TPNC 2016

5th International Conference on Practice of Natural Computing

Is the information extracted enough to apply machine learning algorithms?

Antonio J. Tallón-Ballesteros[1], Cristina Urbano-Sánchez[2], María Rodríguez-Romero[2] and Luís Correia[3]

[1] Department of Languages and Computer Systems, University of Seville, Spain. atallon@us.es
[2] Higher Technical School of Computer Science Engineering, University of Seville, Spain
[3]Department of Computer Science, University of Lisbon, Portugal

Index

1. Introduction

- Data versus information.
- Machine learning refers to algorithms that allow a computer to learn from experience [4]
- This contribution focuses on supervised machine learning.
- Currently, the number of features is extremely large and a data pre-processing [2] is a requirement.
- Our goal is to improve the accuracy and the Cohen's kappa.

2. Classifiers

- Classifiers also called supervised machine learning algorithms could be divided into:
  - Rule-based classifiers
  - Neural networks
  - Decision trees
  - Classifiers based on nearest neighbours (*k*NN)
  - Support vector machines

3. Experimental results

- Our proposal is based on the application of two kind of data pre-processing procedures:
  1. The former carries out feature subset selection based on a consistency-based measure [1] with a scatter search to guide the exploration [3].
  2. The latter takes as input the reduced training subset according to the previous step and create new attributes with the arithmetical combination of the feature values.

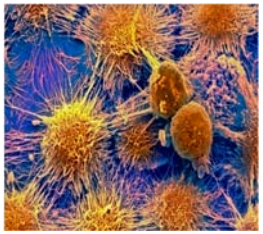
3. Experimental results (cont.)

- Two real-world and complex data sets from Bioinformatics with at least 7 classes

  ... and more than 12 000 of features [5].

Table 1. Data Sets

Data Set	No. Features	Classes	Size
GCM	16063	14	83
SALL	12558	7	327



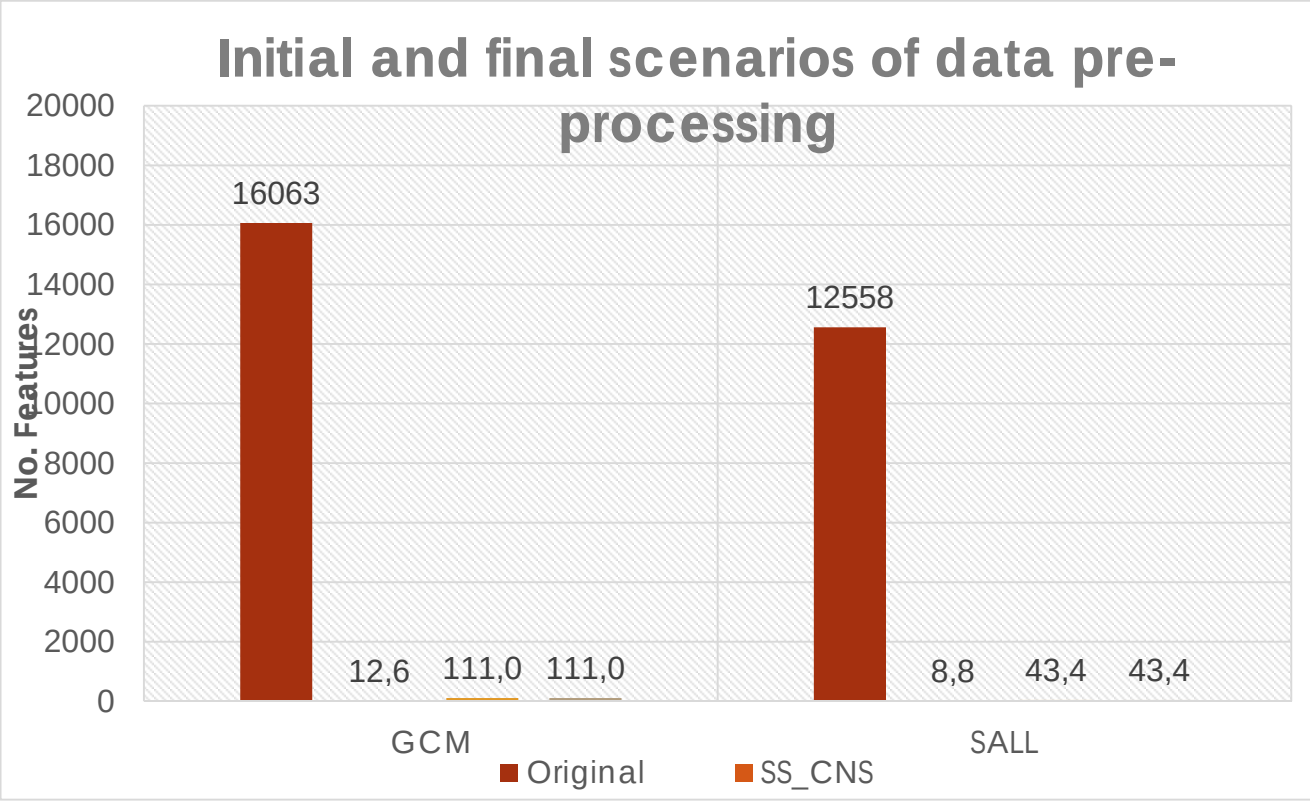
- Validation procedure: holdout with training and test sets [4].

- Two classifiers: *k*NN (with *k*=1) and SVM.

3.1. Feature selection

- Step 1: Feature subset selection via consistency-based measure with scatter search (SS_CNS) [3]
- Step 2: Generation of new attributes taking as starting point all the attributes selected previously

Figure 1. Resulting number of features for each data set



3.2. Performance results

- Global classification results
- Two performance measures: accuracy and Cohen's kappa.
- Cohen's kappa is essential in the context of multi-class problems.

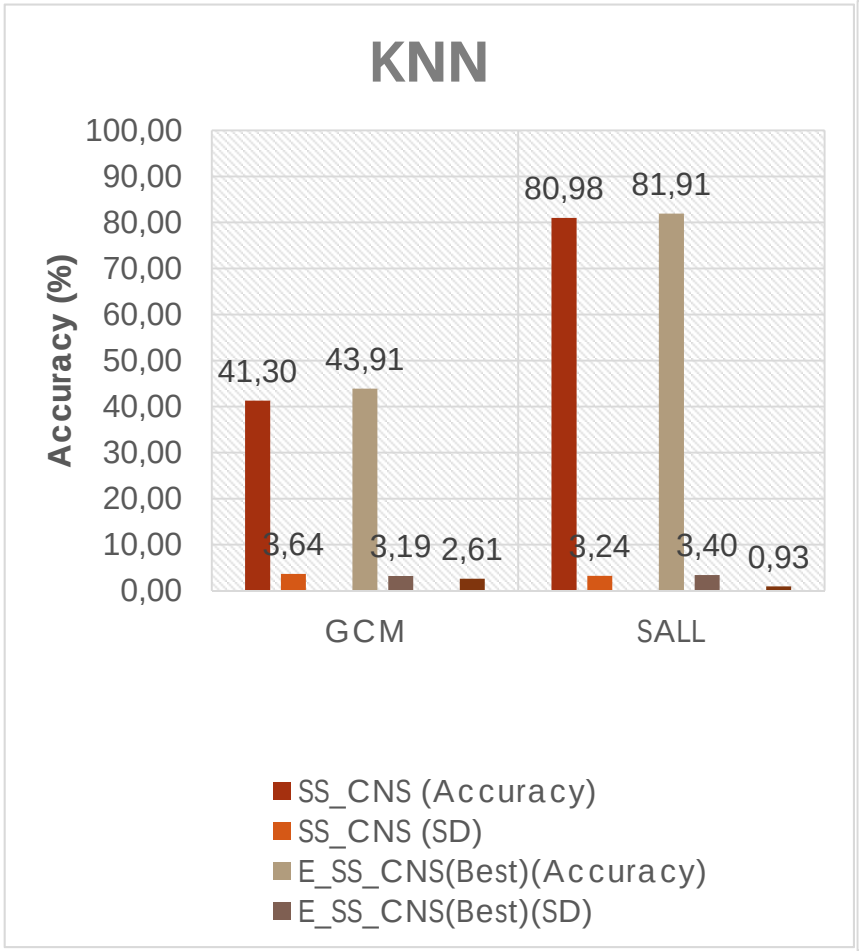
Table 3. KNN and SVM results

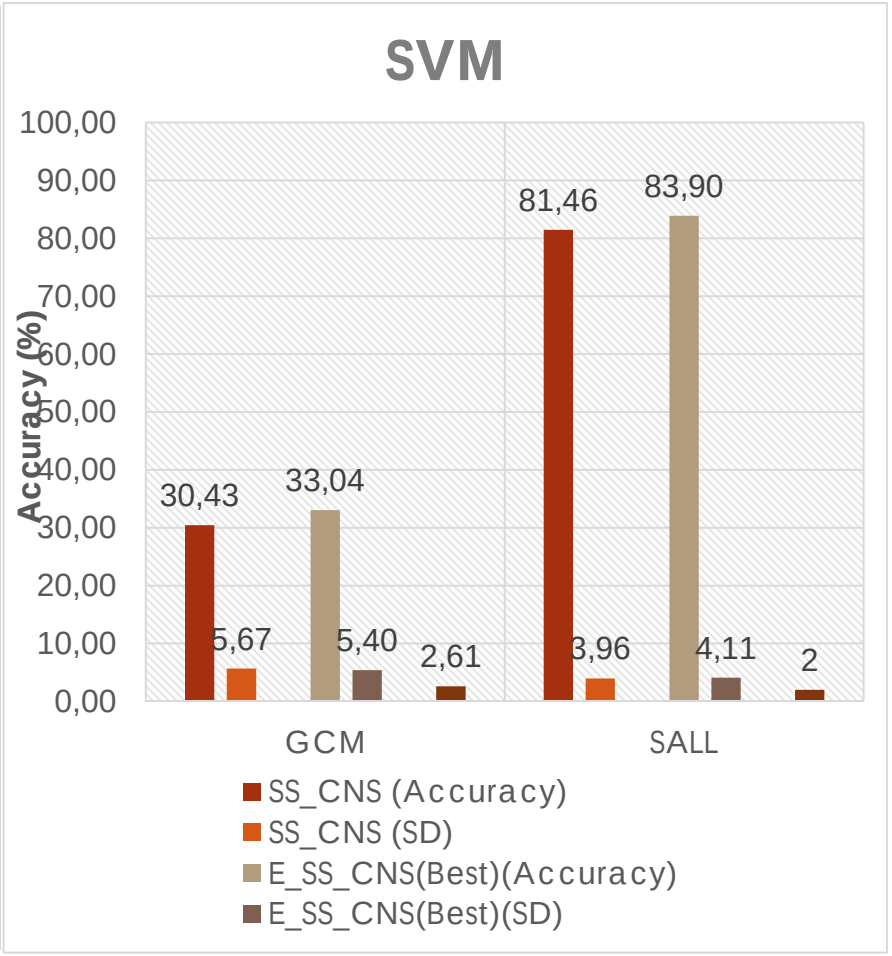
Data Set	Feature selection method	Performance measure			
		Accuracy±SD	Cohen's kappa±SD	Accuracy±SD	Cohen's kappa±SD
		Classifier			
		KNN		SVM	
GCM	SS_CNS	41,30±3,64	0,3630±0,0390	30,43±5,67	0,2200±0,0750
	E_SS_CNS (+)	43,91±3,19	0,3910±0,0340	33,04±5,40	0,2690±0,0570
	E_SS_CNS (*)	41,30±3,37	0,3640±0,0350	31,30±6,24	0,2350±0,0770
SALL	SS_CNS	80,98±3,24	0,7666±0,0038	81,46±3,96	0,7670±0,0490
	E_SS_CNS (+)	80,73±3,03	0,7630±0,0360	83,90±4,11	0,7999±0,0520
	E_SS_CNS (*)	81,91±3,40	0,7770±0,0410	82,93±4,29	0,7870±0,0550
Average	SS_CNS	61,14	0,5648	55,95	0,4935
	E_SS_CNS (+)	62,32	0,5770	58,47	0,5345
	E_SS_CNS (*)	61,61	0,5705	57,12	0,5110

3.2. Performance results (cont.)

- Test accuracy

Figure 2. Test accuracy for each classifier

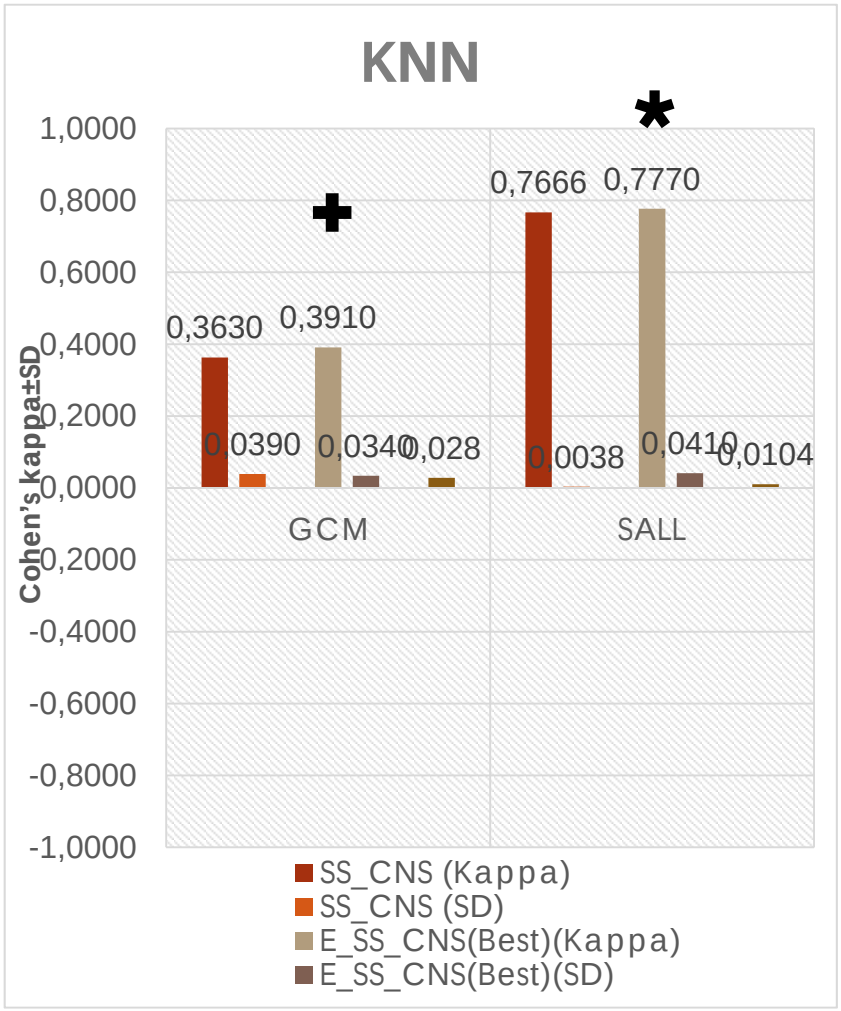


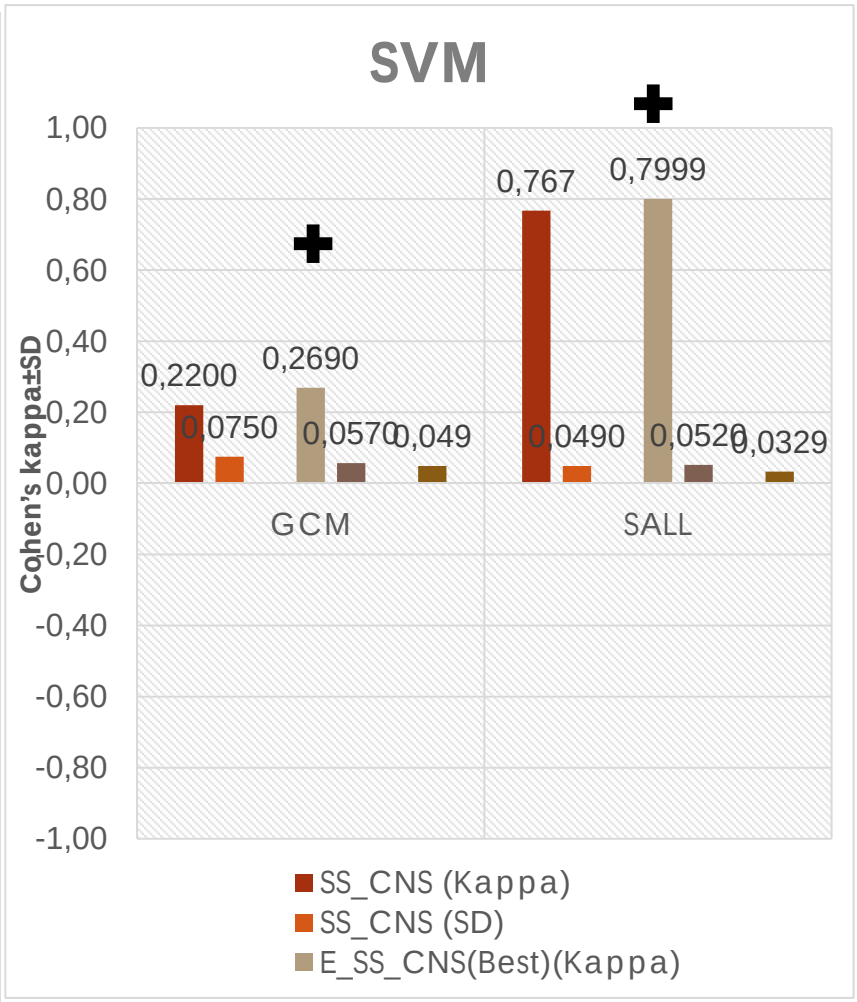


3.2. Performance results (cont.)

- Test Cohen's kappa

Figure 3. Test Cohen's kappa for each classifier

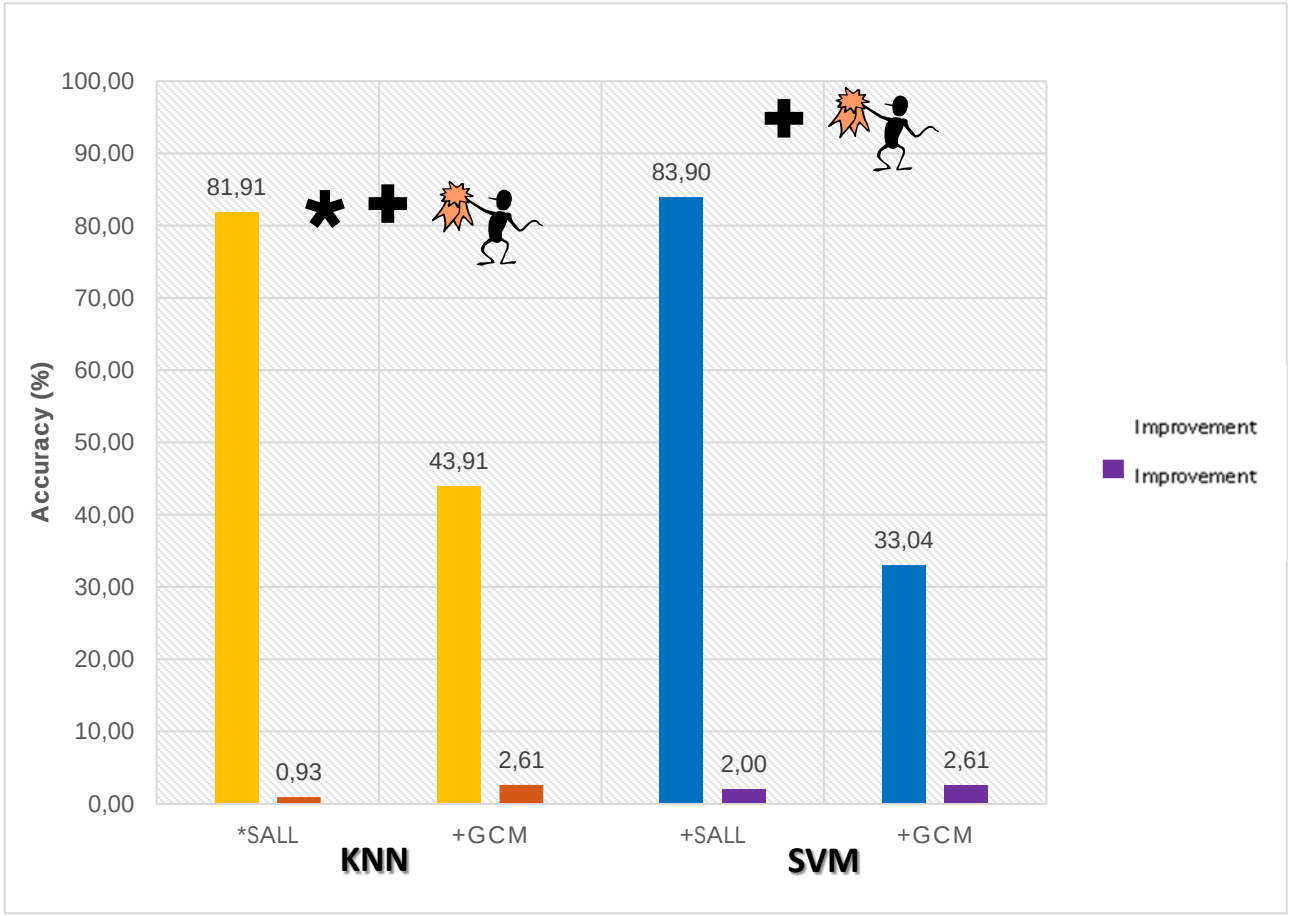




3.2. Performance results (cont.)

- Best results for test accuracy

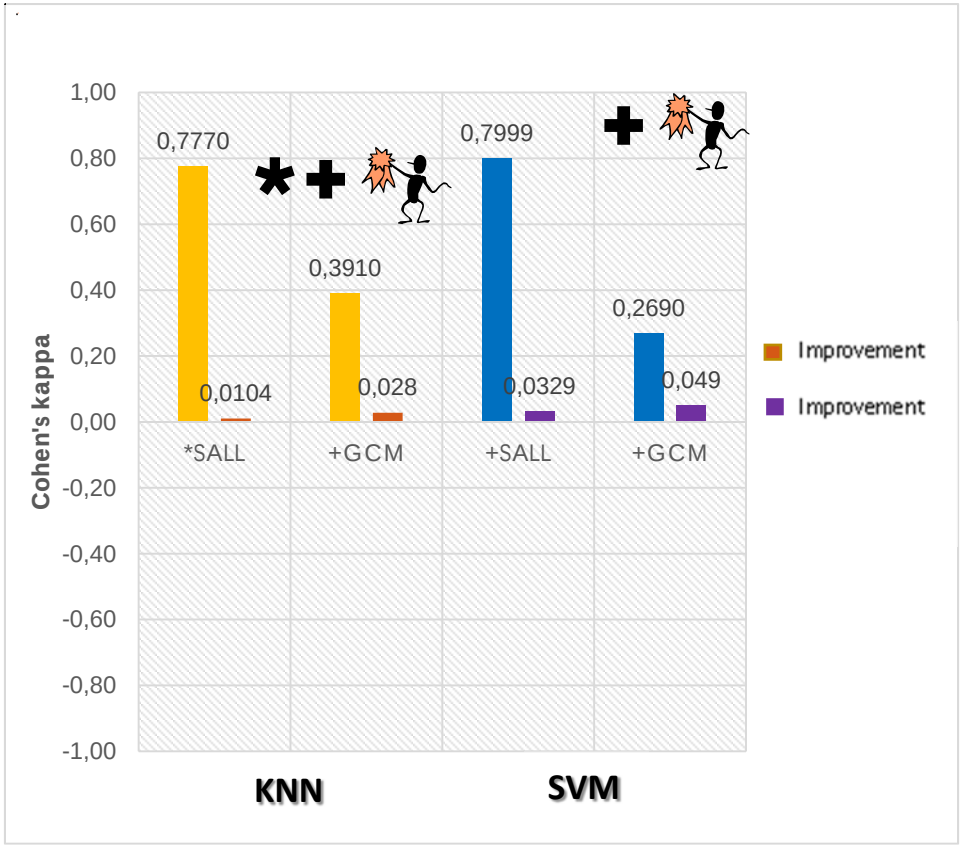
Figure 4. Best approaches for each classifier



3.2. Performance results (cont.)

- Best results for test Cohen's kappa

Figure 5. Best approaches for each classifier



4. Conclusions

- Additive and multiplicative relations are very noticeable. In KNN there are some cases that is better the sum, whereas in others the multiplication is better, while in SVM the sum is substantially better in all cases.

- Concretely, in GCM the classifier KNN obtained an improvement of 2,61 and in SALL it is obtained an improvement of 0,93. With the classifier SVM, in GCM it is obtained an improvement of 2,61 as well, and in SALL an improvement of 2.

5. References

[1] M. Dash, H. Liu (2003). Consistency-based search in feature selection.: **Artif Intell** 151(1-2), 155-176.

[2] Aggarwal, C. C. (2015). **Data preparation**. In C. C. Aggarwal (Ed.), **Data mining** (pp. 27-62). Cham: Springer.

[3] Tallón-Ballesteros, A. J., & Ibiza-Granados, A. (2016). Simplifying pattern recognition problems via a scatter search algorithm. **International Journal for Computational Methods in Engineering Science and Mechanics**, 17 (5-6), 315-321.

[4] Dougherty, G (2012). **Pattern recognition and classification: An introduction**. New York, NY: Springer

[5] McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. (2006)**.** Machine learning for detecting gene-gene interactions**. Applied bioinformatics**, 5(2), 77-88.

TPNC 2016

5th International Conference on Practice of Natural Computing

Is the information extracted enough to apply machine learning algorithms?

Thanks for your patience!

Antonio J. Tallón-Ballesteros[1], Cristina Urbano-Sánchez[2], María Rodríguez-Romero[2] and Luís Correia[3]

[1] Department of Languages and Computer Systems, University of Seville, Spain. atallon@us.es
[2] Higher Technical School of Computer Science Engineering, University of Seville, Spain
[3]Department of Computer Science, University of Lisbon, Portugal