Expert Systems with Applications 42 (2015) 4632–4642

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Similarity of feature selection methods: An empirical study across data intensive classification tasks

Nicoletta Dessì, Barbara Pes *

Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

ARTICLE INFO

Article history:
Available online 7 February 2015

Keywords:
Data mining
Knowledge discovery
Feature selection
Similarity measures

ABSTRACT

In the past two decades, the dimensionality of datasets involved in machine learning and data mining applications has increased explosively. Therefore, feature selection has become a necessary step to make the analysis more manageable and to extract useful knowledge about a given domain. A large variety of feature selection techniques are available in literature, and their comparative analysis is a very difficult task. So far, few studies have investigated, from a theoretical and/or experimental point of view, the degree of similarity/dissimilarity among the available techniques, namely the extent to which they tend to produce similar results within specific application contexts. This kind of similarity analysis is of crucial importance when two or more methods are combined in an ensemble fashion: indeed the ensemble paradigm is beneficial only if the involved methods are capable of giving different and complementary representations of the considered domain. This paper gives a contribution in this direction by proposing an empirical approach to evaluate the degree of consistency among the outputs of different selection algorithms in the context of high dimensional classification tasks. Leveraging on a proper similarity index, we systematically compared the feature subsets selected by eight popular selection methods, representatives of different selection approaches, and derived a similarity trend for feature subsets of increasing size. Through an extensive experimentation involving sixteen datasets from three challenging domains (Internet advertisements, text categorization and micro-array data classification), we obtained useful insight into the pattern of agreement of the considered methods. In particular, our results revealed how multivariate selection approaches systematically produce feature subsets that overlap to a small extent with those selected by the other methods.

1. Introduction

While data intensive applications are fast increasing in scope and sophistication, the extraction of useful knowledge from the large amounts of available data can be a very difficult task (Kumar & Minz, 2014; Liao, Chu, & Hsiao, 2012). One of the most critical issues for data manipulation and analysis is high dimensionality, i.e. the presence of a huge number of attributes (features) that are associated with each problem instance in the dataset. This can cause a number of drawbacks such as reduced performance, large computational time, and the use of features that may be either redundant or irrelevant to the problem at hand.

A lot of research has focused on methods for effectively handling high dimensional datasets (Chandrashekar & Sahin, 2014; Khalid, Khalil, & Nasreen, 2014), with two main approaches existing in literature: mapping the original feature space to a new space with lower dimensions (Wang & Paliwal, 2003) or selecting a meaningful subset of the original features, hence discarding those irrelevant and redundant ones (Guyon & Elisseeff, 2003). This last approach, referred to as feature selection, has proved to be very effective in the context of high dimensional classification problems, enabling to improve predictive performance as well as to obtain faster and more cost-effective predictors, and to achieve a better understanding of the underlying domain.

Though many works have investigated the potential and limits of existing feature selection methods (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Tang, Alelyani, & Liu, 2014), the choice of the most appropriate method for a given task remains difficult. Indeed, while more and more feature selection algorithms are available, there is little theoretical support to find the "right" one for the problem at hand (Liu & Yu, 2005). Increasingly, real-world datasets are being handled by applying a number of selection techniques, instead of a single one, and then integrating their outputs in some way.

* Corresponding author. Tel.: +39 070 6758758; fax: +39 070 6758501.
E-mail addresses: dessi@unica.it (N. Dessì), pes@unica.it (B. Pes).

http://dx.doi.org/10.1016/j.eswa.2015.01.069
0957-4174/

As suggested by recent literature (Dittman, Khoshgoftaar, Wald, & Napolitano, 2012), when choosing a set of techniques for a classification task it is beneficial to evaluate their degree of consistency. Different techniques, indeed, may select different features depending on the search strategy and evaluation criteria adopted in the selection process. However, despite their specificities, two techniques can be similar in their behavior, i.e. they can systematically produce results that overlap to a great extent (Cannas, Dessì, & Pes, 2013).

A similarity-based analysis of feature selection techniques can provide useful insight for domain modeling and understanding: if a set of techniques are dissimilar, i.e. exhibit in general a different behavior, then there is more reason to have confidence in a feature selected by all these techniques. On the other hand, it is not surprising if similar techniques select the same features and it does not help to confirm the relevance of these features for the considered domain (Dessì, Pascariello, & Pes, 2013).

Furthermore, when multiple feature selection methods are systematically combined in an ensemble fashion (Altidor, Khoshgoftaar, Van Hulse, & Napolitano, 2011), a similarity evaluation of the methods in the ensemble should not be neglected: it would not be indeed beneficial to combine two or more methods that give almost identical results. Though it is recognized that diversity has a crucial role for the success of an ensemble learning strategy (Dietterich, 2000), most research work on ensemble feature selection has so far not given due consideration to this important issue. Existing ensemble approaches are mainly built on an "ad hoc" basis (Dutkowski & Gambin, 2007; Leung & Hung, 2010; Olsson & Oard, 2006; Yang, Zhou, Zhang, & Zomaya, 2010), depending on the specific problem at hand, and there is a lack of systematic studies aiming at providing insight on which methods should be combined, and how this combination should be made, based on the degree of diversity/similarity of the involved methods.

In this paper, we aim to give a valuable contribution in this direction by investigating the similarity of eight popular feature selection techniques, representatives of different types of selection approaches. Specifically, we consider both univariate methods that evaluate each feature independently from the others as well as multivariate methods that take into account interdependencies among features. The similarity analysis is carried out in two stages: (i) the feature subsets produced by the chosen methods are compared, on a pair-wise basis, using a proper similarity index; (ii) the overall degree of consistency among the eight methods (or a specific group of them) is obtained by averaging similarity values over all the involved pair-wise comparisons. A similarity trend is also derived for feature subsets of increasing size.

The datasets used in the analysis come from three challenging domains: Internet advertisements, text categorization and microarray data classification. To the best of our knowledge, there is no study in literature that performs such a similarity analysis encompassing different real world application scenarios, as we do in this work.

The paper is organized as follows. Section 2 provides a survey of current literature and discusses related works. Section 3 describes all materials and methods involved in our empirical study, i.e. the adopted methodology, as well as the feature selection techniques and the datasets used for the experiments. The results of the analysis are presented and discussed in Section 4. Finally, Section 5 contains concluding remarks and future research directions.

2. Literature survey and related work

Feature selection is crucial to the analysis of high dimensional datasets coming from a number of application areas such as bioinformatics and text processing. It involves the exploration of the original feature space and the selection of the optimal feature subset based on a suitable relevance evaluation criterion (Kumar & Minz, 2014). According to whether the dataset is labeled or not, feature selection algorithms can be categorized into supervised (Song, Smola, Gretton, Borgwardt, & Bedo J., 2007), unsupervised (Dy & Brodley, 2004) and semi-supervised (Xu, King, Lyu, & Jin, 2010).

Supervised selection methods can be further categorized into *filter*, *wrapper* and *embedded* methods, depending on how they interact with the learning algorithm (classifier) that will be ultimately used to infer a model (Tang et al., 2014). Basically, *filter* approaches (Lazar et al., 2012) assess the relevance of features by looking only at the intrinsic properties of the data, without involving the use of a learning algorithm in the selection stage. In contrast, *wrapper* approaches (Guyon & Elisseeff, 2003) perform a search in the space of feature subsets and evaluate each subset by training and testing a specific classification model; hence wrappers are tailored to a specific learning algorithm, and may achieve better performance than filters methods, but at the price of a greater computational cost. Finally, *embedded* approaches (Ma & Huang, 2008) leverage the internal parameters of a classification algorithm to select relevant features, often providing a good trade-off between computational cost and performance.

A wide literature is currently available on the strengths and weaknesses of different feature selection methods (Bolón-Canedo et al., 2013; Hall & Holmes, 2003; Lazar et al., 2012; Saeys, Inza, & Larranaga, 2007), the choice of the "best" method being dependent on the specific problem at hand. Moreover, with the aim of devising suitable solutions for specific problem settings, new proposals are constantly appearing that exploit different strategies, e.g. (i) using different selection approaches (e.g. a filter and a wrapper) in different search stages (Cannas, Dessì, & Pes, 2011; El Akadi, Amine, El Ouardighi, & Aboutajdine, 2011), (ii) combining the outcomes of different feature selectors in an ensemble fashion (Altidor, Khoshgoftaar, Van Hulse, & Napolitano, 2011; Latkowski & Osowski, 2015) or (iii) combining feature selection with other approaches such as feature extraction (Bharti & Singh, 2015).

With such a body of algorithms available, their comparative analysis is a very difficult task. Most of the existing comparative studies focus on a specific application domain, such as text classification (Forman, 2003; Méndez, Fdez-Riverola, Díaz, Iglesias, & Corchado, 2006), genomic analysis (Abusamra, 2013; Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez, & Herrera, 2014), software defect prediction (Khoshgoftaar, Gao, Napolitano, & Wald, 2014), image classification (Staroszczyk, Osowski, & Markiewicz, 2012). A number of studies have been also conducted on artificially generated data (Bolón-Canedo et al., 2013) in order to evaluate the performance of selection methods under specific conditions (e.g. class imbalance, noise, redundancy and interaction between features).

To date, a quite neglected issue in feature selection literature is the theoretical and/or experimental assessment of the degree of consistency among the outputs of different selection methods. Indeed, it is known that different selection techniques may result in different feature subsets, especially when the high dimensionality is coupled with a small sample size (Saeys et al., 2007), but few direct comparisons exist that quantify these differences in a systematic way. Existing studies (as those cited above) mostly focus on comparing the outcomes of different techniques in terms of predictive performance or, less frequently, in terms of stability with respect to sample variation (Haury, Gestraud, & Vert, 2011; Kalousis, Prados, & Hilario, 2007; Wang, Khoshgoftaar, & Liang, 2013). However, as we showed in a previous work (Dessì et al., 2013), selection methods with a similar behavior in terms of accuracy and/or stability do not necessarily select similar feature subsets and, on the other hand, feature subsets with a good degree of overlapping do not necessarily result in similar classification