Expert Systems With Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature selection using Joint Mutual Information Maximisation

Q1 Mohamed Bennasar, Yulia Hicks, Rossitza Setchi*

School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

ARTICLE INFO

Keywords:
Feature selection
Mutual information
Joint mutual information
Conditional mutual information
Subset feature selection
Classification
Dimensionality reduction
Feature selection stability

ABSTRACT

Feature selection is used in many application areas relevant to expert and intelligent systems, such as data mining and machine learning, image processing, anomaly detection, bioinformatics and natural language processing. Feature selection based on information theory is a popular approach due its computational efficiency, scalability in terms of the dataset dimensionality, and independence from the classifier. Common drawbacks of this approach are the lack of information about the interaction between the features and the classifier, and the selection of redundant and irrelevant features. The latter is due to the limitations of the employed goal functions leading to overestimation of the feature significance.

To address this problem, this article introduces two new nonlinear feature selection methods, namely Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information Maximisation (NJMIM); both these methods use mutual information and the 'maximum of the minimum' criterion, which alleviates the problem of overestimation of the feature significance as demonstrated both theoretically and experimentally. The proposed methods are compared using eleven publically available datasets with five competing methods. The results demonstrate that the JMIM method outperforms the other methods on most tested public datasets, reducing the relative average classification error by almost 6% in comparison to the next best performing method. The statistical significance of the results is confirmed by the ANOVA test. Moreover, this method produces the best trade-off between accuracy and stability.

1. Introduction

High dimensional data is a significant problem in both supervised and unsupervised learning (Janecek, Gansterer, Demel, & Ecker, 2008), which is becoming even more prominent with the recent explosion of the size of the available datasets both in terms of the number of data samples and the number of features in each sample (Zhang et al., 2015). The main motivation for reducing the dimensionality of the data and keeping the number of features as low as possible is to decrease the training time and enhance the classification accuracy of the algorithms (Guyon & Elisseeff, 2003; Jain, Duin, & Mao, 2000; Liu & Yu, 2005).

Dimensionality reduction methods can be divided into two main groups: those based on feature extraction and those based on feature selection. Feature extraction methods transform existing features into a new feature space of lower dimensionality. During this process, new features are created based on linear or nonlinear combinations of features from the original set. Principal Component Analysis (PCA) (Bajwa, Naweed, Asif, & Hyder, 2009; Turk & Pentland, 1991) and Linear Discriminant Analysis (LDA) (Tang, Suganthana, Yao, & Qina, 2005; Yu & Yang, 2001) are two examples of such algorithms. Feature selection methods reduce the dimensionality by selecting a subset of features which minimises a certain cost function (Guyon, Gunn, Nikravesh, & Zadeh, 2006; Jain et al., 2000). Unlike feature extraction, feature selection does not alter the data and, as a result, it is the preferred choice when an understanding of the underlying physical process is required. Feature extraction may be preferred when only discrimination is needed (Jain et al., 2000).

Feature selection is used in many application areas relevant to expert and intelligent systems, such as data mining and machine learning, image processing, anomaly detection, bioinformatics and natural language processing (Hoque, Bhattacharyya, & Kalita, 2014). Feature selection is normally used at the data pre-processing stage before training a classifier. This process is also known as variable selection, feature reduction or variable subset selection.

The topic of feature selection has been reviewed in detail in a number of recent review articles (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Brown, Pocock, Zhao, & Lujan, 2012; Chandrashekar & Sahin, 2014; Vergara & Estévez, 2014). Usually, feature selection methods are divided into two categories in terms of evaluation strategy, in particular, classifier dependent ('wrapper' and 'embedded' methods) or classifier independent ('filter' methods). Wrapper methods search the feature space, and test all possible

* Corresponding author. Tel: +44 2920875720; fax: +44 2920874716.
E-mail addresses: BennasarM@cf.ac.uk (M. Bennasar), HicksYA@cf.ac.uk (Y. Hicks), Setchi@cf.ac.uk (R. Setchi).

http://dx.doi.org/10.1016/j.eswa.2015.07.007
0957-4174/

subsets of feature combinations by using the prediction accuracy of a classifier as a measure of the selected subset's quality, without modifying the learning function. Therefore, wrapper methods can be combined with any learning machine (Guyon et al., 2006). They perform well because the selected subset is optimised for the classification algorithm. On the other hand, wrapper methods may suffer from over-fitting to the learning algorithm. This means that any changes in the learning model may reduce the usefulness of the subset. In addition, these methods are very expensive in terms of computational complexity, especially when handling extremely high-dimensional data (Brown et al., 2012; Cheng et al., 2011; Ding & Peng, 2003; Karegowda, Jayaram, & Manjunath, 2010).

The feature selection stage in the embedded methods is combined with the learning stage. These methods are less expensive in terms of computational complexity and less prone to over-fitting; however, they are limited in terms of generalisation, because they are very specific to the used learning algorithm (Guyon et al., 2006).

Classifier-independent methods rank features according to their relevance to the class label in the supervised learning. The relevance score is calculated using distance, information, correlation and consistency measures. Many techniques have been proposed to compute the relevance score, including Pearson correlation coefficients (Rodgers & Nicewander, 1988), Fisher's discriminate ratio "*F* score" (Lin, Li, & Tsai, 2004), the Scatter criterion (Duda, Hart, & Stork, 2001), Single Variable Classifier SVC (Guyon & Elisseeff, 2003), Mutual Information (Battiti, 1994), the Relief Algorithm (Kira & Rendell, 1992; Liu & Motoda, 2008), Rough Set Theory (Liang, Wang, Dang, & Qian, 2014) and Data Envelopment Analysis (Zhang, Yang, Xiong, Wang, & Zhang, 2014).

The main advantages of the filter methods are their computational efficiency, scalability in terms of the dataset dimensionality, and independence from the classifier (Saeys, Inza, & Larranaga, 2007). A common drawback of these methods is the lack of information about the interaction between the features and the classifier and selection of redundant and irrelevant features due to the limitations of the employed goal functions leading to overestimation of the feature significance.

Information theory (Cover & Thomas, 2006) has been widely applied in filter methods, where information measures such as mutual information (MI) are used as a measure of the features' relevance and redundancy (Battiti, 1994). MI does not make an assumption of linearity between the variables, and can deal with categorical and numerical data with two or more class values (Meyer, Schretter, & Bontempi, 2008). There are several alternative measures in information theory that can be used to compute the relevance of features, namely mutual information, interaction information, conditional mutual information, and joint mutual information.

This paper contributes to the knowledge in the area of feature selection by proposing two new nonlinear feature selection methods based on information theory. The proposed methods aim to overcome the limitations of the current state of the art filter feature selection methods such as overestimation of the feature significance, which causes selection of redundant and irrelevant features. This is achieved through the introduction of a new goal function based on joint mutual information and the 'maximum of the minimum' nonlinear approach. As shown in the evaluation section, one of the proposed methods outperforms the competing feature selection methods in terms of classification accuracy, decreasing the average classification error by 0.88% in absolute terms and almost by 6% in relative terms in comparison to the next best performing method. In addition, it produces the best trade-off between accuracy and stability. The statistical significance of the reported results is further confirmed by ANOVA test.

This paper also reviews existing feature selection methods highlighting their common limitations and compares the performance of the proposed and existing methods on the basis of several criteria. For example, a nonlinear approach, which employs the 'maximum of the minimum' criterion, is compared to a linear approach, which employs cumulative summation approximation. To optimise the nonlinear approach, a goal function based on joint mutual information is compared to the goal function based on conditional mutual information. Finally, the effect of using normalised mutual information instead of mutual information is tested.

The rest of the paper is organised as follows. Section 2 presents the principles of the information theory, Section 3 reviews related work, Section 4 discusses the limitations of current feature selection criteria, Section 5 introduces the proposed methods. Section 6 describes the conducted experiments and discusses the results. Section 7 concludes the paper.

2. Information theory

This section introduces the principles of information theory by focusing on entropy and mutual information and explains the reasons for employing them in feature selection.

The entropy of a random variable is a measure of its uncertainty and a measure of the average amount of information required to describe the random variable (Cover & Thomas, 2006). The entropy of a discrete random variable $X = (x_1, x_2, \ldots\ldots, x_N)$ is denoted by $H(X)$, where x_i refers to the possible values that X can take. $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i)), \tag{1}$$

where $p(x_i)$ is the probability mass function. The value of $p(x_i)$, when X is discrete, is:

$$p(x_i) = \frac{number\ of\ instants\ with\ value\ x_i}{total\ number\ of\ instants\ (N)}. \tag{2}$$

The base of the logarithm, log, is 2, so $0 \le H(X) \le 1$. For any two discrete random variables X and $C = (c_1, c_2, \ldots\ldots, c_M)$, the joint entropy is defined as:

$$H(X, C) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log\big(p(x_i, c_j)\big) \tag{3}$$

where $p(x_i, c_j)$ is the joint probability mass function of the variables X and C. The conditional entropy of the variable X given C is defined as:

$$H(C|X) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log\big(p(c_j|x_i)\big) \tag{4}$$

The conditional entropy is the amount of uncertainty left in C when a variable X is introduced, so it is less than or equal to the entropy of both variables. The conditional entropy is equal to the entropy if, and only if, the two variables are independent. The relation between joint entropy and conditional entropy is:

$$H(X, C) = H(X) + H(C|X) \tag{5}$$

$$H(X, C) = H(C) + H(X|C) \tag{6}$$

Mutual Information (MI) is the amount of information that both variables share, and is defined as:

$$I(X; C) = H(C) - H(C|X) \tag{7}$$

MI can be expressed as the amount of information provided by variable X, which reduces the uncertainty of variable C. MI is zero if the random variables are statistically independent. MI is symmetric, so:

$$I(X; C) = I(C; X) \tag{8}$$

$$I(X; C) = H(X) - H(X|C) \tag{9}$$