Expert Systems With Applications 108 (2018) 81–95

Contents lists available at ScienceDirect

Expert Systems With Applications

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

High-dimensional supervised feature selection via optimized kernel mutual information

Ning Bi[a], Jun Tan[a,*], Jian-Huang Lai[b], Ching Y. Suen[c]

[a] School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China
[b] School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China
[c] Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, QC H3G 1M8, Canada

ARTICLE INFO

Article history:
Received 16 July 2017
Revised 29 January 2018
Accepted 27 April 2018
Available online 2 May 2018

Keywords:
Feature selection
Kernel method
Mutual information
Classification
Optimize function
Machine learning

ABSTRACT

Feature selection is very important for pattern recognition to reduce the dimensions of data and to improve the efficiency of learning algorithms. Recent research on new approaches has focused mostly on improving accuracy and reducing computing time. This paper presents a flexible feature-selection method based on an optimized kernel mutual information (OKMI) approach. Mutual information (MI) has been applied successfully in decision trees to rank variables; its aim is to connect class labels with the distribution of experimental data. The use of MI removes irrelevant features and decreases redundant features. However, MI is usually less robust when the data distribution is not centralized. To overcome this problem, we propose to use the OKMI approach, which combines MI and a kernel function. This approach may be used for feature selection with nonlinear models by defining kernels for feature vectors and class-label vectors. By optimizing the objection equations, we develop a new feature-selection algorithm that combines both MI and kernel learning, we discuss the relationship among various kernel-selection methods. Experiments were conducted to compare the new technique applied to various data sets with other methods, and in each case the OKMI approach performs better than the other methods in terms of feature-classification accuracy and computing time. OKMI method solves the problem of computation complexity in the probability of distribution, and avoids this problem by finding the optimal features at very low computational cost. As a result, the OKMI method with the proposed algorithm is effective and robust over a wide range of real applications on expert systems.

1. Introduction

Feature selection is one of the important issues in expert and intelligent system technology, which uses both the input and output variables to predict the relationships between the features and class labels. Prediction models have been used for various expert and intelligent systems applications including multi-agent systems, knowledge management, neural networks, knowledge discovery, data and text mining, multimedia mining, and genetic algorithms. The models generally involve a number of features. However, not all of these features are equally important for a specific task. Some of them may be redundant or even irrelevant. Better performance may be achieved by discarding some features.

Many applications in pattern recognition require the identification of the most characteristic features of a given data set D that contains N samples and M features, with $X = \{x_i, i = 1, \ldots, M\}$. The data set D usually includes a large amount of irrelevant, redundant, and unnecessary information, which degrades recognition performance. The feature-selection method (Cheriet, Kharma, Liu, & Suen, 2007) selects a subset G of features from M ($|G| < M$) such that D is optimized based on G according to criterion J. The goal is to maximize the predictive accuracy of the data within D and minimize the cost of extracting features within G.

Feature selection focuses on finding the best subspace, where the total number of subspaces in the original data set D is 2^M. Given the number k ($k < M$), the number of subspaces with dimension less than k is $\sum_{i=1}^{k} \binom{M}{i}$. Thus, D is high dimensional for a large feature number M, so thoroughly searching the subspace of features is difficult. To address this issue, sequential-search-based methods to select features have been proposed. Blum and Langley (1997) grouped feature-selection methods into three types: filter, wrapper, and embedded. Filter methods (Almuallim & Dietterich, 1994; Kira & Rendell, 1992) provide quick estimates of the value of features and filter the irrelevant or redundant features be-

* Corresponding author.
E-mail address: mcstj@mail.sysu.edu.cn (J. Tan).
URL: http://www.math.sysu.edu.cn (J. Tan)

https://doi.org/10.1016/j.eswa.2018.04.037
0957-4174/

fore they are fed into the classifier. In contrast, wrapper methods (Kohavi & John, 1997) usually interact with a classifier, so the classifier performance will directly affect the quality of the feature subsets. Finally, in embedded methods (Lal, Chapelle, Weston, & Elisseeff, 2006), feature selection is embedded into the classifier, so the two are not independent and execute simultaneously.

Additionally, feature-selection methods can be categorized as unsupervised, supervised, or semi-supervised. Unsupervised methods were developed without using class labels and include joint embedding learning and sparse regression (JELSR) (Hou, Nie, Li, Yi, & Wu, 2014), matrix factorization (MF) (Wang, Pedrycz, Zhu, & Zhu, 2015), k-nearest-neighbor (Chan & Kim, 2015), feature similarity feature selection (FSFS) (Mitra, Murthy, & Pal, 2002), Laplacian score (LS) (He, Cai, & Niyogi, 2005), and regularized self-representation (Zhu, Zuo, Zhang, Hu, & Shiu, 2015), all of which offer many efficient algorithms for unsupervised feature selection. Supervised feature-selection methods search for features of an input vector by predicting the class label, the existing methods include ReliefF (Kira & Rendell, 1992), Fisher score, correlation, kernel optimize (Kopt) (Xiong, Swamy, & Ahmad, 2005), kernel class separability (KCS) (Wang, 2008), generalized multiple kernel learning (GMKL) (Varma & Babu, 2009), scaled class separability selection(SCSS) (Ramona, Richard, & David, 2012), spectral feature with minimum redundancy (MRSF) (Zhao, Wang, & Liu, 2010), Hilbert-Schmidt independence criterion (HSIC) (Gretton, Bousquet, Smola, & Lkopf, 2005), the HSIC-based greedy feature selection criterion (Song, Smola, Gretton, Bedo, & Borgwardt, 2012), the sparse additive models (SpAM) (Ravikumar, Lafferty, Liu, & Wasserman, 2009), Hilbert–Schmidt feature selection (HSFS) (Masaeli, Fung, & Dy, 2010), and centered kernel target alignment (cKTA) (Cortes, Mohri, & Rostamizadeh, 2014), feature-wise kernelized Lasso (HSIC Lasso) (Yamada, Jitkrittum, Sigal, Xing, & Sugiyama, 2014). The method proposed herein and the methods we compare it with are all supervised methods.

The popular MI method (Eriksson, Kim, Kang, & Lee, 2005) constructs decision trees to rank variables and also serves as a metric for feature selection. An MI method based on Shannon information uses information-theoretic ranking with the dependency between two variables serving as metric and uses entropy to represent relationships between an observed variable x and an output result y. The MI of x and y is defined by their probabilistic density functions $p(x)$ and $p(y)$, respectively, and their joint probabilistic density function $p(x, y)$ (Battiti, 1994). To rank features, some works report that the union of individually good features does not necessarily lead to good recognition performance; in other words "the m best features are not the best m features." Although MI can decrease the redundancy with respect to the original features and select the best features with minimal redundancy, the joint probability of features and the target class increases, so the redundancy among features may decrease (Pudil, Novovičová, & Kittler, 1994).

To select good features by using the statistical dependency distribution, Peng, Long, and Ding (2005) proposed the minimal-redundancy-maximal-relevance criterion (mRMR), which uses a feature-selection method based on MI. The method provides maximal dependency, maximal relevance, and minimal redundancy. The selected features have the maximal joint dependency on the target class, which is called "maximal dependency," but it is hard to implement, so the relevant approximate dependency uses the MI between feature and target class. Minimal redundancy reduces the redundancy resulting from maximum relevance, so the redundancy metric is computed from the MI among the selected features. Experiments with mRMR improved the classification accuracy for some data sets. For more details about research into mRMR, see Refs. Ding and Peng (2005) and Zhao et al. (2010).

The definition of MI is based on the feature entropy and class label; however, it favors features with many values. Some features can be very simple, so the feature value is an integer with a small range. However, some feature values are floating points with very wide ranges, which needs more computation to obtain a ratio that reflects the correlation between features and class. Another problem is inconsistency (Dash & Liu, 2003): consider n samples with the same range of feature-value but m_1 of these samples belong to class 1 and the remaining m_i samples belong to class i. The largest feature is m_1, so the inconsistency is $n - m_1$ and the inconsistency rate is the sum of the inconsistencies divided by the size N of the set. Reference Dash and Liu (2003) shows that the time complexity of computing the inconsistency rate is close to $O(N)$, so the rate is also monotonic and is tolerant to noise; however, it is only available for discrete values. Thus, the rate must be discretized for continuous features, which will seriously affect the computation complexity and consume more memory resources. Occasionally, the computation is interrupted when the feature number is too large for the memory. This problem is discussed in detail below.

To resolve these drawbacks of the MI method, the kernel-based methods (Gretton, Herbrich, & Smola, 2003; Lin, Ying, Chen, & Lee, 2008; Sakai & Sugiyama, 2014) are imported to enhance the MI, the Hilbert–Schmidt independence criterion (HSIC) using kernel-based independence measures is introduced in Refs. Gretton et al. (2005) and Song et al. (2012). These approaches are popular for mapping the data to a nonlinear high-dimensional space (Alzate & Suykens, 2012; Schölkopf, Smola, & Müller, 1998). Multi-kernel learning (Wang, Bensmail, & Gao, 2014) has been applied to feature selection with a sparse representation on the manifold can handle noise features and nonlinear data. The kernel-based feature-selection method integrates a linear combination of features with the criterion. Real applications of the kernel concern the type of kernel and parameters, so while cross validation may optimize the kernel, it consumes more time and is easy to over-fit.

Traditional feature selection methods (Kira & Rendell, 1992) based on the assumption of linear dependency between input features and output values, they cannot capture non-linear dependency. KCS (Wang, 2008) cKTA (Cortes et al., 2014) are not necessarily positive definite, and thus the objective functions can be non-convex. Furthermore, for the kernel-based methods (Gretton et al., 2003; Varma & Babu, 2009; Xiong et al., 2005), output y should be transformed by the non-linear kernel function $\phi()$, this highly limits the flexibility of capturing non-linear dependency, an advantage of the formulation is that the global optimal solution can be computed efficiently. Thus, it is scalable to high-dimensional feature selection problems. Finally, an output y should be a real number in SpAM (Ravikumar et al., 2009), meaning that SpAM cannot deal with structured outputs such as multi-label and graph data. Greedy search strategies such as forward selection/backward elimination are used in mRMR (Peng et al., 2005) HSIC (Gretton et al., 2005). However, the greedy approaches tend to produce a locally optimal feature set. To the best of our knowledge, the convex feature selection method is able to deal with high-dimensional non-linearly related features. In addition, the output Gram matrix L is used to select features in HSIC Lasso (Yamada et al., 2014), which can naturally incorporate structured outputs via kernels. All feature methods are summarized in Table 1.

To address this problem, we propose herein an approach that combines the goodness of the kernel function and the MI method to obtain a high-dimensional supervised feature-selection framework called optimized kernel mutual information (OKMI) with joint kernel learning, maximum relevance, and minimum redundancy. Instead of using MI to characterize high-dimensional data by the feature and class probability, we embed a kernel function into the MI to form a new framework. Widely used types of kernel functions, including polynomial, Gaussian, exponential, and sigmoid, can be seen as special cases in the OKMI framework. Af-