I Workshop on Knowledge Extraction based on Evolutionary Learning

Granada, May 16th, 2008

Low Quality Data

Luciano Sánchez
Universidad de Oviedo

Part I

Introduction to Low Quality Data

Summary

- ***Part I: Introduction to Low Quality Data***
  - *Future directions*
  - *Novel types of data (issues with the representation, fuzzy random variables)*
  - *Preprocessing*
  - *Learning (Inference, Fitness, Genetic optimization)*
  - *Validation*
- ***Part II: Some results on the use of Evolutionary Algorithms for knowledge extraction from low quality data***

Future directions in 2005

- *Trade-off interpretability vs. precision. Use of MOGAs*
- *FRBS for high dimensional problems*
- *GFSs in Data Mining and Knowledge Discovery*
- *Learning genetic models based on* ***vague data***

Herrera, F. **Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions**. International Journal of Computational Intelligence Research. 1 (1). 2005. 59-67

Future directions in 2008

- Multiobjective genetic learning of FRBSs: interpretability-precision
- GA-based techniques for mining fuzzy association rules and data mining
- Learning genetic models based on **low quality data**
- *Genetic learning of fuzzy partitions and context adaptation*
- *Genetic adaptation of inference engine components*
- *Revisiting Michigan style GFSs*

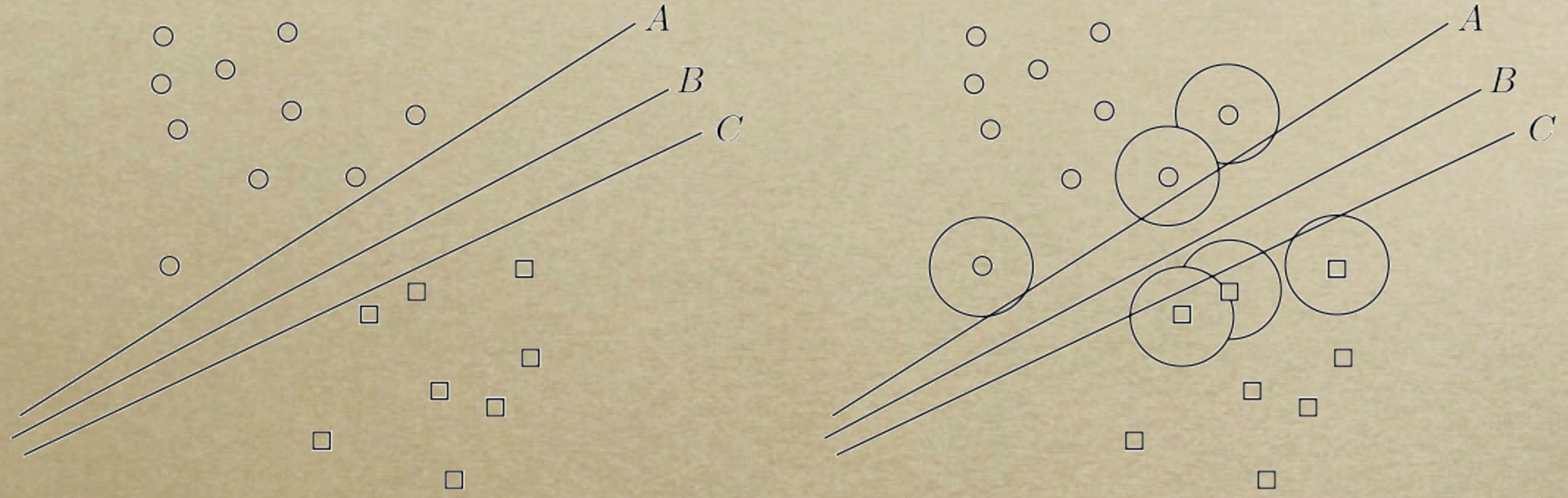
Herrera, F. Genetic Fuzzy Systems: **Taxonomy, Current Research Trends and Prospects**. Evolutionary Intelligence 1 (2008) 27-46

Rationale behind low quality data

- *Crisp data-based GFS are standard statistical classifiers and models:*
    - *Genetic fuzzy classifiers minimize a biased estimation of the classification error (the training error)*
    - *Genetic fuzzy models minimize an estimate of the squared error of a model*
- *There might be theoretical differences if artificial imprecision is added to crisp data and a fuzzy fitness-based GFS is used*

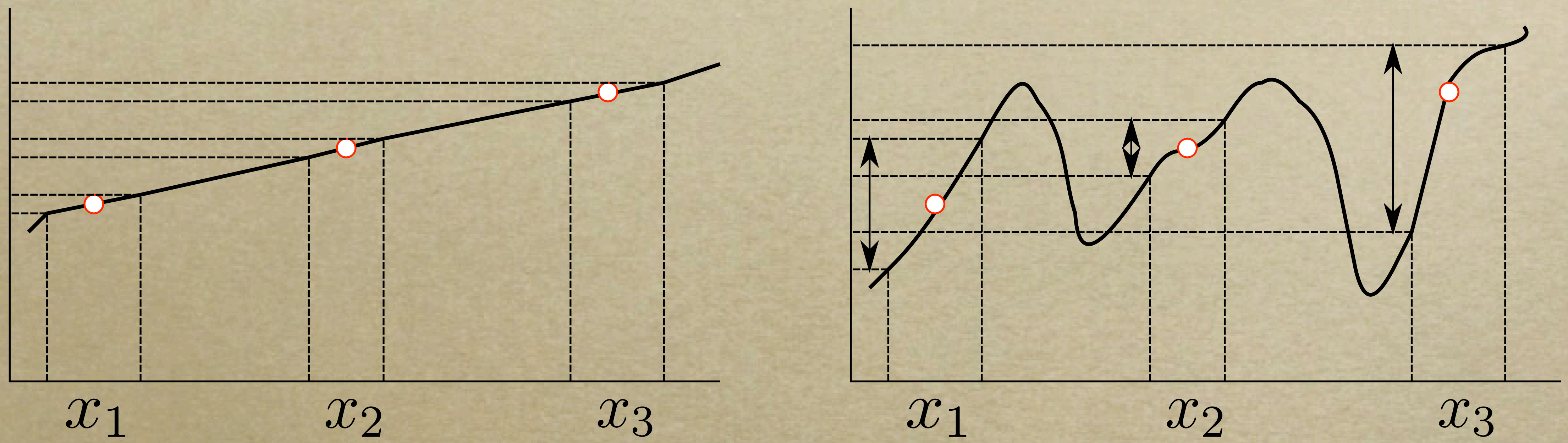
[4] Sánchez, L., Couso, I., **Advocating the use of imprecisely observed data in Genetic Fuzzy Systems**. IEEE Trans Fuzzy Sys 15(4). 2007. 551-562

Addition of imprecision to crisp data (I)



- *There might be a relation between fuzzy fitness-based GF classifiers and SVM.*

Addition of imprecision to crisp data (II)

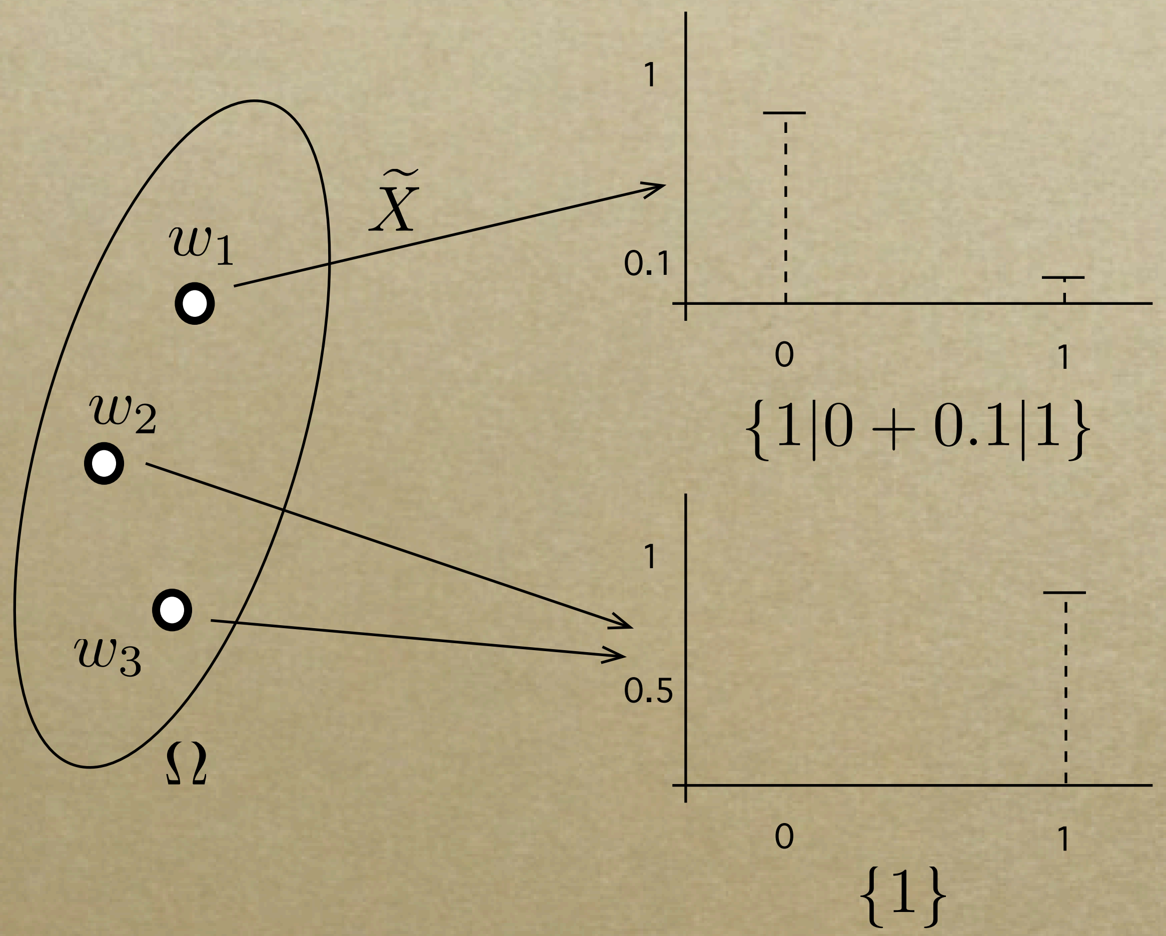


- *The same happens with models - the more regular the model, the more specific its fuzzy fitness is, pointing again to relations between regularized models and fuzzy fitness-based GF models.*

Further uses of low quality data

- ***Imprecise / low quality***
  - *Imprecisely measured data*
  - *Coarsely discretized data, censored data*
  - *Missing values*
- ***Novel representations***
  - *Crisp data + tolerance*
  - *Aggregated values, lists, conflicting data*
  - *Addition of imprecision to crisp data*

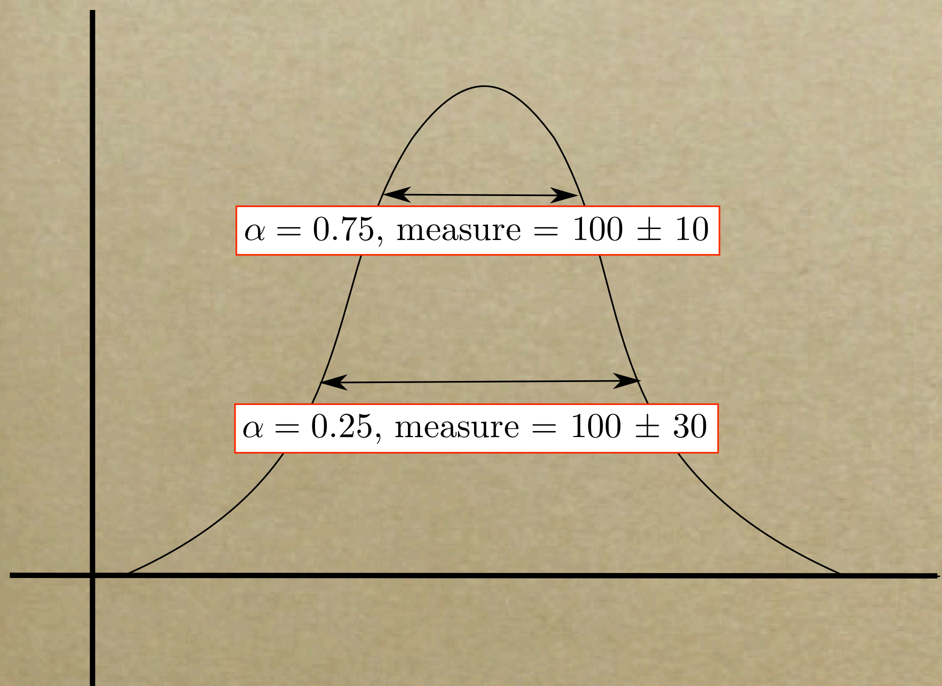
Fuzzy Random Variables

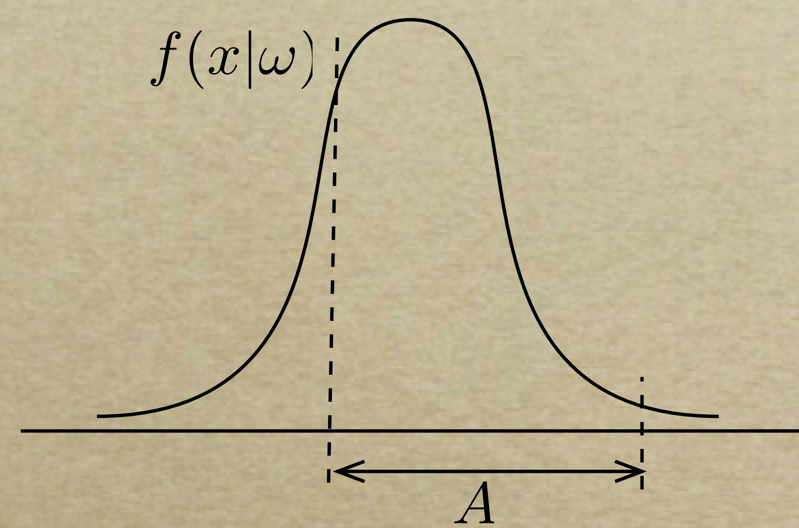


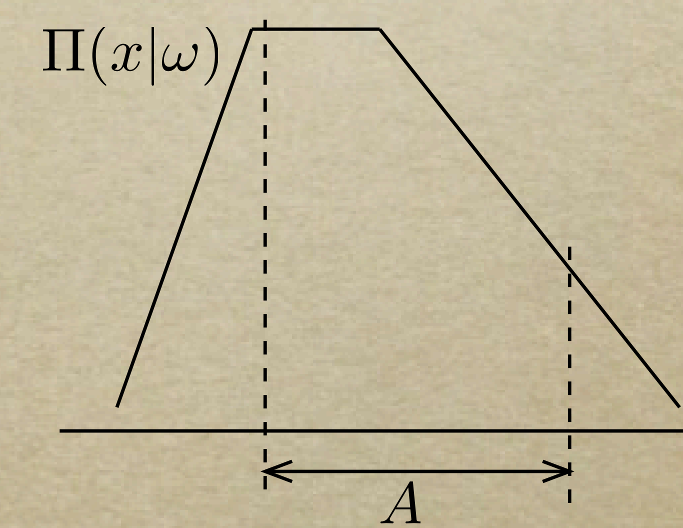
- *Classical model*
- *Second order model*
- *First order, imprecise probabilities-based*

[8] Couso, I., Sánchez, L., **Higher order models for fuzzy random variables**. Fuzzy Sets and Systems 159, 3, 237-258, 2008

Representation



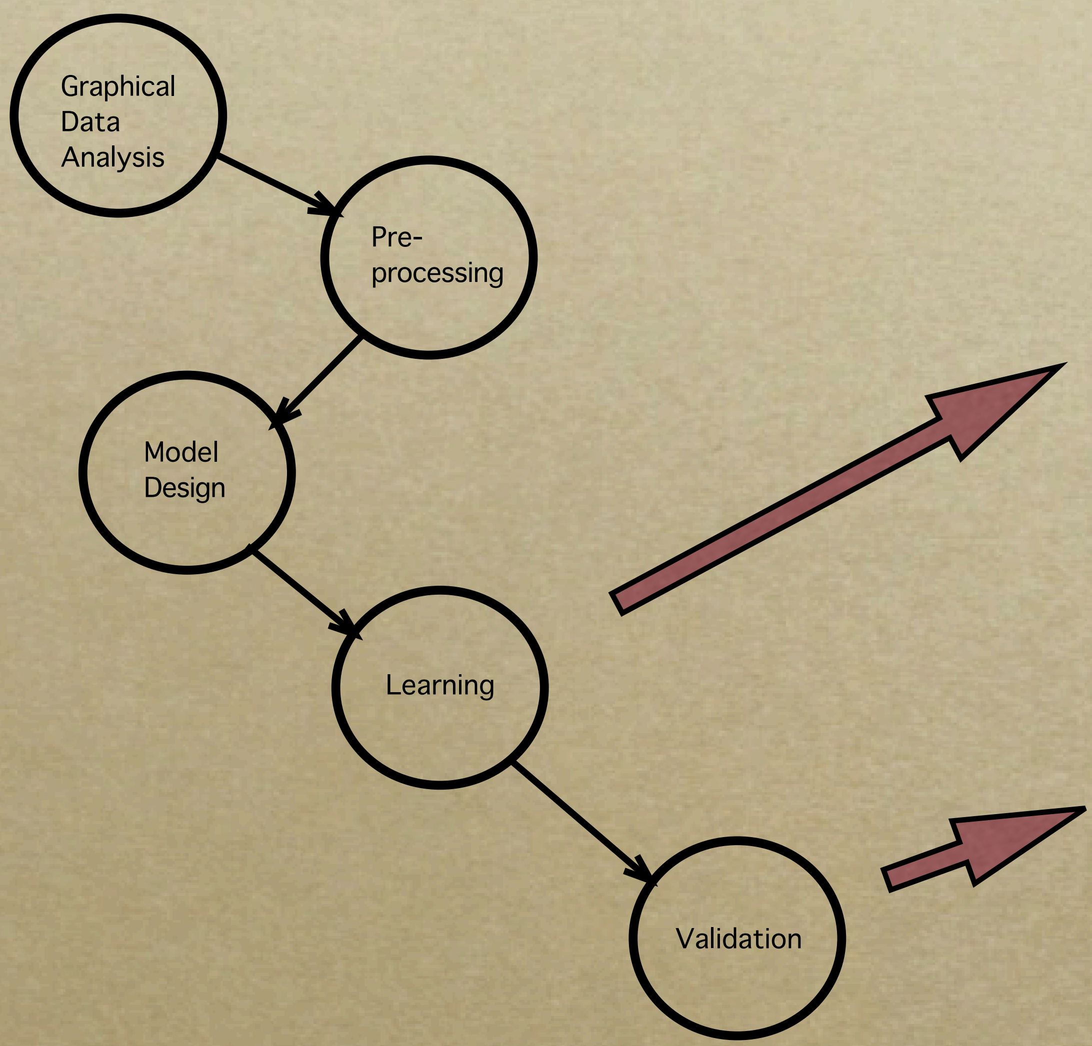




$$P(A) = \int_A f(x|\omega)dx \qquad P(A) \leq \max_A \Pi(x|\omega)$$

- *The first order model has a possibilistic interpretation, coherent with the view of a fuzzy set as a familiy of confidence intervals*

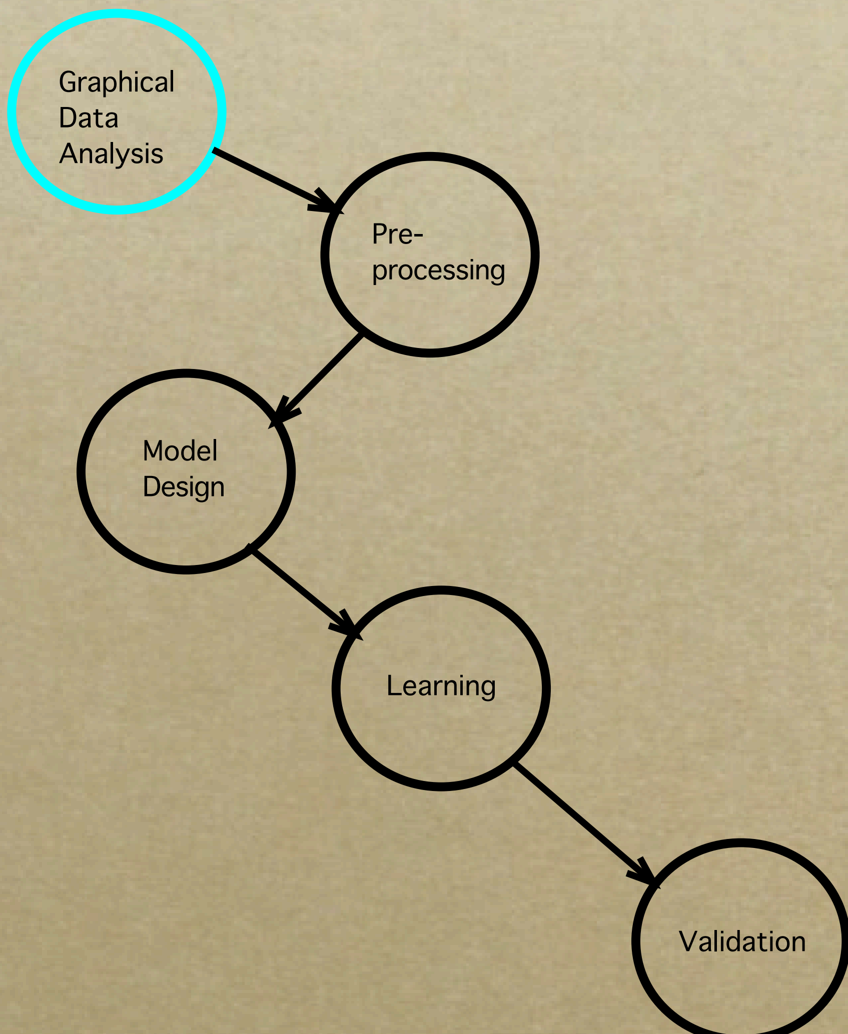
Critical considerations

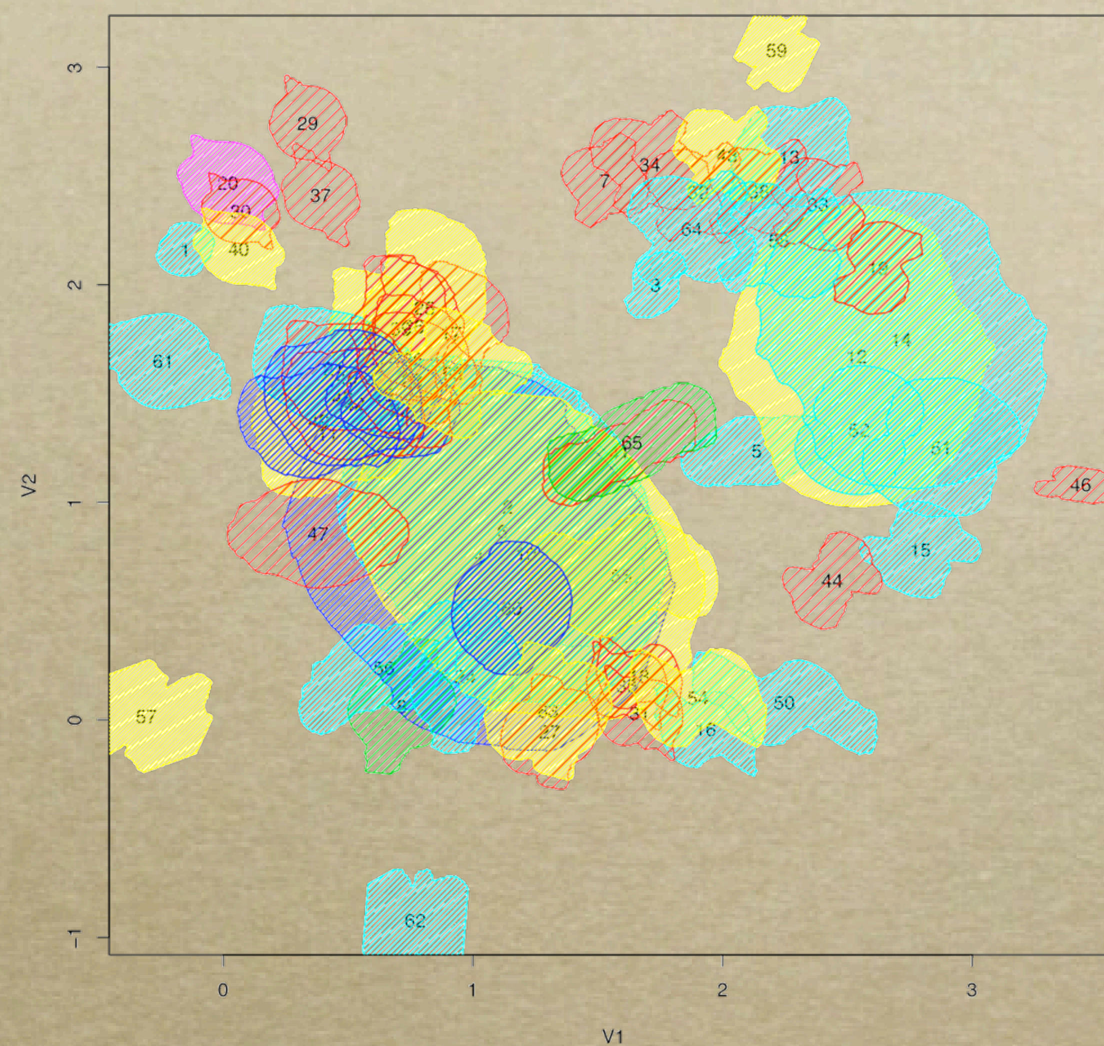


- **EAs used in GFS**
  - *Simple GAs*
  - *The use of novel EAs*
- **Experimental study**
  - *Benchmark problems and reproducibility*
  - *Lack of experimental statistical analysis*
  - *Comparison with the state of the art*

Herrera, F. Genetic Fuzzy Systems: **Status, Critical Considerations and Future Directions.** International Journal of Computational Intelligence Research. 1 (1). 2005. 59-67

Graphical Analysis

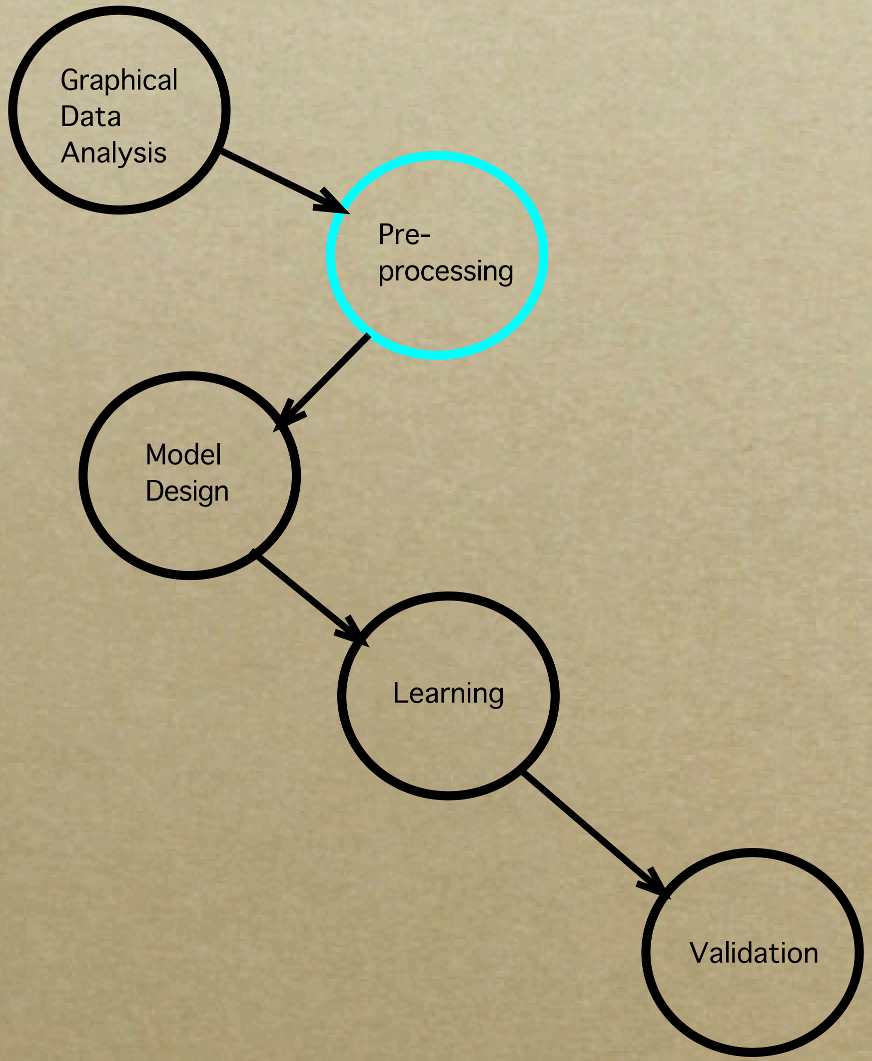




- *PCA, ICA, MDS, for interval and fuzzy data*

[9] Sánchez, L., Palacios, A., Suárez, M. R., Couso, I. **Graphical exploratory analysis of vague data in the early diagnosis of dyslexia**. IPMU 08. Málaga.

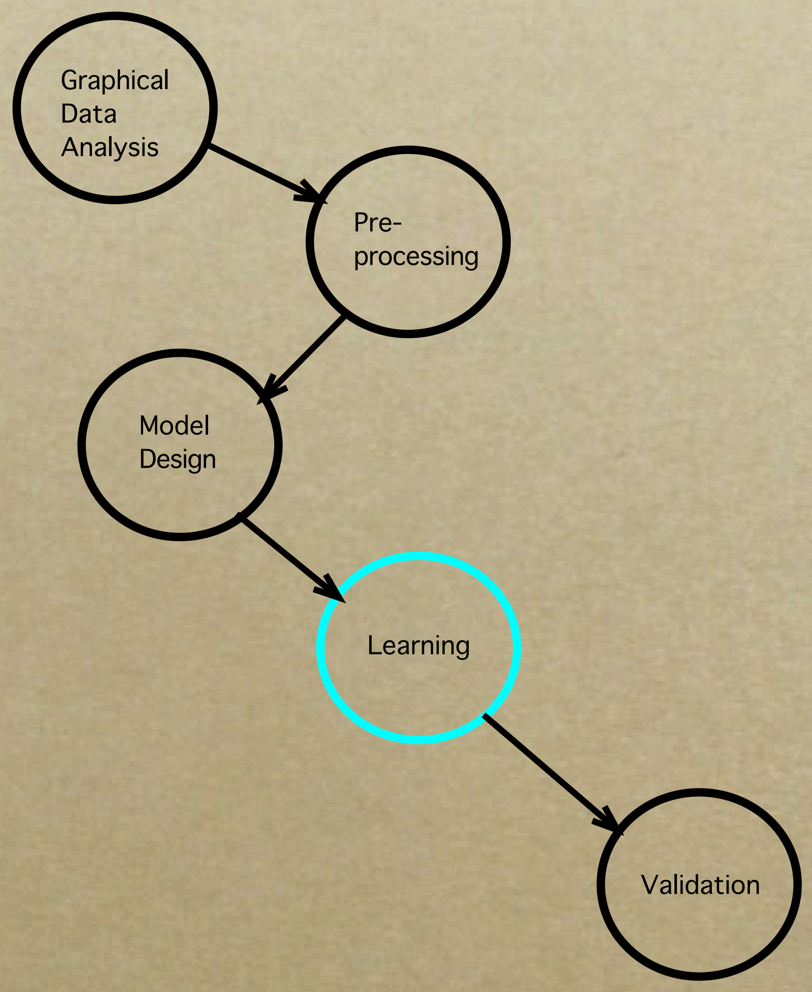
Preprocessing



- *Feature selection (mutual information between frv, and other wrapper / filter methods)*
- *Instance selection*
- *Transformations of sets variables*

[7] Sanchez, L., Suarez, M. R., Villar, J. R., Couso, I. **Some results about mutual information based feature selection and fuzzy discretization of vague data**. Intl. Conf. Fuzzy Systems FUZZ-IEEE 2007, pp. 1-6. 2007.

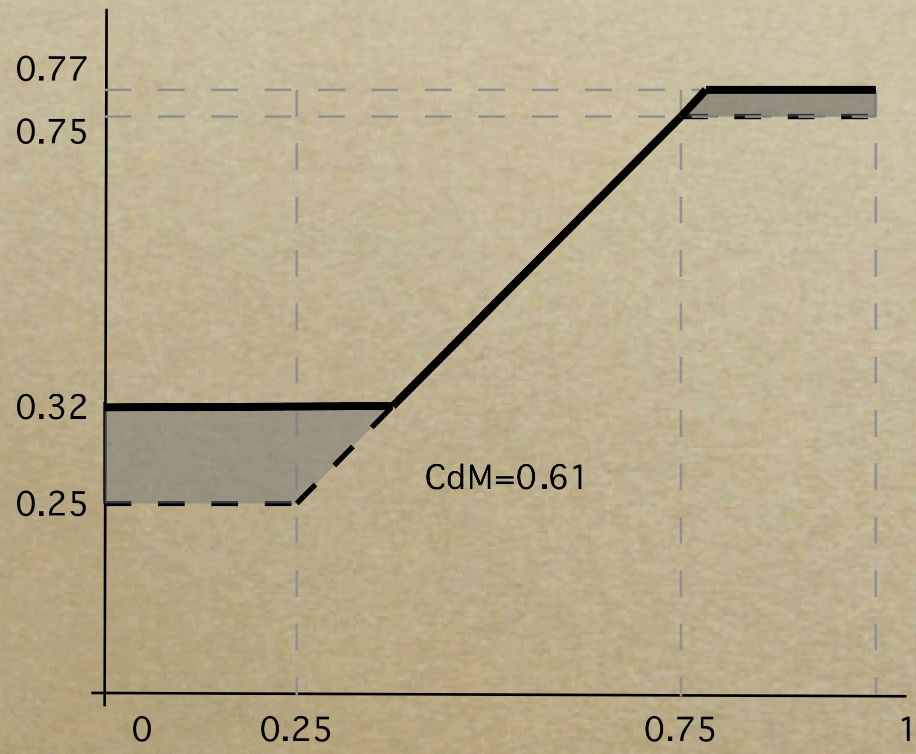
Fuzzy fitness-based GFS

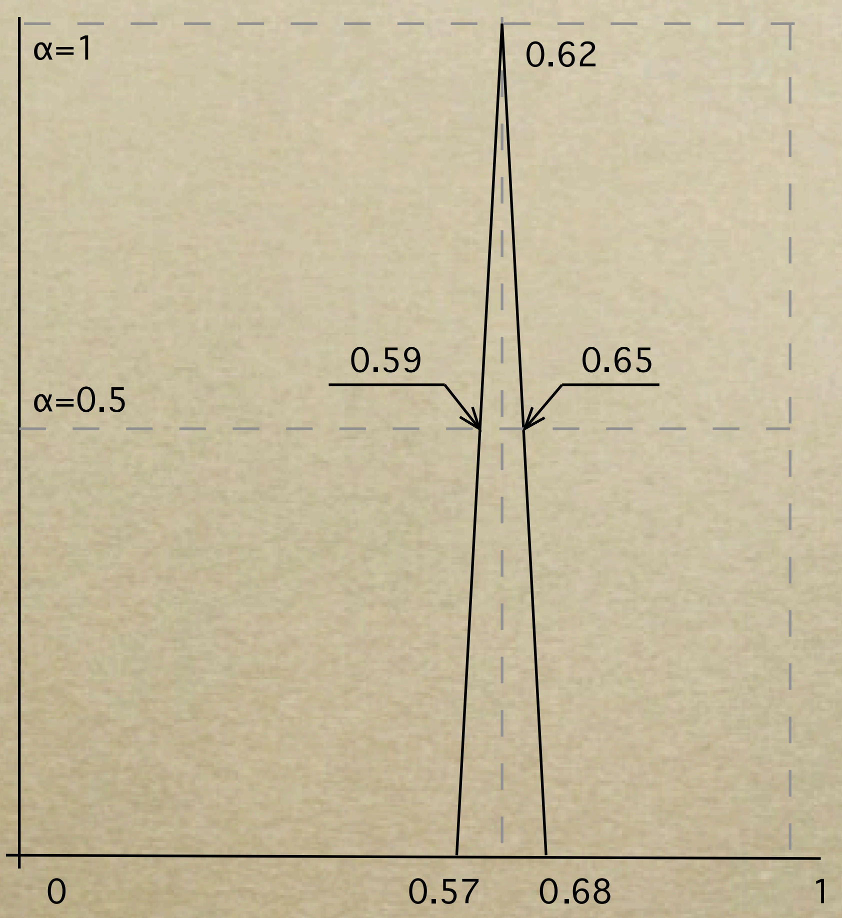


- *New inference mechanisms, that are coherent with the representation of interval or fuzzy data*
- *New measures of fitness (variance of an frv, error of a classifier on imprecise data, etc.)*

Sanchez, L., Couso, I., Casillas, J., **Genetic Learning of Fuzzy Rules based on Low Quality Data**. Submitted to Fuzzy Sets and Systems.

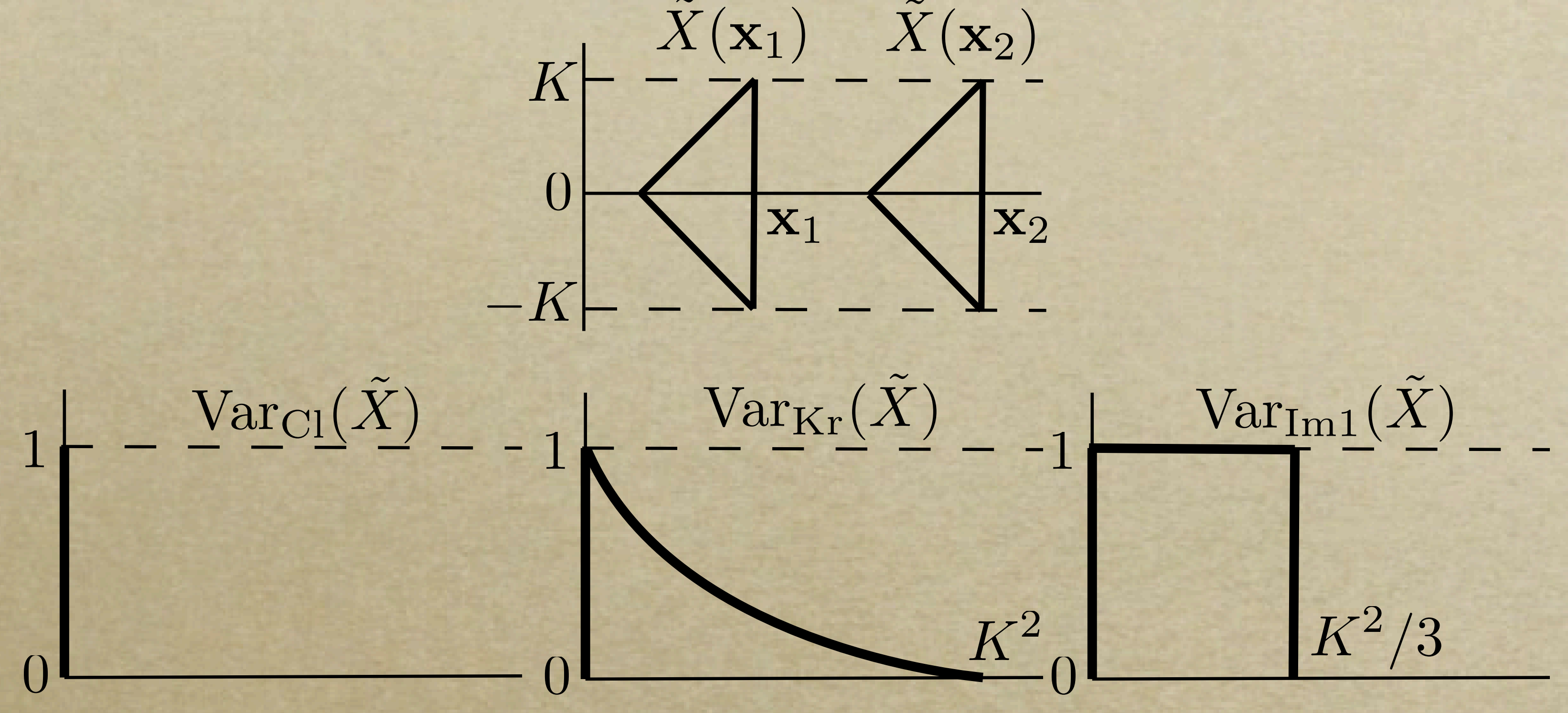
Inference





- *The reasoning method must be coherent with the possibilistic representation of the data.*

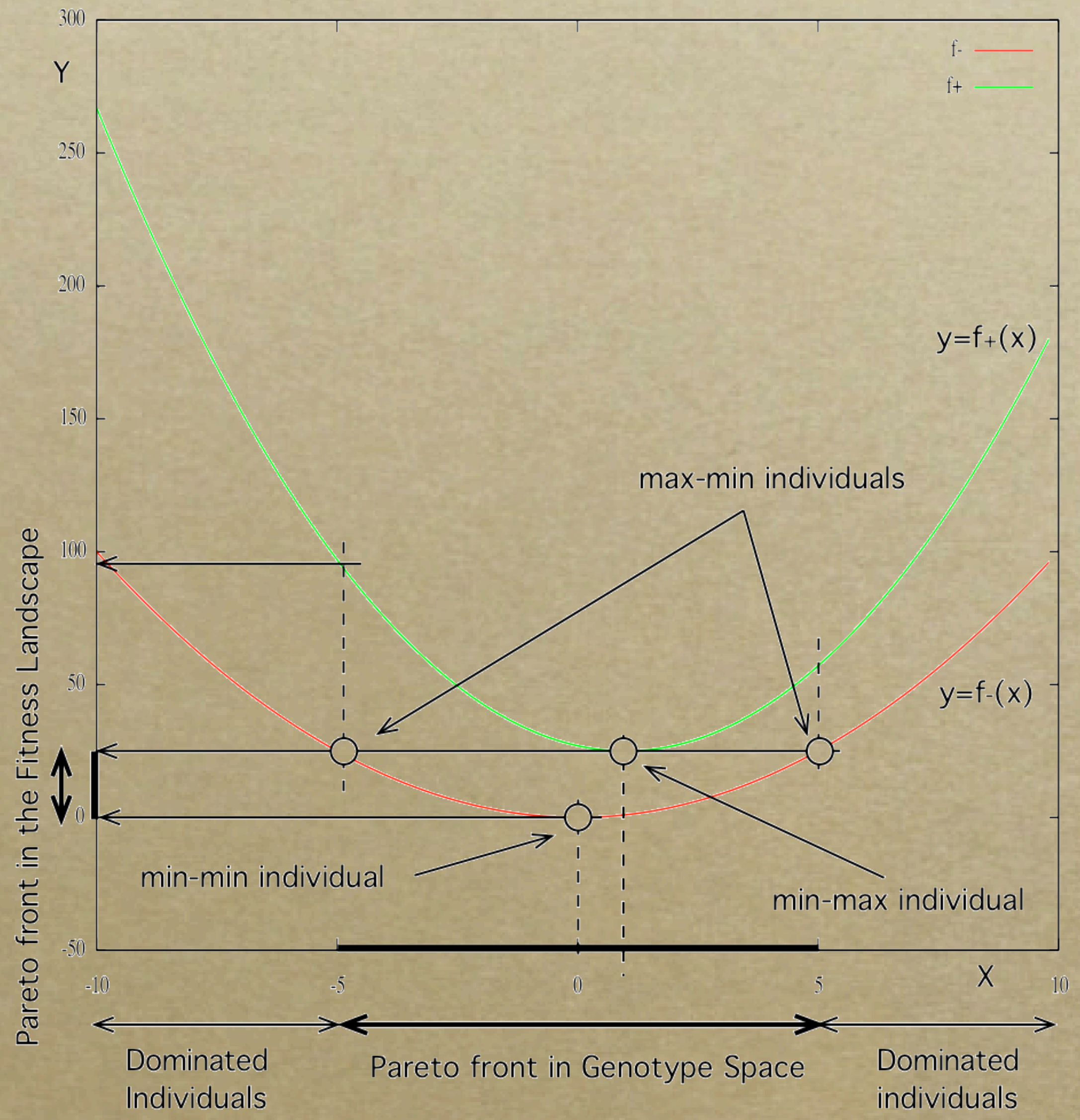
Fitness Function



- *The 1st order imprecise model has an interval-valued (not fuzzy) fitness*

[5] Couso, I., Dubois, D., Montes, S., Sanchez, L., **On various definitions of the variance of a fuzzy random variable**. 5th International Symposium on Imprecise Probability: Theories and Applications. ISIPTA'07. 16-19 July 2007

Optimization of interval-valued fitness functions



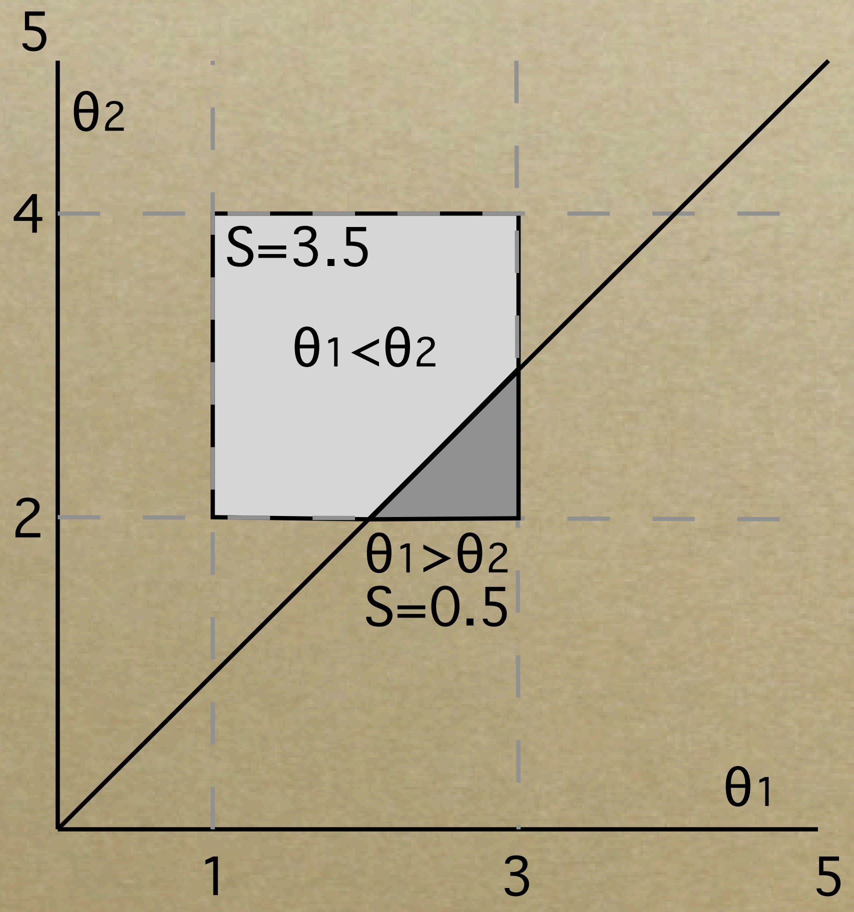
- *GAs and Metaheuristics should minimize interval valued functions*
- *Optimizing an interval valued funcion is analogous to solving a multicriteria problem: the same algorithms (i.e. NSGA-II, SPEA, etc) can be adapted.*

NSGA-II for interval-valued fitness

- *Dominance relation (non dominated sorting)*
- *Crowding distance*

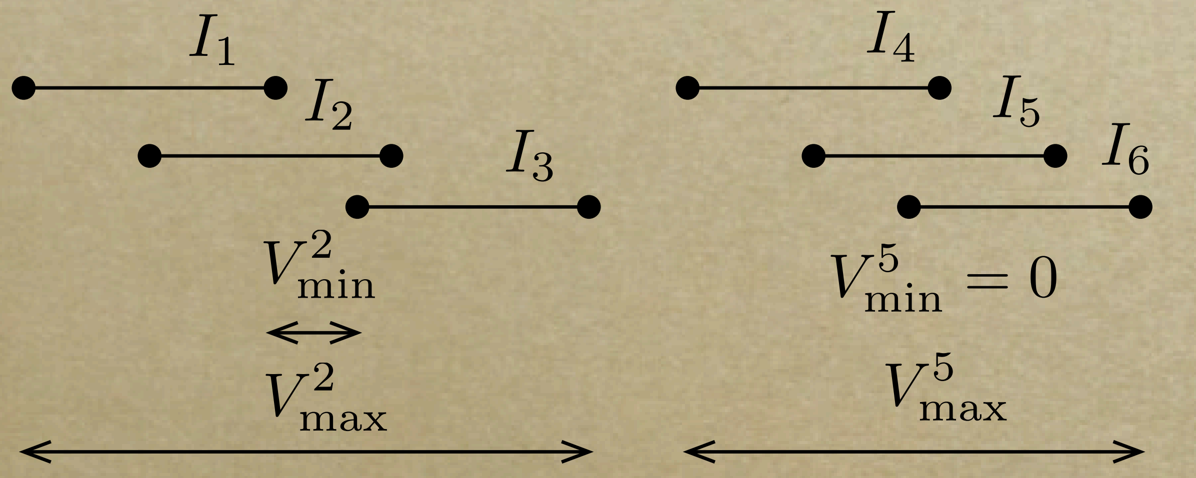
[2] Sánchez, L. Couso, I. Casillas, J. **Modelling Vague Data with Genetic Fuzzy Systems under a Combination of Crisp and Imprecise Criteria**, in *Proc. of the 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making (MCDM'2007)*, pp. 30--37, Honolulu, Hawaii, USA, April 2007

Dominance



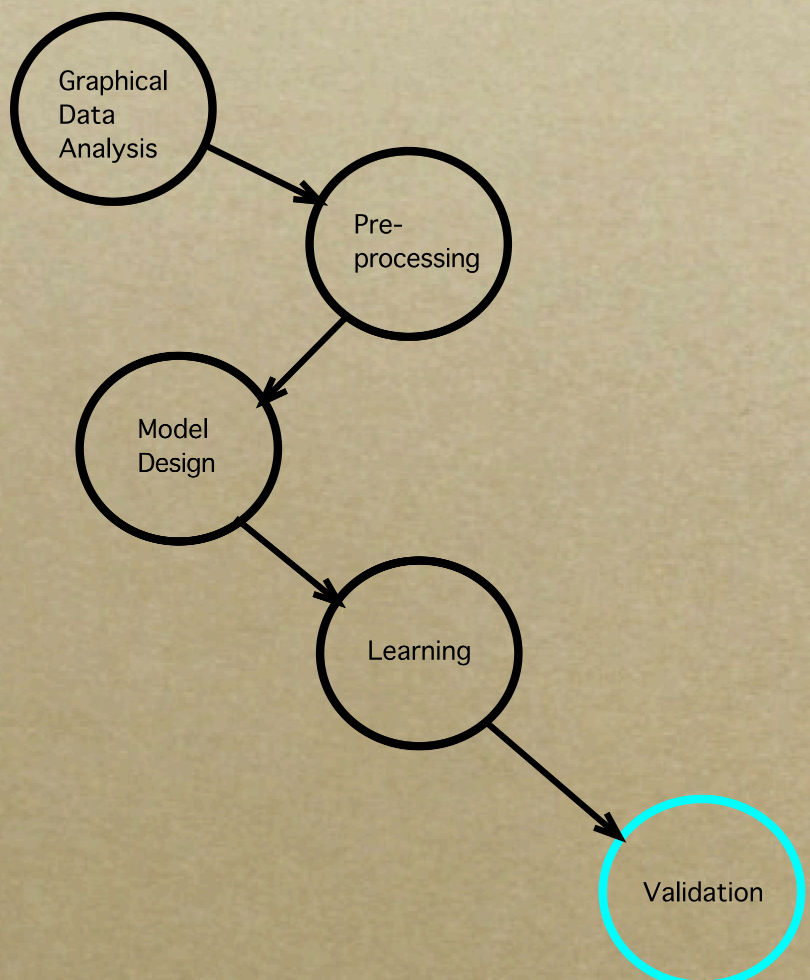
- *Strong dominance*
- *Uniform prior*
- *Imprecise probabilities based prior*

Crowding



- *The Haussdorff distance is between the minimum and the maximum distance between the individuals*

Validation



- *Benchmarks with interval and fuzzy data*
- *Statistical tests for comparing samples of fuzzy random variables*

Couso, I., Sanchez, L., **Mark-recapture techniques in statistical tests for imprecise data.** International Journal of Approximate Reasoning (submitted)

[10] Couso, I., Sanchez, L., **Defuzzification of fuzzy p-values**. Fourth International Workshop on Soft methods in probability and statistics SPMS'08 (admitted)

Summary of Part I

- *Lots of open problems. Some to mention:*
  - *Theoretical study of the relations with **SVM***
  - ***Graphical analysis tools** for coarse data (MDS, ICA, PCA, etc.)*
  - ***Feature selection**, instance selection, transformations*
  - ***New inference mechanisms** that match the representation of data*
  - ***New measures of fitness** for classifiers and models*
  - *Theoretical and practical studies relating MOEAs and optimization of interval and fuzzy-valued fitness functions, **new MOEAs and MO metaheuristics** for solving problems with mixtures of objectives*
  - ***Statistical tests** (bootstrap, t-test for imprecise data, etc.)*
  - ***Benchmarks** for comparing the new algorithms.*

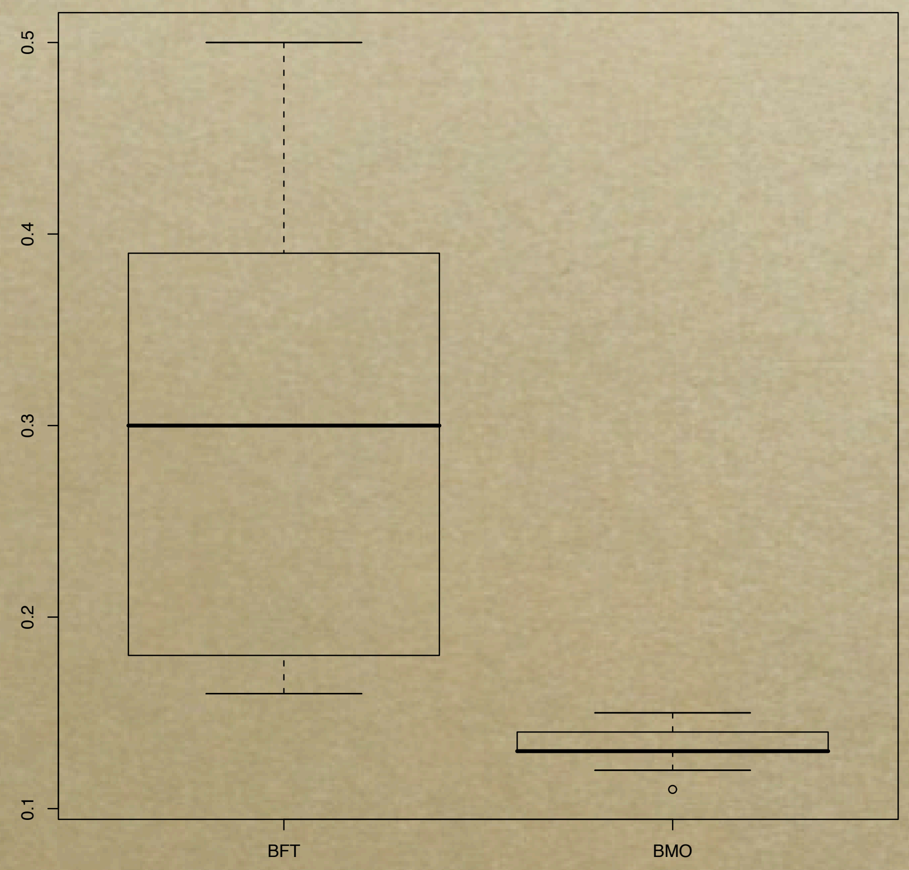
Part II

Some results on the use of Evolutionary Algorithms for extracting knowledge from low quality data

Examples of successful applications

1. *Artificial addition of imprecision to crisp data, learning with IRL and GCCL (synthetic datasets)*
2. *Aggregates of conflicting data, learning with Pittsburgh-style GFS (marketing models)*
3. *Crisp data + tolerance, evolutionary filtering (GPS trajectories)*
4. *Interval-valued data, Pittsburgh-style GFS (Diagnosis of dyslexia)*
5. *Preprocessing and evolutionary graphical analysis*

1. IRL (Boosting) - Extended Data



- *The addition of fuzzy imprecision to crisp data, followed by a fuzzy fitness-based GFS, improves the generalization*

[1] L. Sánchez, J. Otero, and J. R. Villar, "**Boosting of fuzzy models for high-dimensional imprecise datasets**," in Proc. IPMU 2006, Paris, France, 2006.

1. GCCL - Extended data

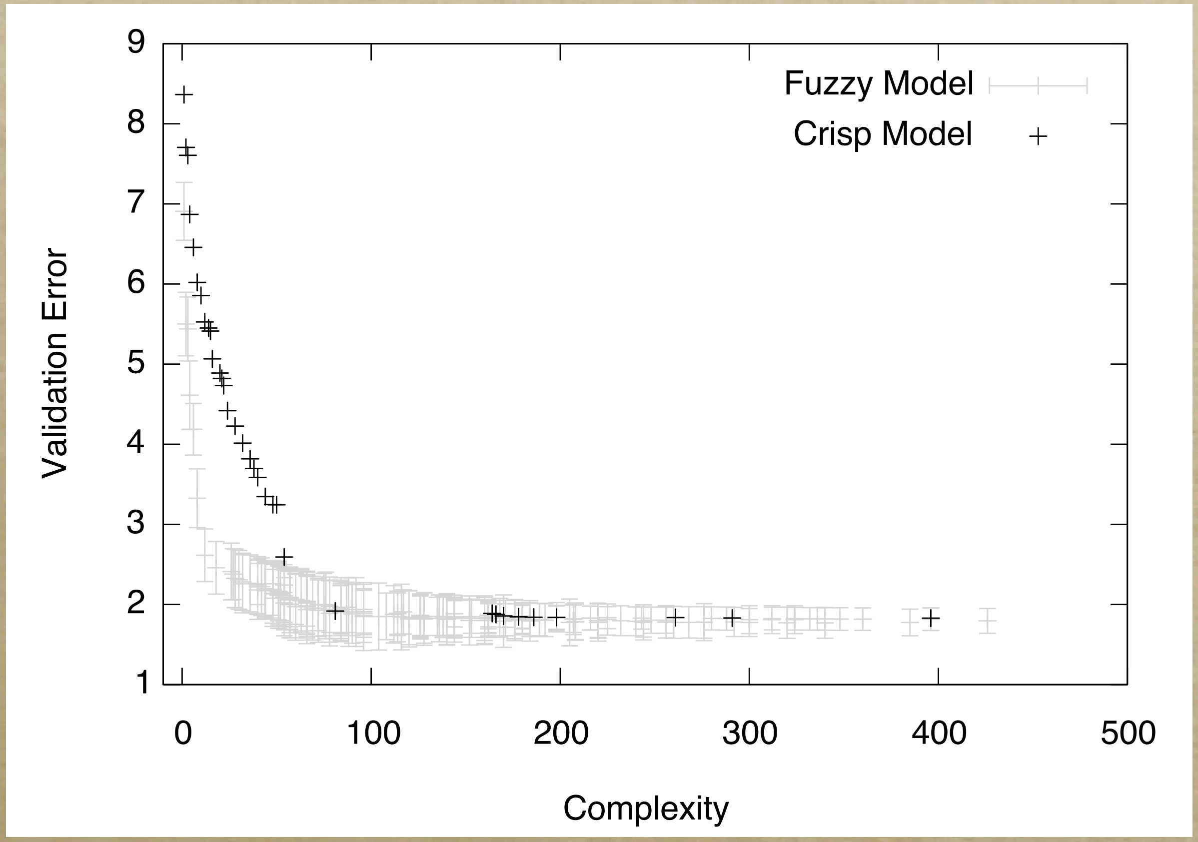
	1% BFT	1% FGCCL	5% BFT	5% FGCCL	10% BFT	10% FGCCL
f_1	0.89	**0.35**	6.64	**6.25**	24.82	**24.77**
f_1-10	2.66	**1.86**	9.39	**8.31**	29.03	**28.60**
f_2	0.52	**0.23**	0.60	**0.47**	1.41	**1.23**
f_2-10	0.56	**0.37**	0.97	**0.68**	**1.67**	1.70
elec	440	**421**	581	**558**	1003	**988**

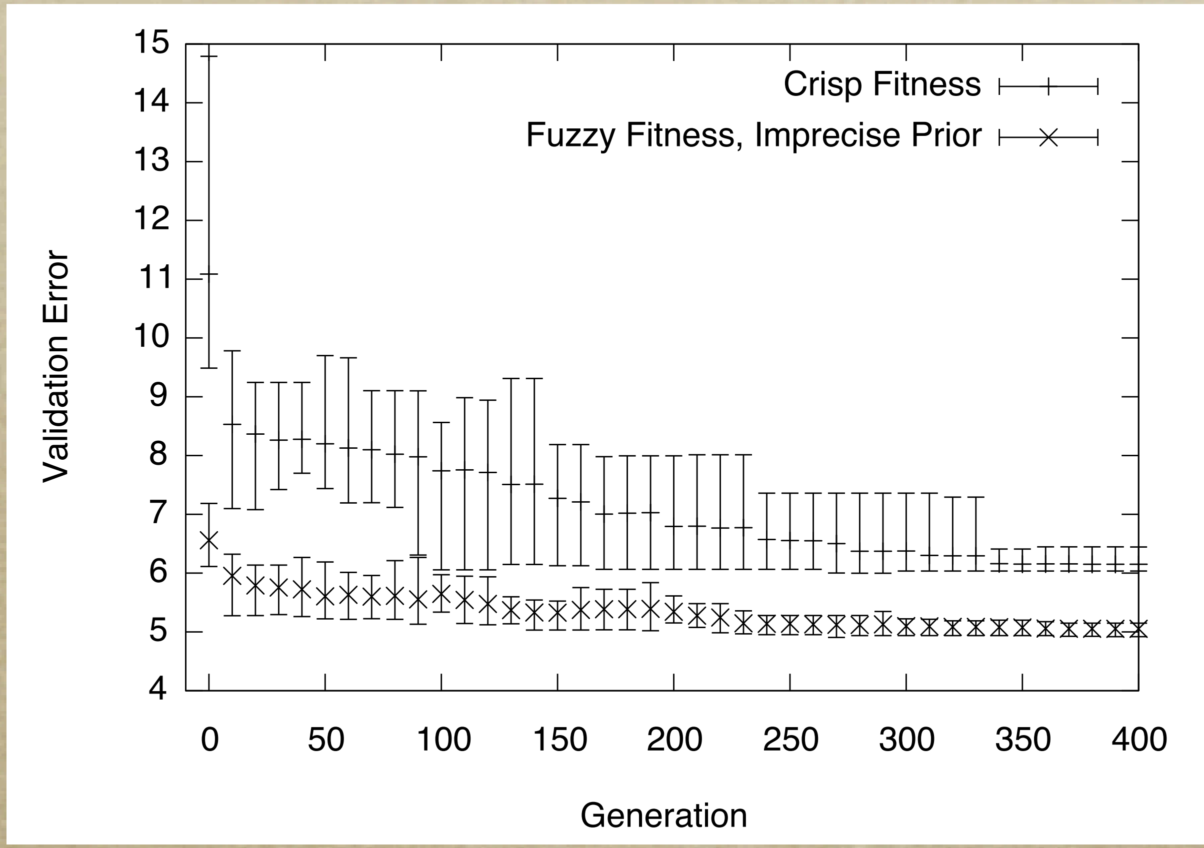
	Labels	WM	WLS-TSK	BFT	FGCCL
elec	3	7	8	10	**4**
machine-CPU	3	20	91	25	**4**
daily-elec	3	64	427	25	**5**
Friedman	3	192	242	25	**10**
building	3/2*	789	896*	30	**20**

- *GCCL is better than IRL for low quality data, in complexity and accuracy*

[6] Sanchez, L., Otero, J. **Learning fuzzy linguistic models from low quality data by genetic algorithms** FUZZ-IEEE 2007, London. pp 1-6, 2007

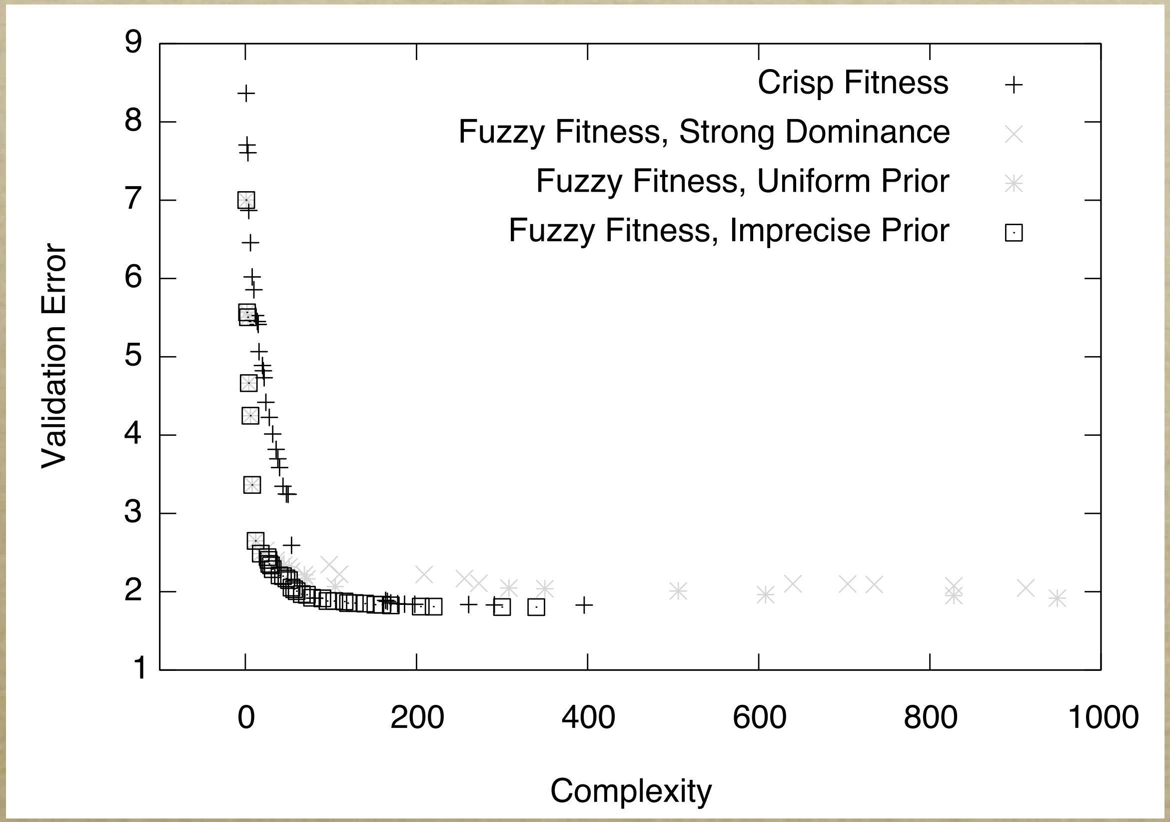
2. Pittsburgh - Aggregated data (I)

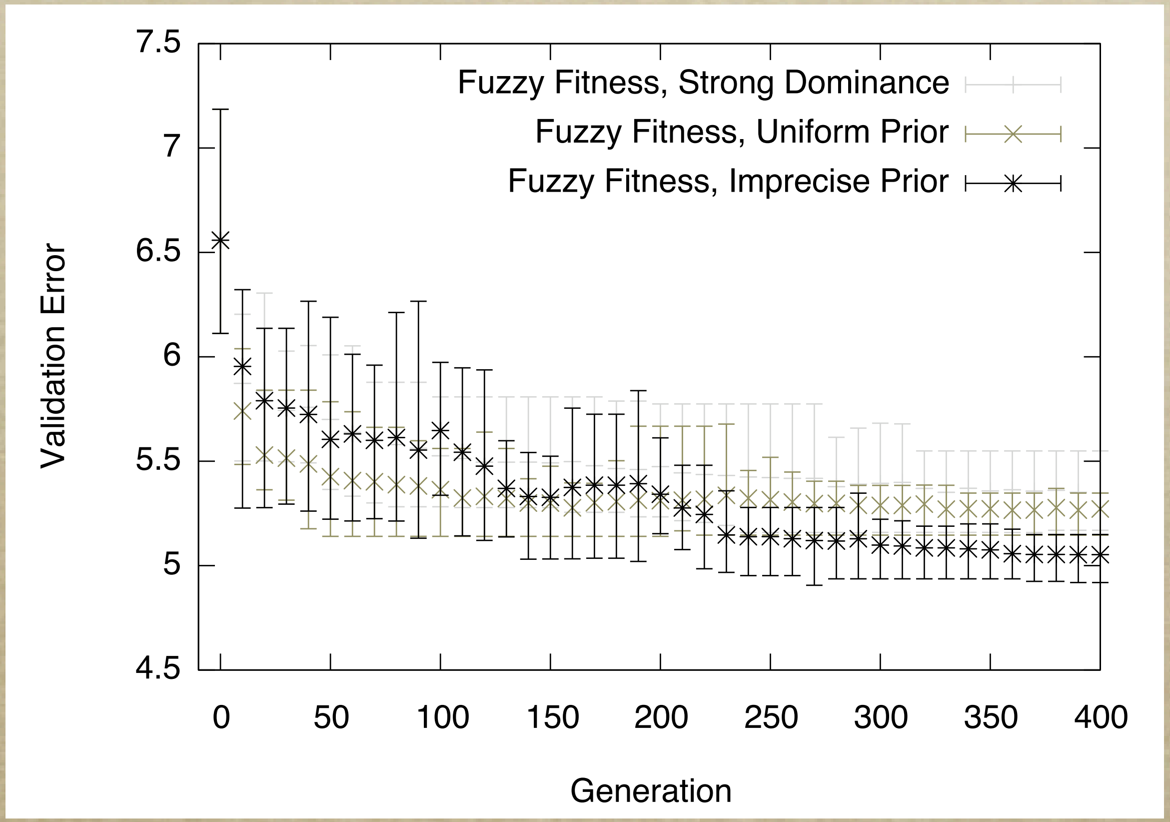




- *The fuzzy fitness-based GFS improved the accuracy of the crisp version.*

2. Pittsburgh - Aggregated data (II)

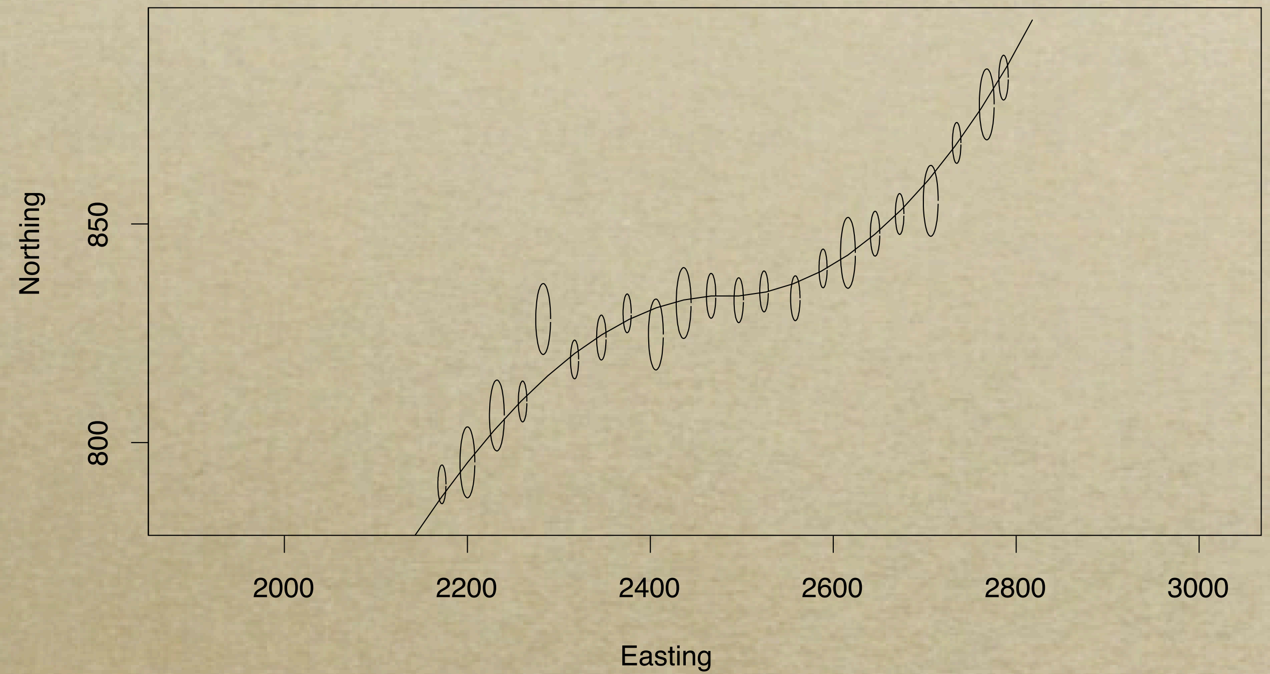




- *The precedence operator based on imprecise probabilities is more efficient in the latter generations of the GA*

[2] Sánchez, L. Couso, I. Casillas, J. **Modelling Vague Data with Genetic Fuzzy Systems under a Combination of Crisp and Imprecise Criteria**, in *Proc. of the 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making (MCDM'2007)*, pp. 30--37, Honolulu, Hawaii, USA, April 2007

3. Composite data (GPS)

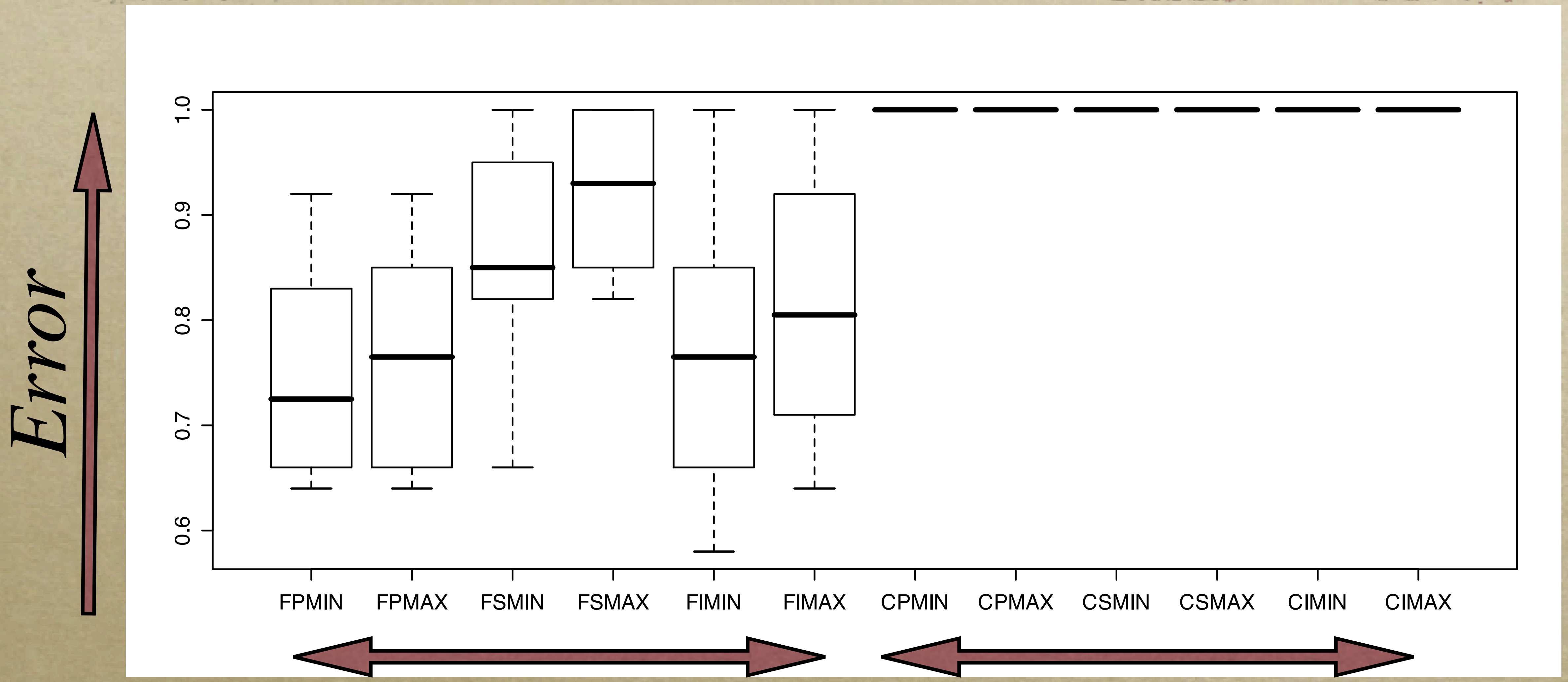


Dataset	True Length	MOSA Best	MOSA Mean	NSGA-II Best	NSGA-II Mean
1	3228.574	3234.86	3235.83	3242.64	3243,45
2	2741.306	2767.06	2782.77	2765.63	2798.01

- *Population-based SA and NSGA-II were used to find the lowest upper bound of the length of a trajectory*

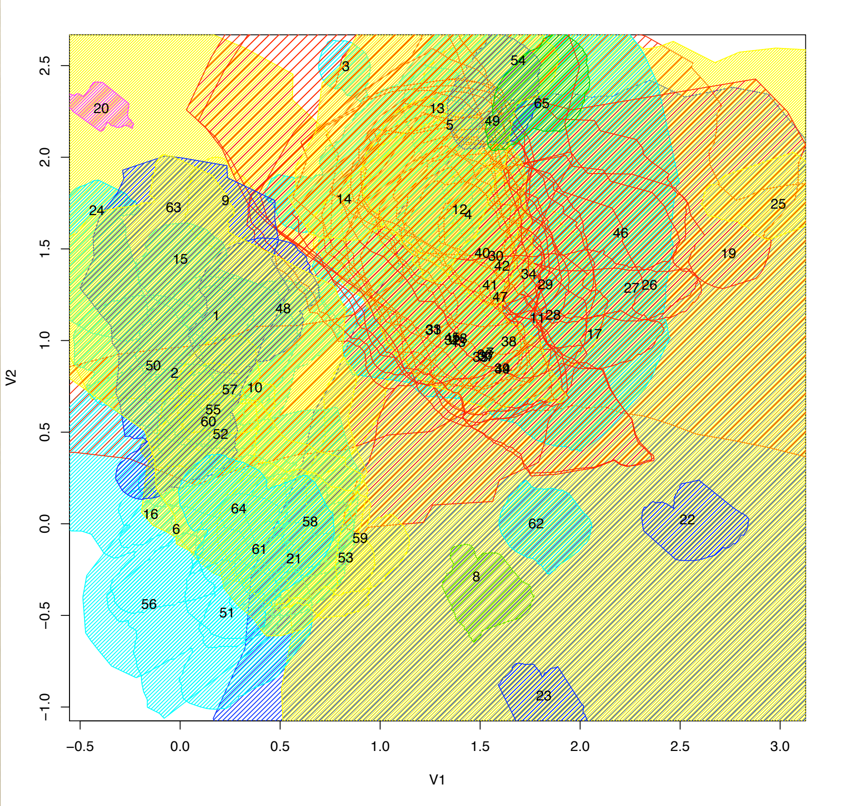
[3] Sanchez, L., Couso, I., Casillas, J. **A Multiobjective Genetic Fuzzy System with Imprecise Probability Fitness for Vague Data**. Int. Symp. on Evolving Fuzzy Systems (EFS 2006), pp. 131-136, 2006.

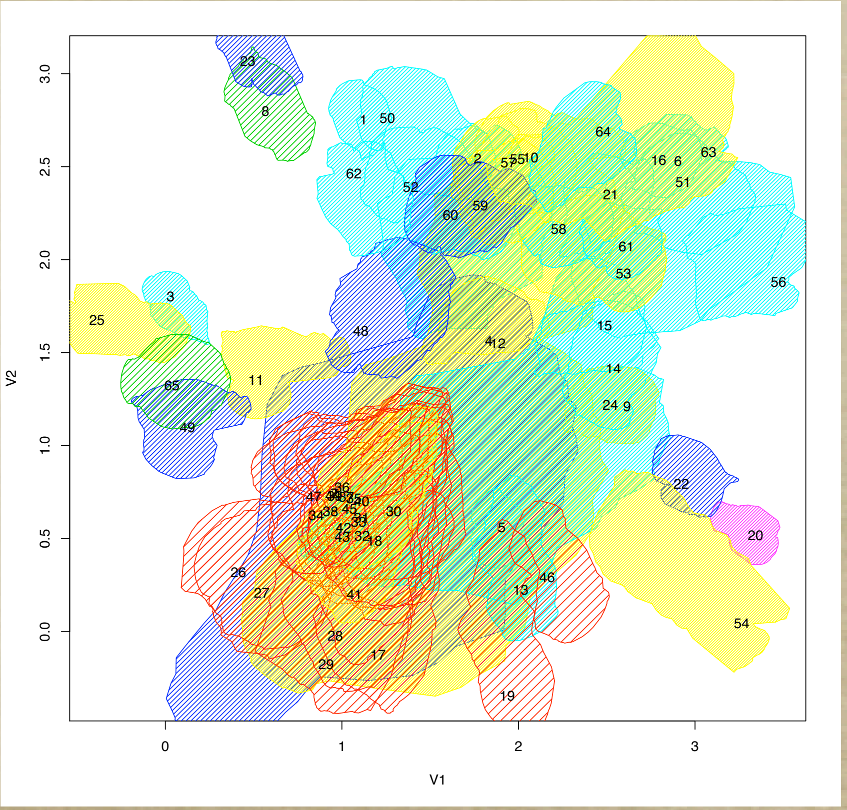
4. Feature Selection + Pittsburgh

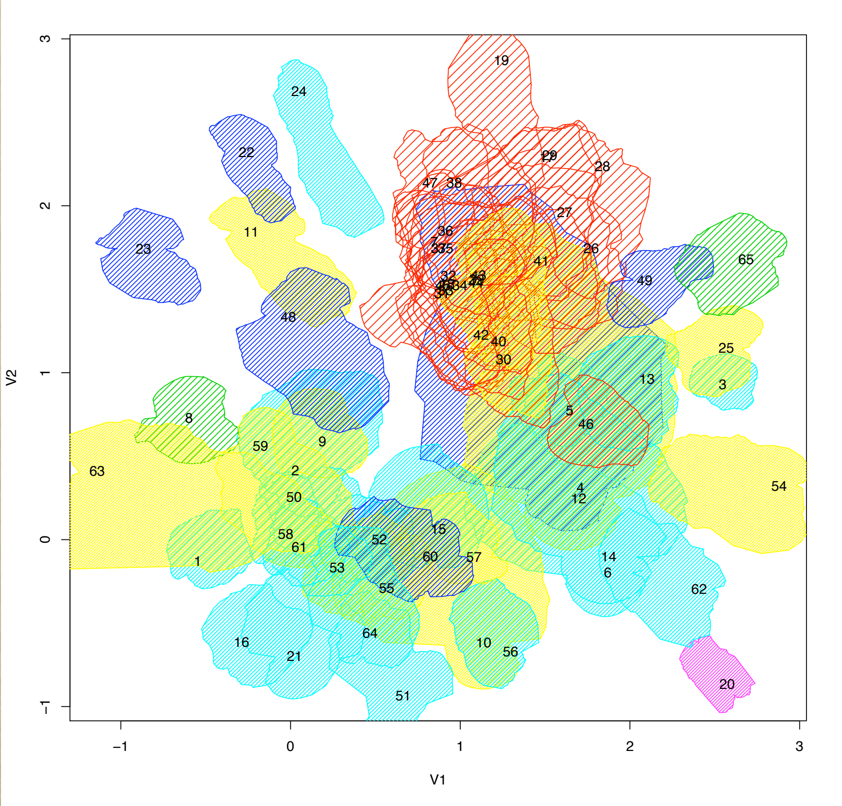


- *Dyslexia diagnosis with interval data. The set of variables selected based on fuzzy techniques are uniformly better than those found by crisp techniques*

5. MDS, Interval & Missing data



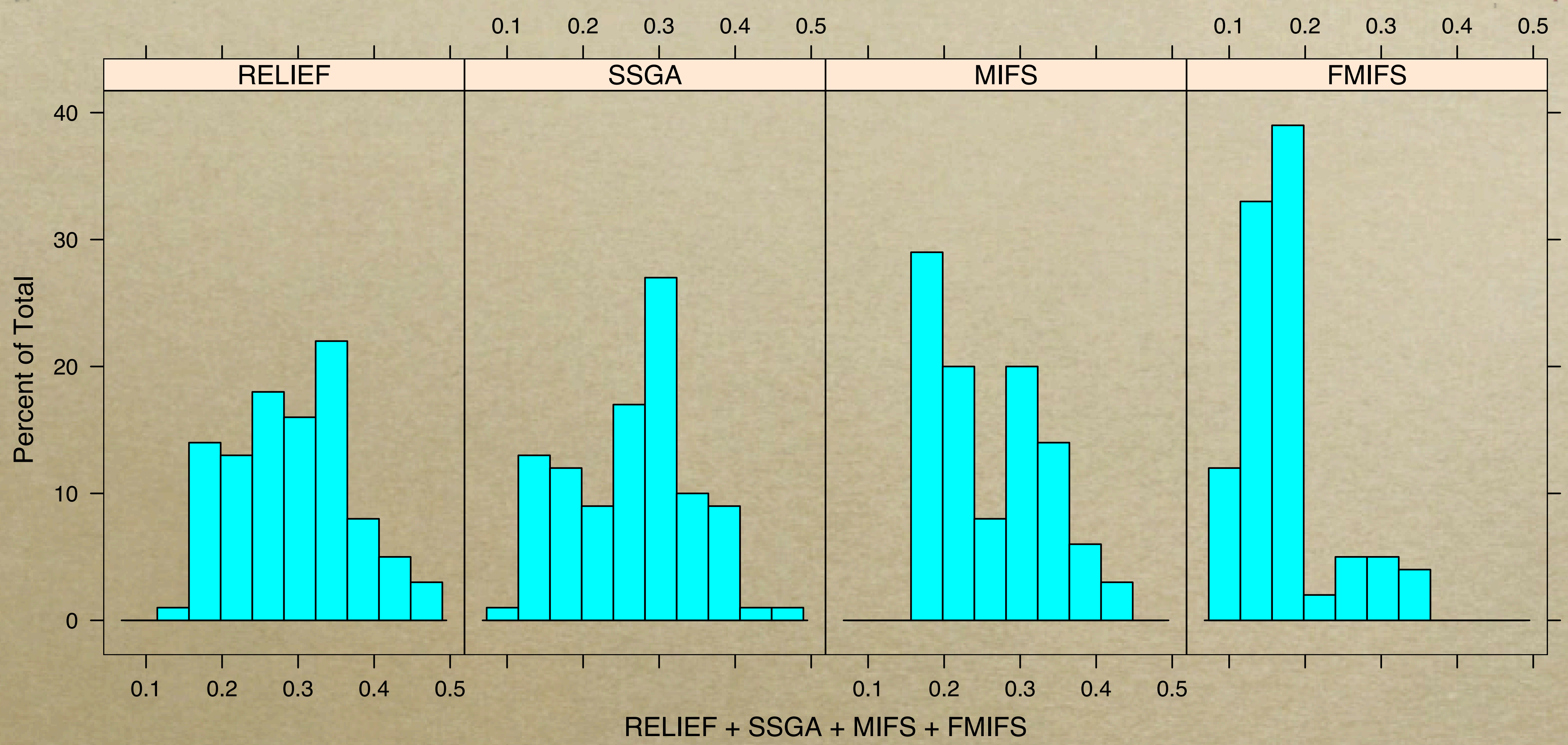




- *MDS analysis for different granularities of the linguistic partition*

[9] L. Sánchez, J. Otero, and J. R. Villar, "**Graphical exploratory analysis of vague data in the early diagnosis of syslexi**
in Proc. IPMU 2008, Málaga, Spain, 2008.

5. Feature selection, fuzzified data



- *The ranking of the features depends on the linguistic partition of the input variables*

[7] Sanchez, L., Suarez, M. R., Villar, J. R.; Couso, I. **Some Results about Mutual Information-based Feature Selection and Fuzzy Discretization of Vague Data**. FUZZ-IEEE 2007, London. pp 1-6, 2007.