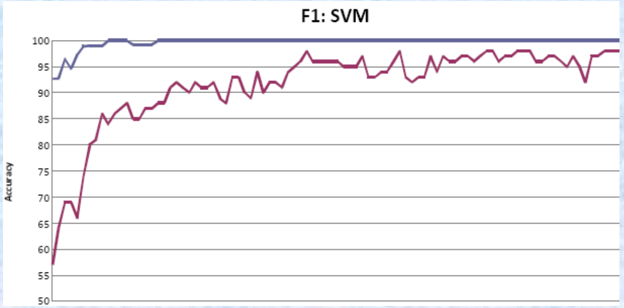
2.1. Fisher's Discriminant Ratio

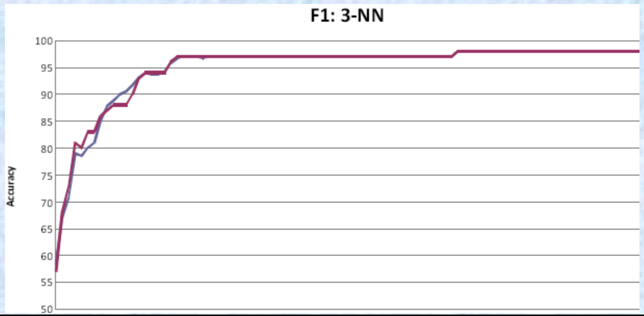
Measure: **Fisher's Discriminant Ratio** (F1)

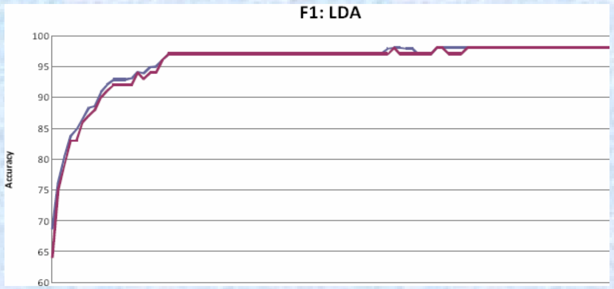
Classificator Behaviour with Artificial Datasets:

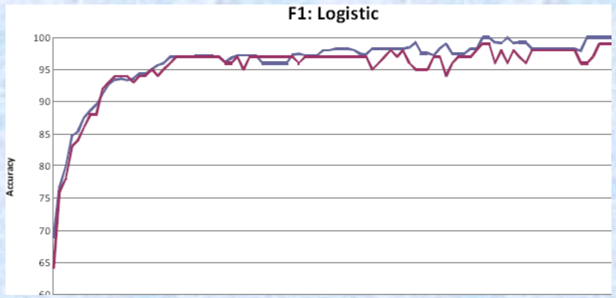
Instances=100, Features=4, Classes=2











2.1. Fisher's Discriminant Ratio

Measure: **Fisher's Discriminant Ratio** (F1)

Generalization (Multiclass extension):

Considering [Ho et al. 2006]:

$$f = \frac{\sum_{i=1,j=1,i\neq j}^{C} p_i p_j (\mu_i - \mu_j)^2}{\sum_{i=1}^{C} p_i \sigma_i^2}$$

- Considering [Mollineda et al 2005], [Sotoca et al 2006]:

They propose a measure of the separability among the classes based in nearest neighbor distance.

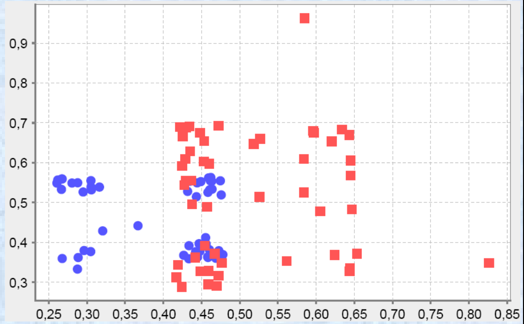
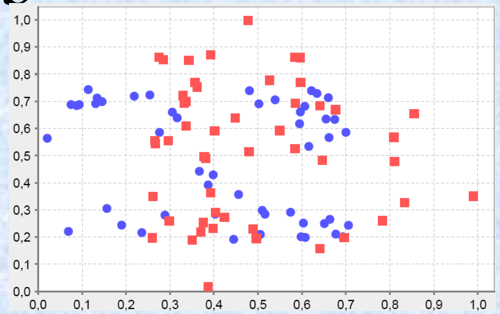
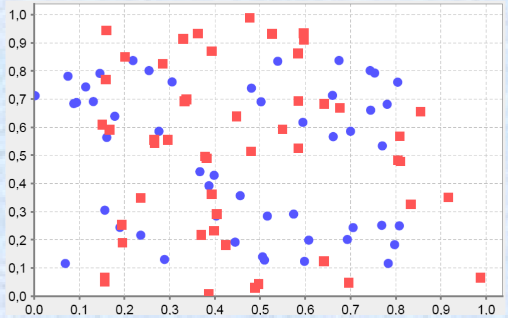
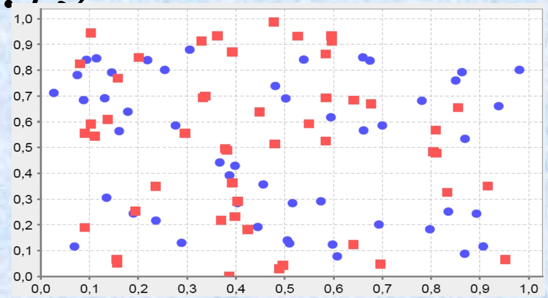
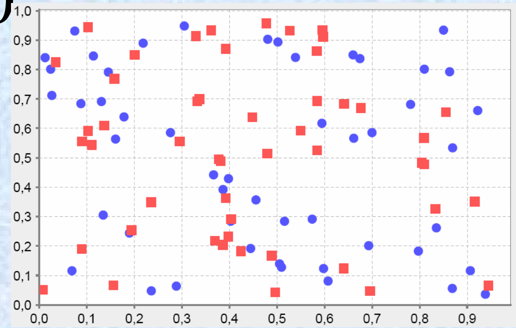
$$F1 = \frac{\sum_{i=1}^{C} n_i \cdot \delta(m, m_i)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \delta(x_j^i, m_i)}$$

where n_i denotes the number of samples in class i, δ is a metric, m is the overall mean, m_i is the mean of class i, and x_j^i represents the sample j belonging to class i.

2.2. Volume of Overlap Region

Measure: **Volume of Overlap Region** (F2)
Behaviour: Small value indicate Small overlap
Cites: [Bernadó et al. 2005] , [Dong et al. 2003] , [Hernandez et al. 2005], [Ho et al. 2000], [Ho et al. 2002a], [Ho et al. 2002b], [Ho et al. 2006], [Mollineda et al 2005], [Sotoca et al 2006]
Definition: $$\mathrm{F2} = \prod_i \frac{MIN(max(f_i, c_1), max(f_i, c_2)) - MAX(min(f_i, c_1), min(f_i, c_2))}{MAX(max(f_i, c_1), max(f_i, c_2)) - MIN(min(f_i, c_1), min(f_i, c_2))}$$.

2.2. Volume of Overlap Region

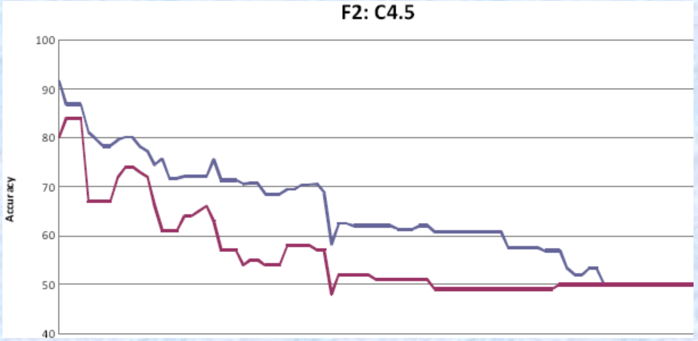
Measure: **Volume of Overlap Region** (F2) *Artificial datasets:* Instances=100, Features=2, Classes=2	F2≈0 
F2=0.25 	F2=0.5 
F2=0.75 	F2=1.0 

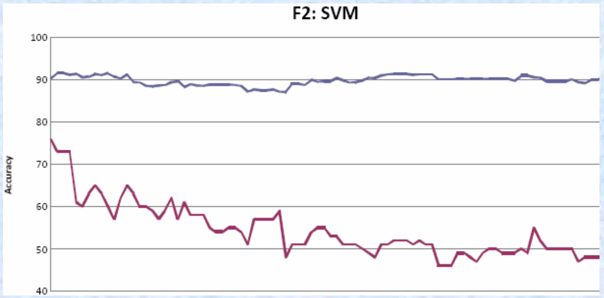
2.2. Volume of Overlap Region

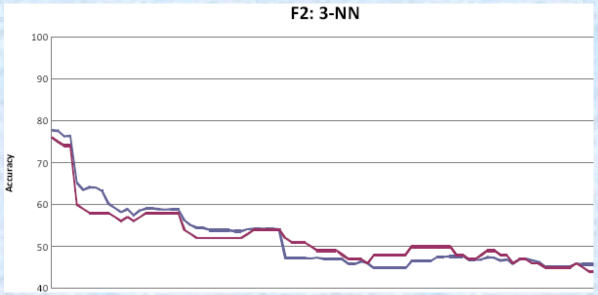
Measure: **Volume of Overlap Region** (F2)

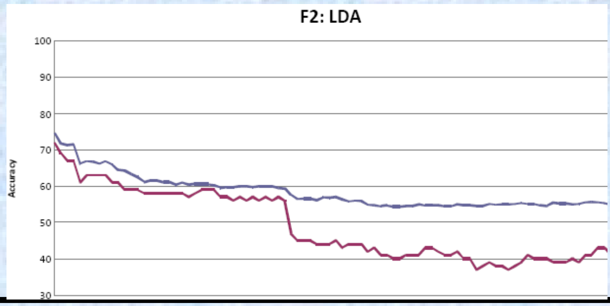
Classificator Behaviour with Artificial Datasets:

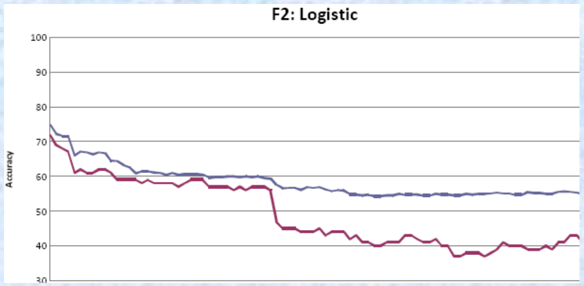
Instances=100, Features=4, Classes=2











2.2. Volume of Overlap Region

Measure: **Volume of Overlap Region** (F2)

Generalization (Multiclass extension):

• Considering [Ho et al. 2006], [Mollineda et al 2005], [Sotoca et al 2006]:

$$f = volume \quad of \sum_{i,j,i \neq j} V_i \cap V_j$$

Being V_i the hyperrectangular region spanned by the *i*th class.

2.3. Feature Efficiency

Measure: **Feature Efficiency** (F3)
Behaviour: Small values indicate High overlap
Cites: [Bernadó et al. 2005] , [Dong et al. 2003] , [Hernandez et al. 2005], [Ho et al. 2000], [Ho et al. 2002a], [Ho et al. 2002b], [Ho et al. 2006], [Mollineda et al 2005], [Sotoca et al 2006]
Definition: The efficiency of each feature is the fraction of all remaining points separable by that feature. It is used the maximum feature efficiency to represent the contribution of the feature most usefull.

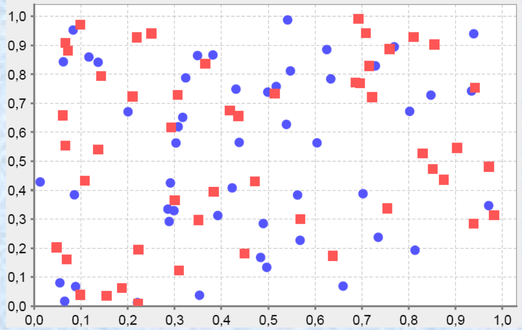
2.3. Feature Efficiency

Measure: **Feature Efficiency** (F3)

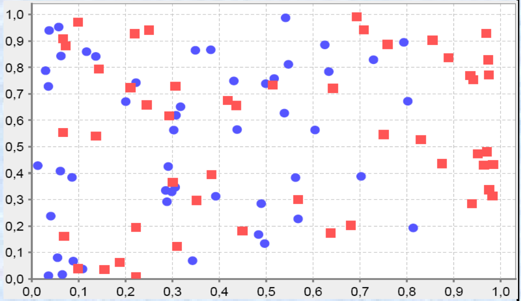
Artificial datasets:

Instances=100, Features=2, Classes=2

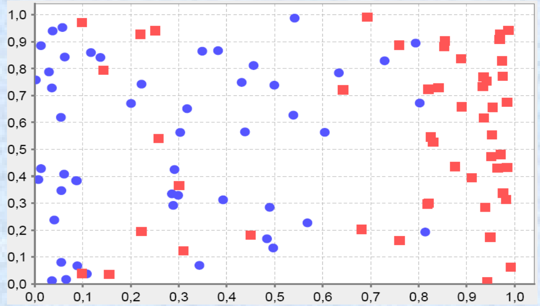
F2=0



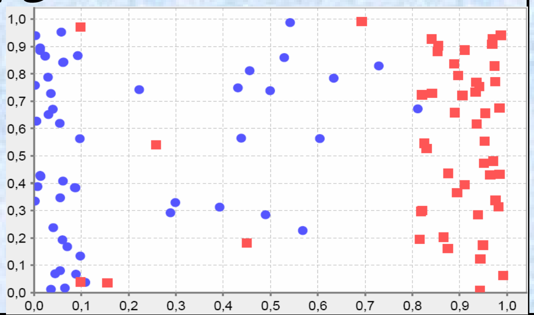
F2=0.25



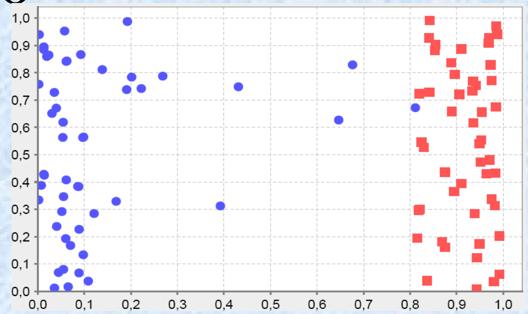
F2=0.5



F2=0.75



F2=1.0

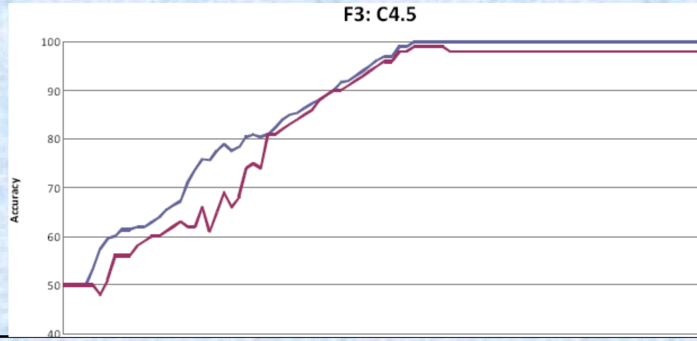


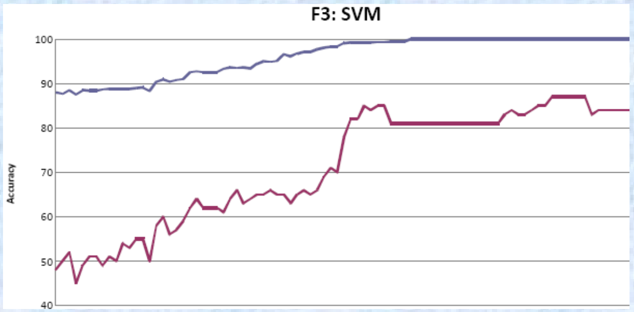
2.3. Feature Efficiency

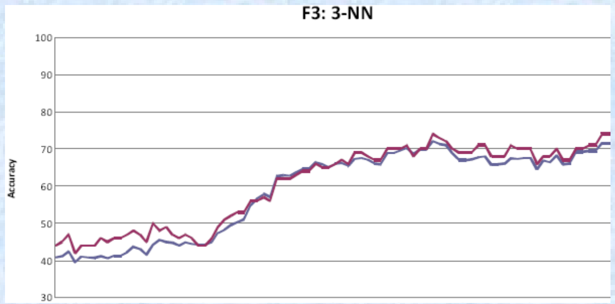
Measure: **Feature Efficiency** (F3)

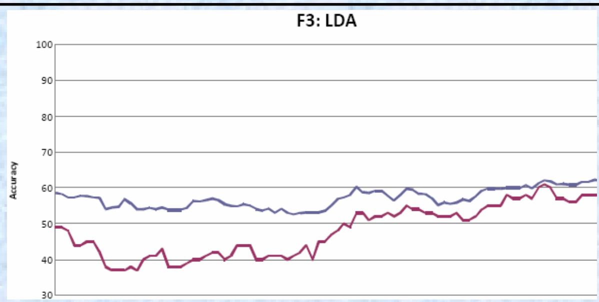
Classificator Behaviour with Artificial Datasets:

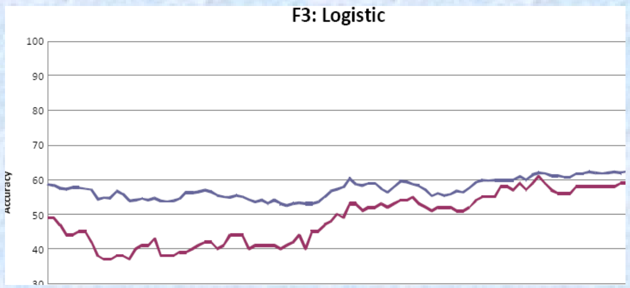
Instances=100, Features=4, Classes=2











2.3. Feature Efficiency

Measure: **Feature Efficiency** (F3)

Generalization (Multiclass extension):

- Considering [Ho et al. 2006], [Mollineda et al 2005], [Sotoca et al 2006]:

"The measure value for C classes is the overall fraction of points in some overlap range of any feature for any pair of classes. Points in more than one range is counted once."

3.- Conclusions

- **The graphical instance distribution helps to understand the effect of the measures variation in instances distribution.**

- **SVM seems to be very sensible to overlapping.**

- **C4.5, when F3 is higher, is the classifier which offers the most interesting behaviour.**

- **It would be needed a more complex environment (with higher number of instances, features and classes) for the artificial data sets to analyze the rest of the classifiers.**

4.- Future Works.

- **Increase the environment (complexity) of the artificial data sets.**
- **Increase the number of measures considered.**
- **Extend the measures and their analysis to multiclass context.**

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.

[Basu et al. 2006] M. Basu, T.K. Ho (2006). Data Complexity in Pattern Recognition. Series: Advanced Information and Knowledge Processing. Springer.

[Ho et al. 2006] T.H. Ho, M. Basu, M.H. Chung (2006). Measures of Geometrical Complexity in Classification Problems. In: Data Complexity in Pattern Recognition, Eds. M. Basu, T.K. Ho, pag. 3-25.

[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.