Scalability analysis of filter-based methods for feature selection

Diego Peteiro-Barral, Veronica Bolon-Canedo, Amparo Alonso-Betanzos,
Bertha Guijarro-Berdiñas, and Noelia Sanchez-Maroño

Faculty of Informatics, University of A Coruña, A Coruña 15071, Spain,
{dpeteiro, vbolon, ciamparo, cibertha, nsanchez}@udc.es,
WWW home page: http://www.lidiagroup.org/

Abstract. Researchers in machine learning are now interested not only in accuracy but also in scalability of methods. Although scalability of learning algorithms is a trending issue, scalability of feature selection methods has not received the same amount of attention. In this research, a preliminary attempt to study the scalability of three well-known filter-based feature selection methods will be done. For this sake, several new measures are introduced, based not only in accuracy but also in execution time and stability and the results will be presented according to them.

Keywords: feature selection, machine learning, scalability

1 Introduction

The proliferation of high-dimensional data within many diverse domains has posed an unprecedented challenge to researchers [1]. This challenge can be twofold: (a) an enormous number of samples or (b) an enormous number of features. In the first case, the problem is that the performance of learning algorithms likely degenerates, whilst in the second case, the problem lies in the fact that with such a large number of features, the interpretability of a learning model decreases, as well as their computational efficiency declines.

For all these reasons, scaling up learning algorithms is a trending issue, as pointed out the workshop "Big Learning" at the conference of the Neural Information Processing Systems Foundation (NIPS'11). However, although scalability of learning algorithms has been recently the focus of much attention, scalability of feature selection algorithms has not received the same consideration by the scientific community.

Feature selection is a case of data partitioning which consists of selecting a subset of features [2] for reducing the size problem, that is often forgotten in discussions of scaling. Feature selection helps to avoid over-fitting (especially with small-size datasets), train more reliable learners or provide more insights into the underlying causal relationships. Moreover, as the number of samples increases, feature selection becomes more necessary from both run-time and spatial-complexity perspectives. Therefore, this preliminary research will be focused on the scalability of feature selection methods, paving the way to their application on extremely large datasets.

2 Feature selection

Feature selection is a technique which consists of selecting the relevant features and discarding the irrelevant ones in order to obtain a subset of features that describes properly the given problem. Among the different feature selection techniques available, this work will be focused on *filters* [2], since they rely on the general characteristics of training data and carry out the feature selection process as a pre-processing step with independence of the induction algorithm, making them computationally inexpensive.

Three filters will be considered in this work to study their scalability. All of them follow the subset evaluation approach, which consists of producing candidate feature subsets based on a certain search strategy. Each candidate subset is evaluated by a specific evaluation measure [3].

- **Correlation-based Feature Selection** (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function [4]. Theoretically, irrelevant features should be ignored and redundant features should be screened out.
- The **Consistency-based Filter** [5] evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of attributes.
- The **INTERACT** algorithm [6] is a subset filter based on symmetrical uncertainty (SU). Their authors stated that this method can handle feature interaction, and efficiently selects relevant features.

3 Experimental section

3.1 Materials

Two synthetic datasets were chosen to evaluate the scalability of feature selection methods. The main advantage of artificial scenarios the set of optimal features that must be selected is known and, therefore, the results of the filters can be easily evaluated [7]. Two different datasets were chosen for this research, following described (f_i stands for feature number i).

- A modified version of the CorrAL dataset [8] will be used. Its class value is $(f_1 \wedge f_2) \vee (f_3 \wedge f_4)$. The correct behavior for a given feature selection method is to select the four relevant features and to discard the irrelevant ones.
- The LED problem [9] is a simple classification task that consists of identifying the digit that the display is representing. Given the active leds described by seven binary attributes $f_1, \ldots, f_7$ (seven segments display), the task to be solved is its classification in one of the ten classes.

For assessing the scalability of the methods, different configurations of these datasets were used. In particular, the number of features ranges from 8 to 128 whilst the number of samples ranges from 8 to 1024 (all pairwise combinations). Notice that the number of relevant features is fixed (4 for CorrAL and 7 for LED) and it is the number of irrelevant features the one that varies. When the number of samples increases, the new instances are randomly generated.

3.2 Evaluation metrics

The goal is to determine the best method is terms of some evaluation measures. In this research, F-score, Jaccard-index, and training time were considered,

- The *F-score* is defined as the harmonic mean between precision and recall,

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

  Precision is computed as the number of relevant features selected divided by the number of features selected, and *recall* is the number of relevant features selected divided by the total number of relevant features.
- The *Jaccard-index*, or Jaccard similarity coefficient, is a metric used for comparing the diversity of a set of samples (in this case, set of features). It is defined as the cardinality of the intersection divided by the cardinality of the union of the sets A and B,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  Stability (similarity) of the selected features is an important aspect when the task is machine learning, not merely returning an accurate classifier [10].

Motivated by the methodology proposed in [11], we define three figures from which eight scalar measures are extracted. Note that the evaluation of feature selection algorithms relies on the bi-dimensional features-samples space (X-Y-axes). So, these evaluation measures shape a surface (Z-axis) in a three-dimensional space.

- F-score surface: *Feature size* vs *Sample size* vs *F-score.* It is obtained by displaying the evolution of the F-score across the features-samples space. The following scalar measures are computed,
  1. *F-score*: maximum F-score.
  2. *Fs95%*: the minimum amount of data (*features* × *samples*) for which the F-score rises above a threshold (95% of *F-score*).
  3. *VuFs*: volume under the F-score surface.
- Jaccard-index surface *Feature size* vs *Sample size* vs *Jaccard-index.* It is obtained by displaying the evolution of the Jaccard-index across the features-samples space.
  4. *Jaccard-index*: maximum Jaccard-index.
  5. *Ji95%*: the minimum amount of data (*features* × *samples*) for which the Jaccard-index rises above a threshold (95% of *Jaccard-index*).
  6. *VuJi*: volume under the Jaccard-index surface.
- Training time surface: *Feature size* vs *Sample size* vs *Training time.* It is obtained by displaying the evolution of the training time across the features-samples space.
  7. *Training time*: training time in seconds.

8. *VuTt*: volume under the training time surface.

Those measures related to F-score and Jaccard-index (i.e. F-score, VuFs, Jaccard-index, and VuJi) are desirable to be maximized, whilst those related to amount of data and time (i.e. Fs95%, Ji95%, Training time, and VuTt) are desirable to be minimized.

3.3 Results

This section shows the scalability results according to the measures explained above. Figure 1 plots the measures of scalability of CFS, consistency-based and INTERACT. In general terms, F-score and Jaccard-index are more influenced by sample size whilst training time is more affected by feature size (notice that in Figures 1(e) and 1(f) the X-Y axes are shifted for purposes of visualization).

Figures 1(a) and 1(c) show a better performance of consistency-based filter in comparison with the others in both F-score and Jaccard-index. This filter maintains its performance with the increase of the feature size whereas CFS and INTERACT deteriorate. With regard to the LED dataset, the three filters show a constant behavior in both F-score and Jaccard-index (see Figures 1(b) and 1(d)).

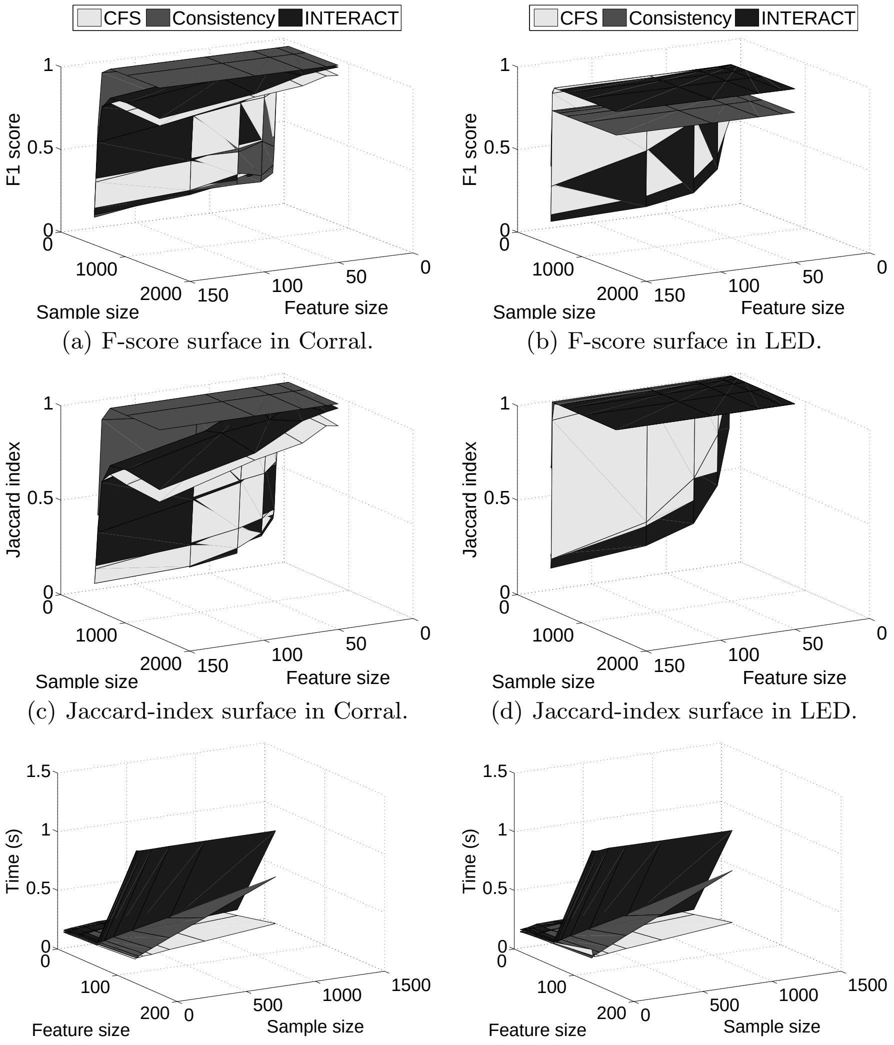
Regarding the training time (see Figures 1(e) and 1(f)), INTERACT is sharply affected by the feature size (remaining almost constant with respect to the sample size). On the contrary, consistency-based is mostly influenced by the sample size. Finally, CFS is very efficient with respect to the training time (remaining almost constant for both feature and sample size).

Table 1. Evaluation metrics.

Dataset	Filter	F-score	Fs95%	VuFs	J-index	Ji95%	VuJi	Time	VuTt
Corral	CFS	0.96	1024	17.80	0.93	1024	14.76	0.32	5.30
	Consistency	1.00	1024	19.03	1.00	2048	16.89	0.72	6.19
	INTERACT	0.99	1024	18.29	0.98	2048	15.54	1.11	9.06
LED	CFS	0.85	256	21.32	1.00	128	24.74	0.33	5.52
	Consistency	0.71	128	19.15	1.00	128	26.68	0.77	6.46
	INTERACT	0.85	256	21.28	1.00	256	24.63	1.12	9.22

Table 1 depicts the eight scalar measures related with Figure 1. These results confirm the trends seen in Figure 1, reflecting the adequacy of these measures which are reliable and confident and can give an idea of the scalability properties of the filter methods.

In light of these results, INTERACT shows the worst behavior in terms of scalability (it needs more data than the others and shows the longest training time), CFS is more efficient in terms of time than the others, and consistency-based filter exhibits a good tradeoff between training time and the rest of the measures.



(a) F-score surface in Corral.

(b) F-score surface in LED.

(c) Jaccard-index surface in Corral.

(d) Jaccard-index surface in LED.

(e) Training time surface in Corral.

(f) Training time surface in LED.

Fig. 1. Measures of scalability of CFS, consistecy-based, and INTERACT filters in the Corral (Figures a, c, and e) and LED datasets (Figures b, d, and f).

4 Conclusions

An algorithm is said to be scalable if it is suitable efficient and practical when applied to large databases. However, the current state is that the issue of scalability is far from being solved although it is present in a diverse set of problems such as learning, clustering, or feature selection.

In this research, we focus our attention on the scalability of feature selection that has not received much consideration in the literature. Three well-known filter-based feature selection methods were evaluated over two synthetic datasets in terms of several new measures proposed in this paper. They are not only based in accuracy but also in execution time and stability, and their adequacy was demonstrated.

For future work, we plan to extend this research to other datasets and feature selection methods (filters, wrappers, and embeded) in order to draw reliable conclusions. It is also interesting to check how different search strategies affect to scalability. Finally, a methodology for fusing the proposed measures seems to be necessary to rank the methods.

Acknowledgements

This work was supported in part by the Secretaría de Estado de Investigación of the Spanish Goverment (Grant code TIN2009-10748) and by the Xunta de Galicia (project CN2011/007 and *Plan I2C* Grant Program), and the European Union by FEDER funds.

References

1. Z.A. Zhao and H. Liu. *Spectral Feature Selection for Data Mining.* CRC Press, 2012.
2. I. Guyon. *Feature extraction: foundations and applications*, volume 207. Springer Verlag, 2006.
3. L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.
4. M.A. Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.
5. M. Dash and H. Liu. Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2):155–176, 2003.
6. Z. Zhao and H. Liu. Searching for interacting features. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1156–1161. Morgan Kaufmann Publishers Inc., 2007.
7. LA Belanche and FF González. Review and evaluation of feature selection algorithms in synthetic problems. *Arxiv preprint arXiv:1101.2320*, 2011.
8. G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the eleventh international conference on machine learning*, volume 129, pages 121–129. San Francisco, 1994.
9. L. Breiman. *Classification and regression trees.* Chapman & Hall/CRC, 1984.
10. L.I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, pages 390–395. ACTA Press, 2007.
11. S. Sonnenburg, V. Franc, E. Yom-Tov, and M. Sebag. PASCAL Large Scale Learning Challenge. *Journal of Machine Learning Research*, 2009.