Computational Statistics and Data Analysis 71 (2014) 832–848

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimating mutual information for feature selection in the presence of label noise

Benoît Frénay [*], Gauthier Doquire, Michel Verleysen

Machine Learning Group, ICTEAM Institute, Université catholique de Louvain, Place du Levant 3, BE 1348, Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 4 April 2012
Received in revised form 1 May 2013
Accepted 1 May 2013
Available online 9 May 2013

Keywords:
Label noise
Mutual information
Entropy estimation
Feature selection

ABSTRACT

A way to achieve feature selection for classification problems polluted by label noise is proposed. The performances of traditional feature selection algorithms often decrease sharply when some samples are wrongly labelled. A method based on a probabilistic label noise model combined with a nearest neighbours-based entropy estimator is introduced to robustly evaluate the mutual information, a popular relevance criterion for feature selection. A backward greedy search procedure is used in combination with this criterion to find relevant sets of features. Experiments establish that (i) there is a real need to take a possible label noise into account when selecting features and (ii) the proposed methodology is effectively able to reduce the negative impact of the mislabelled data points on the feature selection process.

1. Introduction

Performing feature selection is an essential preprocessing step for many data mining and pattern recognition applications, including classification (Guyon and Elisseeff, 2003; Dash and Liu, 1997). The objective is to determine, among the original set of features of a data set, which are the most relevant ones to achieve a particular task. In practice, the benefits of feature selection are numerous. First, it helps reducing the dimensionality of a data set. This aspect is particularly important when the data are high-dimensional. Indeed, learning in this context is a hard task, due to many difficulties known under the generic term *curse of dimensionality* (Bellman, 1961). In addition, it is likely that for a specific problem, some features are either irrelevant or redundant. Discarding these features generally improves the performances of classification models. Last, feature selection has the advantage over other dimensionality reduction strategies, such as feature extraction (Guyon et al., 2006), that it preserves the original features. This is of crucial importance in many industrial and medical applications, where the interpretation of the models is important.

Among the different possible solutions, filter methods are often preferred to achieve feature selection. Filters are based on the optimisation of a criterion which is independent of any prediction model; in practice, this makes them particularly fast compared to wrapper methods, which directly optimise the performances of a specific prediction model. Moreover, filter methods can be used in combination with any prediction model; for these reasons, they will be considered in this work. As a criterion of relevance, Shannon's mutual information (MI) (Shannon, 1948) is one of the most popular and successful choices for filter-based feature selection. Due to several reasons described in Section 2.1, MI possesses many required qualities for this task and has strong advantages over other well-known criteria such as the correlation coefficient.

The major problem when using MI is that, in general, it cannot be computed analytically but has to be estimated from the available data. Even if estimating MI has been intensively studied for one-dimensional features, estimating the MI between high-dimensional groups of features still remains a challenging task; however, it can prove to be very useful in practice

[*] Correspondence to: ICTEAM/ELEN, Université catholique de Louvain, Place du Levant 3, BE 1348, Louvain-la-Neuve, Belgium. Tel.: +32 10 478133; fax: +32 10 472598.
E-mail address: benoit.frenay@uclouvain.be (B. Frénay).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.001

for feature selection. Recent works have addressed this problem, by showing the interest of a nearest-neighbours based MI estimator (Kraskov et al., 2004; Gómez-Verdejo et al., 2009).

Even if feature selection for traditional classification problems has been widely studied in the literature, it is somehow surprising that the impact of label noise on this task has not been investigated yet. To our knowledge, problems with feature selection were only mentioned by Zhang et al. (2006) and Shanab et al. (2012). In the particular context of gene selection, they show that only a few mislabelled samples cause a large percentage of the most discriminative genes to be not identified and that label noise decreases the stability of feature rankings. It is quite common when working with real-world datasets that some of the class labels are wrong (Brodley and Friedl, 1999). This can be due to the fact that, for many applications, human expertise is needed to assign class labels. Moreover, some errors can be made when labels are encoded in a data set. As label noise is known to have a negative impact on the performances of supervised classification algorithms, it is reasonable to assume that it will also degrade the performances of supervised feature selection algorithms. In this case, a label noise-tolerant feature selection algorithm would undoubtedly be of great interest.

First, the impact of label noise on a traditional MI-based filter feature selection algorithm is analysed, which shows how the performances of such an algorithm can decrease when the label noise increases. A solution to make a nearest neighbours based entropy estimator less sensitive to errors in the class labels is then proposed; the solution is based on a statistical model of the label noise and an expectation-maximisation algorithm.

The rest of the paper is organised as follows. Section 2 briefly reviews basic notions about MI-based feature selection and about the label noise problem; the impact of label noise on feature selection is also illustrated. Section 3 introduces a label noise-tolerant entropy estimator, assuming the true class memberships are known. An expectation-maximisation algorithm to estimate these memberships is derived in Section 4. The complete label noise-tolerant feature selection procedure is introduced in Section 5 and its interest is experimentally illustrated in Section 6. Section 7 concludes the paper.

2. Imprecise labels and feature selection

This section reviews basic concepts about mutual information (MI)-based feature selection and methods to handle label noise. The impact of the label noise on the performances of a classical MI-based supervised feature selection algorithm is eventually illustrated in an example.

2.1. Mutual information: definitions and interest for feature selection

Filter-based feature selection requires the use of a statistical criterion, measuring the relevance of a feature set for predicting the class labels. In this work, the mutual information (MI) (Shannon, 1948) criterion is considered. Let X denote a (group of) real-valued random variable(s) on domain $\mathcal{X}$ and Y a discrete random variable on domain $\mathcal{Y}$. In a feature selection context, X is a (group of) feature(s) and Y the associated class label. The MI between X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y), \tag{1}$$

where $H(X)$ is called the entropy of X. The entropy is

$$H(X) = -\int_{\mathcal{X}} p_X(x) \log p_X(x) dx, \tag{2}$$

p_X being the probability density function of X. In Eq. (1), $H(X|Y)$ is the conditional entropy of X given Y:

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y), \tag{3}$$

where p_Y is the probability mass function of Y. In the last equation, $H(X|Y = y)$ is the classical entropy of X, but limited to the points whose class label is y.

The MI criterion has many desirable properties for feature selection. First it has a natural interpretation in terms of uncertainty reduction. Indeed, it is symmetric and Eq. (1) can be equivalently rewritten as

$$I(X;Y) = H(Y) - H(Y|X). \tag{4}$$

Since the entropy measures the uncertainty on the observed values of a random variable, the MI can be seen as the reduction of uncertainty on the class labels once a (group of) feature(s) is known. This is obviously a sound criterion to assess the interest of a subset of features. Moreover, the MI has the advantage over other well-known criteria (such as the popular correlation coefficient, see e.g. Yu and Liu (2003)) that it is able to detect non-linear relationships between variables; it is thus more powerful in practice. Eventually, the MI can be naturally defined for multidimensional variables, which again is not the case for other popular criteria. This property can be particularly helpful for feature selection, since some features are often only relevant or redundant when considered together.

2.2. Search procedures

The objective of the feature selection method that is considered in this paper is to find the subset of the original features which together maximise the MI with the output Y. The most straightforward strategy is to try all possible feature subsets. However, such an exhaustive search is intractable in practice as the number of features gets large.