An Experimental Comparison of Feature Selection Methods on Two-Class Biomedical Datasets

P. Drotár, J. Gazda, Z. Smékal

PII: S0010-4825(15)00291-7
DOI: http://dx.doi.org/10.1016/j.compbiomed.2015.08.010
Reference: CBM2217

To appear in: *Computers in Biology and Medicine*

Received date: 18 June 2015
Revised date: 5 August 2015
Accepted date: 12 August 2015

Cite this article as: P. Drotár, J. Gazda and Z. Smékal, An Experimenta Comparison of Feature Selection Methods on Two-Class Biomedical Datasets *Computers in Biology and Medicine* http://dx.doi.org/10.1016/j.compbiomed.2015.08.010

An Experimental Comparison of Feature Selection Methods on Two-Class Biomedical Datasets

P. Drotár[a*], J. Gazda[b*], Z. Smékal[a*]

[a] Department of Telecommunications, Brno University of Technology, Technická 12, 61200 Brno, Czech Republic
[b] Department of Computers and Informatics, Technical University of Kosice, Letna 9, 0401 Kosice, Slovakia

Abstract

Feature selection is a significant part of many machine learning applications dealing with small-sample and high-dimensional data. Choosing the most important features is an essential step for knowledge discovery in many areas of biomedical informatics. The increased popularity of feature selection methods and their frequent utilisation raise challenging new questions about the interpretability and stability of feature selection techniques. In this study, we compared the behaviour of ten state-of-the-art filter methods for feature selection in terms of their stability, similarity, and influence on prediction performance. All of the experiments were conducted on eight two-class datasets from biomedical areas. While entropy-based feature selection appears to be the most stable, the feature selection techniques yielding the highest prediction performance are minimum redundance maximum relevance method and feature selection based on Bhattacharyya distance. In general, univariate feature selection techniques perform similarly to or even better than more complex multivariate feature selection techniques with high-dimensional datasets. However, with more complex and smaller datasets multivariate methods slightly outperform univariate techniques.

Keywords: feature selection, stability, classification performance, univariate FS, multivariate FS

1. Introduction

Classification tasks in which the number of features is much larger than the number of subjects are becoming more and more abundant in many research areas. High dimensional datasets frequently occur in text processing, combinatorial chemistry or bioinformatics. These datasets can contain tens of thousands of features while having available only hundreds (or usually less than hundred) samples. High dimensionality can negatively impact the performance of classifier by increasing risk of overfitting and prolonging the computational time. Moreover, there are the applications where one's intention is to identify a small group of features that may be descriptive for some phenomenon or may shed the light on some underlaying process.

There are two approaches to reduce the dimensionality of dataset: feature extraction and feature selection [1], [2], [3]. In case of the feature extraction, the new basis is chosen for the data and new features are derived from the original input. On the other hand, the main goal of the feature selection techniques is to reduce effects of high dimensionality on dataset and to find a subset of features from the entire feature set that can efficiently describe the data. Reducing the dimensionality of data helps to avoid the effects of the *curse of dimensionality* [4] that seriously degrades the ability of learning algorithms to develop robust models. As the reduced subset is usually significantly smaller than the set of the input features, the computation time of subsequent analysis is greatly reduced. In this study, we will focus only on feature selection.

When facing the issue of choosing a feature selection (FS) algorithm for some machine learning problem, the usual approach is first to try the simple univariate techniques and if these do not provide desired result move on and try more complex FS methods. In fact, there is no procedure or systematic approach for choosing the most suitable FS method for particular problem. The increasing number of FS techniques that are indeed very effective and sophisticated makes the problem of selecting the most suitable FS even more apparent [5], [6], [7]. The only available guidelines are previous experiences and comparative studies from literature [8], [9]. When evaluating the suitability of FS method we are concerned with two aspects: (i) stability of FS i.e. how the output of FS algorithm changes when the data change [10] and (ii) how efficiently data are described by the selected subset of features, i.e. what is the influence on prediction accuracy. The motivation for evaluating stability comes from the domain experts requiring a small set of discriminatory features that are robust to variations in the training dataset [10], [11].

There are several studies comparing FS techniques from different aspects, however the literature on the subject is rather limited. The first study introducing the term *stability* of FS that also evaluated stability of five FS methods was done by Kalousis in [10]. Similarly, Molous et al. [12] evaluated stability of five univariate FS techniques. These studies focus mainly on univariate methods and do not analyze new and more advanced techniques. A deeper analysis of the performance of feature selection in high-dimensional setting with focus on stability was provided by [13], however, here the authors again

*peter.drotar84@gmail.com