Feature Selection in Statistical Classification

Matthias Kohl*

Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany

Abstract: We give a brief overview of feature selection methods used in statistical classification. We cover filter, wrapper and embedded methods.

Keywords: Statistical classification, supervised statistical learning, machine learning, curse of dimensionality, over-fitting, feature selection, filter, ranker, wrapper, embedded methods.

INTRODUCTION

We consider the following model

$$Y = f(X_1, X_2, \ldots X_k) + \varepsilon$$

i.e., the outcome Y (*dependent variable*) depends in some unknown way f from the input X_1, X_2, ..., X_k (*independent or explanatory variables* also called *features*) where the observation of Y is additionally disturbed by some random error ε. The goal is to predict Y given X_1, X_2, ..., X_k; that is, given some (training) dataset we try to find a good approximation (estimate) for the unknown function f. This is also called *supervised statistical learning* or *machine learning* where in machine learning one typically does not explicitly consider random errors. If Y is a quantitative measurement this is called *regression*. In case Y is a qualitative output (categorical or discrete variable) where the values of Y correspond to certain categories or classes, it is called *classification*; for more details and examples see [1]. The main goal in statistical classification is to classify subjects with the highest possible accuracy where accuracy is measured in terms of so-called performance measures; see [2] for an overview of important performance measures for binary classification.

In the case of high dimensional data, i.e. with a large number of explanatory variables, most classification algorithms cannot be directly applied. The main reason is the *curse of dimensionality*: with increasing dimensions the volume of the search space increases very fast and the data becomes sparse. As a consequence the distances among the observations assimilate and the classification algorithm is not able to establish boundaries between the classes. This effect is further intensified by noisy and uninformative features. A noisy feature is a variable that is not related to Y. In case of high-dimensional data the number of noisy variables is often orders of magnitude greater than the number of informative variables. Furthermore, given k features the set of all subsets (power set) is of order 2^k, hence already for a moderate number k it becomes practically impossible to compute the classifier for all possible combinations of features. A solution to these problems is to apply the learning algorithm only to appropriately chosen subsets of the explanatory variables. Such a feature selection often improves the classification accuracy and reduces the risk of *over-fitting* [3, 4]. When there are a large number of explanatory variables, maybe even orders of magnitude more features than observations, there is a good chance to find some complex model that achieves a high accuracy for the training data. However, such a complex model is often very specific for the training data but its accuracy for new external data is poor. This is called over-fitting or lack of generalisability.

FEATURE SELECTION METHODS

Feature selection methods try to find the subset of features with the highest predictive power. In case of high-dimensional data, e.g. microarray data, one often uses non-specific or independent filtering, using criteria such as overall variance, as a first step in order to remove noisy and uninformative features. A non-specific or independent filtering is independent of the observed classes (categories) and can increase the power of subsequent analyses [5].

The (specific) filtering strategies are usually divided into *filters*, *wrappers*, and *embedded methods*; see [3] for a comprehensive overview. Filters consist of procedures that operate independently of the actual

*Address correspondence to this author at the Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany; Tel: +49 (0) 7720 307-4746; Fax: +49 (0) 7720 307-4727; E-mail: Matthias.Kohl@hs-furtwangen.de

E-ISSN: 1929-6029/12

learning algorithm by using general characteristics of the data to judge the predictive power of the features. They can further be divided into rankers and feature subset evaluation methods. Rankers generate a ranked list of the features by evaluating each feature independently using; e.g., statistical tests, fold changes, single variable classifiers, or information theoretic criteria. These techniques are simple and computationally very efficient. Their main disadvantage is that they neglect potential interactions and correlations between explanatory variables making it very likely that they fail in situations where it is not individual features but only certain combinations of them are informative. One way out is through the use of subset evaluation methods which judge the usefulness of subsets of the features using for example correlation based metrics or information theoretic methods such as Markov blanket algorithms [6]. As there are 2^k subsets an exhaustive search of the feature space with increasing number of features quickly becomes computationally intractable and only efficient search strategies can be applied in the case of high-dimensional data; examples are a forward selection, backward elimination, branch-and-bound, simulated annealing, or genetic algorithm.

The wrapper feature selection methods use the actual learning algorithm to measure the predictive power of feature subsets. They combine an efficient iterative search algorithm on feature subsets with performance measures for classification where the assessment of the performance is based on a validation set or by using re-sampling methods such as cross-validation or bootstrap. The combination of several algorithms makes wrappers the most time-consuming of all feature selection strategies [3].

In the case of embedded methods such as decision trees feature selection is fully incorporated into the learning algorithm yielding computationally more efficient procedures than wrappers. In addition, the available data is used in a better way as there is no need to split the training data into a training and validation set [3].

The three feature selection strategies are quite different and each method has its own strengths and weaknesses. The choice of the feature selection method is problem dependent; see [3] for a check list. Filters are expected to perform well for small training datasets where they yield stable models. For larger sample sizes, wrappers and embedded methods, which in principle incorporate interactions and correlations between features, often lead to better predictive models [7, 8].

ACKNOWLEDGEMENTS

We would like to thank two anonymous referees for valuable comments on the manuscript.

REFERENCES

[1] Hastie T., Tibshirani R., and Friedman J. The elements of statistical learning. 2nd edition. Springer 2009. http://dx.doi.org/10.1007/978-0-387-84858-7

[2] Kohl M. Performance measures in binary classification. Int J Stats Med Res 2012 Sep; 1(1): 79-81.

[3] Guyon I. and Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research 2003 Mar; 1157–1182, Mar 2003.

[4] Navot A., Gilad-Bachrach R., Navot Y. and Tishby N. mIs Feature Selection Still Necessary? In: Saunders C., Grobelnik M., Gunn S., and Shawe-Taylor J., editors. mSubspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science, volume 3940; Springer 2006: p. 127-138.

[5] Bourgon R., Gentleman R., and Huber, W. Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. U.S.A. 2010 May; 107(21): 9546-51. http://dx.doi.org/10.1073/pnas.0914005107

[6] Koller D. and Sahami M. Toward optimal feature selection. In: 13th International Conference on Machine Learning, July 1996: p. 284-292.

[7] Hall M.A. and Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans. Know. Data Eng. 2003 May/Jun; 15(3): 1437-47. http://dx.doi.org/10.1109/TKDE.2003.1245283

[8] Pochet N., De Smet F., Suykens J.A.K., and De Moor B.L.R. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004 Nov; 20(17): 3185-95. http://dx.doi.org/10.1093/bioinformatics/bth383

Received on 18-11-2012 Accepted on 12-12-2012 Published on 21-12-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.02.11