Computational Statistics and Data Analysis 71 (2014) 681–693

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Analysis of feature selection stability on high dimension and small sample data

David Dernoncourt[a,b,*], Blaise Hanczar[c], Jean-Daniel Zucker[a,d]

[a] Institut National de la Santé et de la Recherche Médicale, U872, Nutriomique, Équipe 7, Centre de Recherches des Cordeliers, 75006, Paris, France

[b] Université Pierre et Marie-Curie - Paris 6, 75006, Paris, France

[c] LIPADE, Université Paris Descartes, 45 rue des Saint-Pères, Paris, F-75006, France

[d] Institut de Recherche pour le Développement, IRD, UMI 209, UMMISCO, France Nord, F-93143, Bondy, France

ARTICLE INFO

Article history:
Received 30 June 2012
Received in revised form 7 July 2013
Accepted 7 July 2013
Available online 18 July 2013

Keywords:
Feature selection
Small sample
Stability
Low N/D ratio

ABSTRACT

Feature selection is an important step when building a classifier on high dimensional data. As the number of observations is small, the feature selection tends to be unstable. It is common that two feature subsets, obtained from different datasets but dealing with the same classification problem, do not overlap significantly. Although it is a crucial problem, few works have been done on the selection stability. The behavior of feature selection is analyzed in various conditions, not exclusively but with a focus on t-score based feature selection approaches and small sample data. The analysis is in three steps: the first one is theoretical using a simple mathematical model; the second one is empirical and based on artificial data; and the last one is based on real data. These three analyses lead to the same results and give a better understanding of the feature selection problem in high dimension data.

1. Introduction

Classification tasks in which the number of features D is much larger than the number of samples N are an increasingly frequent problem and became recently a research area of its own (Hastie et al., 2009). For instance, in computational biology, microarray data contain the simultaneous expression of tens of thousands of genes, and metagenomic data contain in the order of a few millions of genes…usually measured on (at most) a few hundreds patients. High dimensionality and small sample size pose a challenge to classification techniques, since they both increase the risk of overfitting and decrease the accuracy of classifiers (Jain and Chandrasekaran, 1982). Moreover, high dimensionality can increase computation time beyond reasonable limits, as classifiers usually do not scale too well to huge numbers of features. To deal with these problems, feature selection is used to reduce data dimensionality.

Feature selection refers to the process of removing irrelevant or redundant features from the original set of features $\mathcal{F} = \{f_1, f_2, \dots, f_{|\mathcal{F}|=D}\}$, so as to retain a subset $S \subset \mathcal{F}$ containing only informative features useful for classification. Feature selection methods can be broken down into three categories: filter, wrapper and embedded methods (Saeys et al., 2007). It is generally agreed that wrappers or embedded methods should be preferred if technically feasible (Pudil and Somol, 2008), however, on very high dimensional data, filters remain the method of choice for tractability reasons, which is why we will focus on them.

[*] Correspondence to: Centre de Recherches des Cordeliers, Équipe 7 Nutriomique, 15 rue de l'École de Médecine, 75006 Paris, France. Tel.: +33 1 44 27 80 76.

E-mail addresses: me@daviddernoncourt.com, david.dernoncourt@crc.jussieu.fr (D. Dernoncourt).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.012

Beyond classification performance, the other main objective of feature selection is to obtain a reliable and robust list of predictive variables (signature). A good signature must not overfit the available data and be exportable to other datasets related to the same classification problem. These conditions cannot be respected if the subset of selected features is highly variable. A lot of examples in the literature show that in small-sample or high dimension settings, the feature selection is not stable. For instance, in Miecznikowski et al. (2010), five classification tasks dealing with a similar problem (breast cancer prognosis prediction from gene expression data) were performed on five different datasets, leading to highly variable results of the individual gene analysis. Several other studies, such as Ioannidis (2005), Michiels et al. (2005), Ein-Dor et al. (2006) and Haury et al. (2011), emphasized the difficulty to obtain a reproducible gene signature on high-dimension small-sample data. This difficulty to find a common subset of predictors between such different but similar datasets, or even between different sample subsets from a same dataset, raises the problem of feature selection stability.

Few studies have already dealt with this problem, and most of them have focused on comparing the stability of different, pre-existing or new feature selection methods, without exploring how different types of variations in the training sets affect this stability (for instance, Kalousis et al., 2005; Somol et al., 2009 and Yao and Wang, 2013). Moreover, they most often used stability measures which could be biased by the proportion of selected features (most stability measures artificially increase when the proportion of selected features increases) or by the amount of non-selected features (some stability measures take into account the stability of both selected and unselected features, so can be excessively high on datasets containing a large proportion of easy to exclude, irrelevant features). In this work, we investigate the behavior of the feature selection stability and its impact on the classifiers. We first present the main measures of selection stability used in machine learning and propose corrections of some of them that are biased. Then we present our analysis of the behavior of feature selection in three steps. In the first step, we present a theoretical analysis of the performance and stability of feature selection on a simple Gaussian model. The second step is an empirical analysis performed on a large number of simulations based on artificial data. In the last step we present results of selection stability on real data. These three analyses lead to the same conclusions: in high dimensions feature selection is not stable and the probability for relevant features to be selected can be very low.

2. Stability measures

The stability of a feature selection method was defined in Kalousis et al. (2007) as *the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution*. To evaluate this robustness, quite a few different stability measures have already been described. We follow the taxonomy presented by Somol and Novovičová (2010), who distinguished:

- *feature-focused* versus *subset-focused* measures: the former evaluate feature selection frequencies over all feature subsets considered together as a whole, while the latter evaluate similarities within every pairs of selected feature subsets. Both types provide complementary information, so we want to have at least one of each.
- *selection-registering* versus *selection-exclusion-registering* measures: the first only considers the stability of selected features while the latter also measures the stability of excluded features. On large datasets where a huge number of features are irrelevant and easy to exclude, *selection-exclusion-registering* measures will be strongly upward biased, so we will only be interested in *selection-registering* measures here.
- *subset-size-biased* versus *subset-size-unbiased* measures: the first yield values bounded more tightly than [0; 1], with most notably the lower bound strongly increasing with the proportion of selected features, the latter are adjusted to be actually bounded by [0; 1]. Obviously, for better generalization, we want to use *subset-size-unbiased* measures.

2.1. Relative weighted consistency, an unbiased feature-focused measure

Among the stability measures sorted in the above-mentioned taxonomy, only one was both *selection-registering* and *subset-size-unbiased*: the relative weighted consistency CW_{rel} (Somol and Novovičová, 2010). It was based on a *subset-size-biased* measure, the weighted consistency CW, corrected to be actually bounded by [0; 1] no matter the proportion of selected features. A value of 0 indicates the highest possible instability, while a value of 1 indicates the highest possible stability, i.e., if all feature subsets have the same cardinality, all subsets are identical.

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_\omega\}$ be a system of ω feature subsets obtained from ω runs of the feature selection routine on different samplings, $\Omega = \sum_{i=1}^{\omega} |S_i|$ be the total number of occurrences of any feature in $\mathcal{S}$ and F_f be the number of occurrences of feature $f \in \mathcal{F}$ in system $\mathcal{S}$. CW was defined as follows:

$$CW(\mathcal{S}) = \sum_{f \in X} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1}, \tag{1}$$

and CW_{rel} was then derived by adjusting CW on its minimal and maximal possible values $CW_{\min}$ and $CW_{\max}$:

$$CW_{\mathrm{rel}}(\mathcal{S}, \mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{\min}(\Omega, \omega, \mathcal{F})}{CW_{\max}(\Omega, \omega) - CW_{\min}(\Omega, \omega, \mathcal{F})}. \tag{2}$$