Expert Systems With Applications 110 (2018) 11–19

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature selection by integrating two groups of feature evaluation criteria

Wanfu Gao[a], Liang Hu[a,*], Ping Zhang[b], Feng Wang[a]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] College of Software, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 14 September 2017
Revised 3 February 2018
Accepted 27 May 2018
Available online 28 May 2018

Keywords:
Feature selection
Information theory
Classification
Class-independent feature redundancy
Class-dependent feature redundancy

ABSTRACT

Feature selection is a preprocessing step in many application areas that are relevant to expert and intelligent systems, such as data mining and machine learning. Feature selection criteria that are based on information theory can be generally sorted into two categories. The criteria in the first group focus on minimizing feature redundancy, whereas those in the second group aim to maximize new classification information. However, both groups of feature evaluation criteria fail to balance the importance of feature redundancy and new classification information. Therefore, we propose a hybrid feature selection method named Minimal Redundancy-Maximal New Classification Information (MR-MNCI) that integrates the two groups of feature selection criteria. Moreover, according to the characteristics of the two groups of selection criteria, we adopt class-dependent feature redundancy and class-independent feature redundancy. To evaluate MR-MNCI, seven competing feature selection methods are compared with our method on 12 real-world data sets. Our method achieves the best classification performance in terms of average classification accuracy and highest classification accuracy.

1. Introduction

Feature selection plays a critical role in expert and intelligent systems. It not only improves the classification performance but also reduces the computational cost of data analysis (Che et al., 2017; Hernández-Pereira et al., 2016; Peng & Fan, 2017). Therefore, feature selection is increasingly important in classification problems. Feature selection retains relevant features while eliminating irrelevant and redundant features as much as possible. Traditional feature selection methods can be divided into three categories based on different selection strategies (Agnihotri, Verma, & Tripathi, 2017; Yousefpour, Ibrahim, & Hamed, 2017), including filter, wrapper and embedded methods. Filter methods evaluate a feature based on an evaluation function, and are independent of any classifier. Wrapper methods compute the score of a subset according to a specific classifier. Embedded methods estimate a feature subset in the learning stage. Compared with wrapper and embedded methods, filter methods are simple and fast (Hu, Gao, Zhao, Zhang, & Wang, 2018; Li et al., 2016). We focus on filter methods in this work.

Many filter methods have been proposed. These filter methods combine many measurement criteria, such as similarity-, statistical- and information-theory-based methods. Filters based on information theory receive the most attention because information theory can measure nonlinear correlations (Zhao, Zhou, Zhang, & Chen, 2016).

Mutual information is an important concept in information theory. It can evaluate the relevancy between two random variables X and Y (Cover & Thomas, 2012). Mutual information is defined as follows:

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

where $H(Y)$ represents the entropy, which is a measure of the uncertainty of a random variable, and $H(Y|X)$ represents the conditional entropy, which is the uncertainty of Y left when X is introduced.

Filter methods that are based on information theory can be roughly categorized into two groups according to the feature evaluation criteria: those that minimize feature redundancy and those that maximize new classification information (Wang, Wei, Yang, & Wang, 2017).

The feature selection methods in the first group aim at minimizing feature redundancy. We suppose that X_k is a candidate feature, S is the selected feature subset and Y is the class label. The feature selection criterion $J(.)$ of the first group is presented as fol-

* Corresponding author.
E-mail addresses: gaowf16@mails.jlu.edu.cn (W. Gao), 543786450@qq.com (L. Hu), zhangping15@mails.jlu.edu.cn (P. Zhang), wangfeng12@mails.jlu.edu.cn (F. Wang).

https://doi.org/10.1016/j.eswa.2018.05.029
0957-4174/

lows:

$$J(X_k) = I(X_k; Y) - R(S; X_k) \tag{2}$$

$R(S; X_k)$ represents the feature redundancy. In some methods (Battiti, 1994; Ding & Peng, 2005; Estévez, Tesmer, Perez, & Zurada, 2009), $R(S; X_k)$ is calculated by the cumulative summation of the mutual information between a candidate feature X_k and each selected feature X_j:

$$R(S; X_k) = \beta \sum_{X_j \in S} I(X_k; X_j) \tag{3}$$

β is a parameter, for which different feature selection methods provide different values. The indices k and j of X_k and X_j represent the kth and jth features.

Alternatively, many feature selection methods regard interaction information (i.e., $I(X_k; Y; X_j)$) as feature redundancy (Bontempi & Meyer, 2010; Cheng, Qin, Feng, Wang, & Li, 2011; El Akadi, El Ouardighi, & Aboutajdine, 2008). Interaction information is defined as follows:

$$I(X_k; Y; X_j) = I(X_k; Y) + I(X_j; Y) - I(Y; X_k, X_j) \tag{4}$$

where $I(Y; X_k, X_j)$ is the joint mutual information between Y and (X_k, X_j).

The feature selection methods in the second group maximize the new classification information (Fleuret, 2004; Meyer & Bontempi, 2006; Yang & Moody, 2000). The criterion can be generally defined as follows:

$$J(X_k) = \lambda I(X_k; Y|S) \tag{5}$$

where $I(X_k; Y|S)$ is the conditional mutual information, which quantifies the new classification information that is provided by candidate feature X_k when selected feature subset S is known, and λ is a parameter. In practice, Eq. (5) is expressed as follows:

$$J(X_k) = \lambda I(X_k; Y|X_j) \tag{6}$$

Based on these descriptions, we conclude that the first group of feature selection methods focuses on similarities between candidate features and selected features, whereas the second group of feature selection methods pays more attention on discrepancies between candidate features and selected features. In summary, the two groups of feature selection methods aim at minimizing feature redundancy and maximizing new classification information. Nevertheless, high new classification information does not imply low feature redundancy, and vice versa.

To address this issue, a hybrid feature selection method that integrates the two groups of feature evaluation criteria is proposed. The proposed method, which is named Minimal Redundancy - Maximal New Classification Information (MR-MNCI), employs feature redundancy and new classification information.

To evaluate the classification performance of MR-MNCI, MR-MNCI is compared with three traditional methods and four state-of-the-art methods on 12 real-world data sets. Our method achieves the best classification performance in terms of average classification accuracy and highest classification accuracy. Additionally, the Area Under Curves (AUCs) for all 12 benchmark data sets are represented by boxes. The boxes of MR-MNCI are higher than those of the other compared methods on two different classifiers.

The remainder of this work is organized as follows: Section 2 reviews related work. In Section 3, a hybrid feature selection method is proposed. Section 4 describes the experimental evaluation. Section 5 discusses the experimental results. Section 6 presents the conclusions of this work and our plan for future action.

2. Related work

Dimensionality reduction has been a hot topic over the last two decades. Many feature selection methods that are based on information theory have been proposed. Most belong to one of two groups: those that minimize feature redundancy and those that maximize new classification information.

Intuitively, the importance of a feature is measured by mutual information, which quantifies the relevancy between features and classes. Therefore, Mutual Information Maximization (MIM) based on this criterion (Lewis, 1992) is proposed. In MIM, each feature is viewed as an independent individual. However, this method suffers from the limitation of feature redundancy.

To avoid the effect of redundant information, Mutual Information Feature Selection (MIFS) (Battiti, 1994), which considers feature redundancy, is proposed. It is expressed as follows:

$$J(X_k) = I(X_k; Y) - \beta \sum_{X_j \in S} I(X_k; X_j) \tag{7}$$

β is a parameter of feature redundancy. MIFS belongs to the group of methods that minimize feature redundancy.

With more selected features, however, the effect of feature redundancy gradually decreases. Differing from MIFS, Minimum-Redundancy Maximum-Relevance (mRMR) (Peng, Long, & Ding, 2005) defines the value of β as the inverse of the number of selected features:

$$J(X_k) = I(X_k; Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_k; X_j) \tag{8}$$

mRMR and MIFS provide accurate calculation of the relevancy term. However, they always suffer from overestimation of the redundancy term (Che et al., 2017).

Conditional Mutual Information Maximization (CMIM) (Fleuret, 2004) maximizes conditional mutual information by employing the 'maximum of the minimum' criterion. The criterion belonging to the second group, definitely, seek to maximize the new classification information. F denotes the full feature set. CMIM is expressed as follows:

$$J(X_k) = argmax_{X_k \in F-S}(min_{X_j \in S}(I(X_k; Y|X_j))) \tag{9}$$

Similarly, Joint Mutual Information Maximization (JMIM) (Bennasar, Hicks, & Setchi, 2015) maximizes the joint mutual information. This approach belongs to the second group, which consists of the approaches that maximize the new classification information. However, JMIM has the problem of underestimation of feature significance (Che et al., 2017).

$$J(X_k) = argmax_{X_k \in F-S}(min_{X_j \in S}(I(X_k, X_j; Y))) \tag{10}$$

Although the joint mutual information is the mutual information between (X_k, X_j) and the class label Y, the joint mutual information can be rewritten as follows:

$$I(X_k, X_j; Y) = I(X_k; Y|X_j) + I(X_j; Y) \tag{11}$$

Note that $I(X_j; Y)$ is regarded as constant in the feature selection process. Thus, JMIM is a variation of CMIM.

Different from the two groups of feature selection methods that are mentioned above, several feature selection methods that consider feature interdependency have been proposed.

Dynamic Weighting-Based Feature Selection (DWFS) (Sun, Liu, Xu, Chen, Han & Wang, 2013) not only selects the most relevant features and eliminates redundant features but also tries to retain interdependent features.

Zeng et al. proposed Interaction Weight based Feature Selection (IWFS) (Zeng, Zhang, Zhang, & Yin, 2015) in 2015. First, IWFS refines feature relevancy, feature redundancy and feature interaction.