ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Do we need hundreds of classifiers or a good feature selection?

Laura Morán-Fernández and Verónica Bolón-Canedo and Amparo Alonso-Betanzos *

CITIC, Universidade da Coruña, A Coruña, Spain

Abstract.
The task of choosing the appropriate classifier for a problem is not an easy-to-solve question due to the high number of algorithms available belonging to different families. Most of these classification algorithms exhibit a degradation in the performance when faced with many irrelevant and/or redundant features. Thus, in this work we analyze the impact of feature selection in classification. Experimental results over ten synthetic datasets show that the significance of selecting a classifier decreases after applying an appropriate preprocessing step and, not only this alleviates the choice, but it also improves the results in almost all classifiers tested.

1 Introduction

Classification is essential to data analytics, pattern recognition and machine learning. It arises from the need of making predictions of a categorical variable, known as a class variable, from one or more attribute variables which can be either categorical or numeric. A data instance (e.g. a patient potentially having cancer) is characterized by a number of independent variables (features), e.g. tumor markers (substances found in the blood, urine, stool, other bodily fluids, or tissues of the patient). It also has a response variable, e.g. whether the patient has a benign or a malignant tumor. When a data analyzer or researcher faces the classification of a dataset, the objective is usually to select the classifier which more probably achieves the best performance. However, this is a hard task due to the high number of classifiers arising from many different families. According to the No-Free-Lunch theorem, the best classifier will not be the same for all the datasets [1]. Despite this, Fernández-Delgado et al. [2] presented an exhaustive evaluation of 179 classifiers over 121 datasets. They stated that the classifiers most likely to be the best were random forest and support vector machines. However, Wainberg et al. [3] showed that the previous study's results are biased by the lack of a held-out test set and the exclusion of trials with errors, calling into question that conclusion.

Theoretically, having more features should give more discriminating power. However, experimental evidence has shown that this is not always the case [4]. Decision trees, such as C4.5, exhibit a degradation in the performance when faced with many irrelevant features. Similarly, instance-based features, such as kNN, are also very susceptible to irrelevant features. It has been shown that the number of training samples needed to produce a predetermined level of performance for instance-based learning increases exponentially with the number of irrelevant features [5]. On the other hand,

*This research has been financially supported in part by the Spanish Ministerio de Economía y Competitividad (research project TIN2015-65069-C2-1-R), by European Union FEDER funds and by the Xunta de Galicia (research projects GRC2014 /035 and ED431G/01).

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

algorithms such as naive Bayes are robust with respect to irrelevant features, degrading their performance very slowly when more irrelevant features are added [6]. However, the performance of such algorithms deteriorates quickly by adding redundant features, even if they are relevant to the concept.

For these reasons, researchers began to apply feature selection in a pre-processing phase, with the goal of determining the "best" subset of features that accurately describes a given problem with a minimum degradation of the performance [7]. This arises the question of which is the effect of feature selection in classification. Thus, in this paper we analyze if the application of a good preprocessing step can alleviate the choice of the classification algorithm and also if its impact improves the accuracy over ten synthetic datasets.

2 Feature selection techniques

Feature selection methods have received a great deal of attention in the classification literature [8], which largely reflects filter, wrapper and embedded methods. The essential difference between the first two is that the wrapper methods, unlike the filter methods, make use of the classifier that will be used to build the final classifier. As for embedded methods, these are generally used to classify machine learning models, with the classification algorithm building an optimal subset of features. Since wrapper and embedded methods interact with the classifier, we opted for filter methods.

Filter methods evaluate the goodness of data subsets by observing only intrinsic data characteristics and evaluating a single feature or subset against the class label. Below we describe the five filters used in the experimental study (all implemented in the Weka environment [9]).

- **Correlation-based Feature Selection** (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function [10]. This function is biased towards subsets containing features that are highly correlated with the class and uncorrelated with each other.
- The **Consistency-based Filter** (CONS) evaluates the worth of a features subset according to consistency in class values when training samples are projected onto the features subset [11].
- The **INTERACT** (INT) algorithm is based on symmetrical uncertainty and it also includes the consistency contribution [12].
- **Information Gain** (IG) filter evaluates the features according to their information gain and and considers a single feature at a time [13].
- **ReliefF** (Rf) algorithm estimates features according to how well their values distinguish among the instances that are near to each other [14].

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

3 Synthetic datasets

The first step to test the effectiveness of a feature selection method should be on synthetic data, since the knowledge of the optimal features and the chance to modify the experimental conditions allows to draw more useful conclusions.

The datasets chosen for this study try to cover different problems: increasing number of irrelevant features, redundancy, noise, alteration of the inputs, nonlinearity of the data etc. These factors complicate the task of the feature selection methods, which are very affected by them. Besides, some of the datasets have a significantly higher number of features than samples, which implies an added difficulty for the correct selection of the relevant features.

Table 1 shows a summary of the different problems covered by the synthetic datasets employed, as well as the number of features and samples and the relevant features which should be selected by the feature selection methods.

Table 1: Summary of the synthetic datasets. It shows the number of samples (#sam.), the number of features (#feat.), the relevant features (#rel-feat.) and the number of classes (#cl.), as well as the presence of correlation (#corr.), noise and no linearity. G_i means that the feature selection method must select one feature within the i-th group of features.

Dataset	#sam.	#feat.	#rel-feat.	Corr.	Noise	No linear	#cl.	Ref.
CorrAL-100	99	32	1-4	✓			2	[15]
XOR-100	99	50	1-2			✓	2	[15]
Parity3+3	12	64	1-3			✓	2	[4]
Led-25	24	50	1-7		✓		10	[16]
Led-100	99	50	1-7		✓		10	[16]
Monk3	122	6	2,4,5		✓		2	[17]
SD1	75	4020	G_1, G_2				3	[18]
SD2	75	4040	$G_1 - G_4$				3	[18]
SD3	75	4060	$G_1 - G_6$				3	[18]
Madelon	2400	500	1-5		✓	✓	2	[19]

4 Experimental results

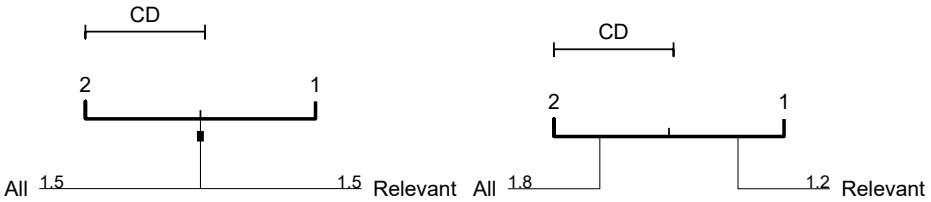
In this section, the results after applying five different feature selection methods over ten synthetic datasets will be presented. While three of the feature selection methods return a feature subset (CFS, CONS and INTERACT), the other two (IG and ReliefF) are ranker methods, so a threshold is mandatory in order to obtain a subset of features. In this work we have opted for retaining the top 10%, 20% and $log_2(n)$ [20] of the most relevant features of the ordered ranking, where n is the number of features in a given dataset. In the case of SD datasets, due to the mismatch between dimensionality and

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

sample size, the thresholds selected the top 5%, 10% and $log_2(n)$ features, respectively. We computed 5-fold cross validation to estimate the error rate.

The behaviour of the feature selection methods will be tested according to the classification error obtained by five different classifiers, each belonging to a different family. The classifiers employed were: two linear (naive Bayes and Support Vector Machine using a linear kernel) and three nonlinear (C4.5, k-Nearest Neighbor with $k = 3$ and Random Forest). All five classifiers were executed using the Weka [9] tool, using default values for the parameters.

In order to check if the importance of choosing a specific classifier decreases after applying a good preprocessing step, we analyzed the standard deviation of the classification error obtained by the five classifiers. We consider that a lower value of standard deviation represents a lower influence of the classifier selected. Thus, to explore the statistical significance of our classification results, we analyzed the standard deviation by using a Friedman test with the Nemenyi post-hoc test. Figure 1 presents the critical different diagrams, introduced by Demšar [21], where groups of methods that are not significantly different (at $\alpha = 0.10$) are connected.



(a) All classifiers (b) Non-linear classifiers

Fig. 1: Critical difference diagram showing the difference in terms of standard deviation between the error obtained by the five classifiers over the ten synthetic datasets.

As we are dealing with synthetic datasets, the relevant features of each dataset are known (Table 1). Thus, firstly we compared the results obtained by the classifiers over the original datasets (All), i.e. without feature selection, and then the datasets with the relevant features (Relevant). As can be seen in Figure 1(a), the classifiers perform better on average over the datasets with only the relevant features but with no statistical significance over the classifier using the original data. However, three nonlinear problems are tested: XOR-100, Parity3+3 and Madelon. Then, for these datasets, classification errors obtained by linear classifiers (naive Bayes and SVM) are not taking into account in Figure 1(b). As a result, statistical significance appeared between the two approaches, which demonstrates our initial hypothesis.

However, we cannot trust that any feature selection method is able to select the relevant features, so the next step is to test the behavior of five different feature selection methods as a preprocessing step before classification. Figure 2 shows that, although not all the feature selection methods are statistically significant with respect to the results obtained over the version using all the features of the dataset (All), they always outperform it.

Finally, we compared the performance of the feature selection methods over the five

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

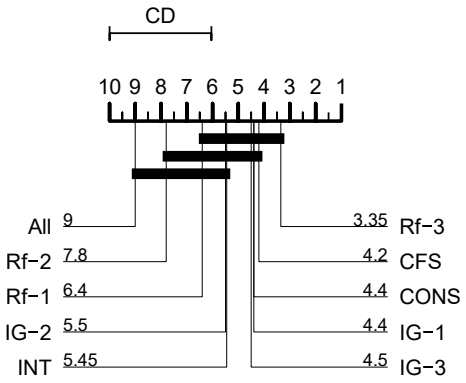


Fig. 2: Critical difference diagram showing the difference in terms of standard deviation between the error obtained by the five classifiers over the ten synthetic dataset. IG/Rf-1, IG/Rf-2, IG/Rf-3 refer to Information Gain/ReliefF filter using three different threshold.

different classifiers, trying to study which classification algorithm benefits more from the preprocessing phase. As can be seen in Table 2, all the classifiers achieved lower classification errors after applying feature selection except classifier C4.5. It has to be noted that this classifier performs an embedded selection of the features; therefore, it may be using a subset of features smaller than the given by the feature selection method.

Table 2: Average classification error. IG/Rf-1, IG/Rf-2, IG/Rf-3 refer to Information Gain/ReliefF filter using three different thresholds. Lower errors obtained by the filters versus the All approach are highlighted in bold.

	CFS	INT	CONS	IG-1	IG-2	IG-3	Rf-1	Rf-2	Rf-3	All
C4.5	39.95	38.87	40.53	48.68	38.95	39.30	48.80	45.00	40.44	38.64
NB	**37.12**	**35.50**	**38.63**	**51.43**	**50.26**	**39.05**	**44.65**	**43.72**	**38.27**	52.67
3-NN	**41.84**	**41.27**	**43.83**	**53.11**	**44.59**	**39.21**	47.64	**51.80**	**42.34**	47.30
SVM	**41.20**	**39.20**	**40.24**	**49.67**	**47.63**	**39.25**	**43.76**	**44.58**	**42.38**	51.77
RF	**37.89**	**37.47**	**39.80**	47.18	**39.65**	**35.38**	**41.81**	**44.58**	**39.53**	44.61

5 Conclusions

Feature selection has been an active and fruitful field of research in machine learning. In this paper, we analyze the effect of this preprocessing task on classification over ten synthetic datasets. The suite of synthetic datasets chosen covers phenomena such as presence of redundant and irrelevant features, interaction between features or non-linearity. In light of the results, we can conclude that: (i) the choice of a classifier is less critical if we apply a good feature selection method before the classification task and (ii) not only alleviates the choice but also improves the results in almost all classifiers.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Moreover, the results obtained by Random Forest support the conclusions of its power in previous studies [2, 3] and, in this case, specially thanks to the impact of feature selection.

As future work, we plan to extend this study to other scenarios such as real datasets.

References

[1] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[2] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[3] Michael Wainberg, Babak Alipanahi, and Brendan J Frey. Are random forests truly the best classifiers? *The Journal of Machine Learning Research*, 17(1):3837–3841, 2016.

[4] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519, 2013.

[5] Pat Langley and Wayne Iba. Average-case analysis of a nearest neighbor algorithm. In *IJCAI*, volume 93, pages 889–889. Citeseer, 1993.

[6] Ron Kohavi, George H John, et al. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[7] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[8] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, 86:33–45, 2015.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[10] Mark Andrew Hall. Correlation-based feature selection for machine learning. 1999.

[11] Manoranjan Dash and Huan Liu. Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2):155–176, 2003.

[12] Zheng Zhao and Huan Liu. Searching for interacting features in subset selection. *Intelligent Data Analysis*, 13(2):207–228, 2009.

[13] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. 1998.

[14] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer, 1994.

[15] Gilhan Kim, Yeonjoo Kim, Heuiseok Lim, and Hyeoncheol Kim. An mlp-based feature subset selection for hiv-1 protease cleavage site analysis. *Artificial intelligence in medicine*, 48(2-3):83–89, 2010.

[16] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[17] Sebastian B Thrun, Jerzy Bala, Eric Bloedorn, Ivan Bratko, Bojan Cestnik, John Cheng, Kenneth De Jong, Saso Dzeroski, Scott E Fahlman, D Fisher, et al. The monk's problems a performance comparison of different learning algorithms. 1991.

[18] Zexuan Zhu, Yew-Soon Ong, and Jacek M Zurada. Identification of full and partial class relevant genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2):263–277, 2010.

[19] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[20] Borja Seijo-Pardo, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos. On developing an automatic threshold applied to feature selection ensembles. *Information Fusion*, 45:227–245, 2019.

[21] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.