Pattern Recognition Letters 92 (2017) 1–8

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A new feature selection method based on a validity index of feature subset

Chuan Liu [a,*], Wenyong Wang [a], Qiang Zhao [a], Xiaoming Shen [b], Martin Konan [a]

[a] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[b] *Zhejiang Electric Power Research Institute, Hangzhou 310014, China*

ARTICLE INFO

Article history:
Received 26 May 2016
Available online 22 March 2017

Keywords:
Feature selection
Wrapper methods
Filter methods

ABSTRACT

The wrapper feature selection method can achieve high classification accuracy. However, the cross-validation scheme of the wrapper method in evaluation phase is very expensive regarding computing resource consumption. In this paper, we propose a new statistical measure named as LW-index which could replace the expensive cross-validation scheme to evaluate the feature subset. Then, a new feature selection method, which is the combination of the proposed LW-index with Sequence Forward Search algorithm (SFS-LW), is presented in this paper. Further, we show through plenty of experiments conducted on nine UCI datasets that the proposed method can obtain similar classification accuracy as the wrapper method with centroid-based classifier or support vector machine, and its computation cost is approximate to the compared filter methods.

1. Introduction

The development of information technology and the various expanded applications available on the Internet have exponentially increased the demands of services such as biological analysis, financial data mining and online information processing. Hence, a dimensionality issue for the learning algorithm arose with the growth of features. In classification problems, the significance and importance of features are different according to the given classification method and criteria. That is, the key feature has a strong distinguishing importance and high correlation with the category label [26]. In contrast, the redundancy features not only affect the performance of classification algorithm but also require an additional computational cost. Therefore, it is significant to eliminate irrelevant and redundant features by feature selection algorithm that selects the best subset of features from the original feature domain. Thus, feature selection is one of the most important issues in machine learning and pattern recognition research [19], which greatly reduces the computational cost, avoids overfitting and improves the generalization ability [31,39].

Feature selection has begun to draw intensive attention among researchers since it brings a lot of benefits for data analysis and data understanding. Hence, several methods have been proposed, which can be generally divided into approaches [6] that are classifier-independent ('filter' methods), and classifier-dependent ('wrapper' and 'embedded' methods).

Filter [12,26,30] methods attempt to assess the importance of features statically according to a heuristic scoring criteria without any particular classifier [30]. Thus, the features with the high score are selected and applied to a classification algorithm. Generally, filter methods have high computational efficiency by reducing the size of the feature subset quickly (fast speed).

Wrapper [11,12,45] methods search the space of feature, using a particular classifier as the measure of importance (significance) for a candidate feature subset. Firstly, the wrapper produces a candidate feature subset by the search strategy, and then the classifier is trained and tested to evaluate the candidate feature subset. This process will be iteratively performed until the selected feature subset meets the specific requirements [44]. Though wrapper methods may guarantee good results, they have the disadvantage to be computationally expensive and become more unfeasible (computationally) as the number of features increases. Additionally, they may produce feature subsets that are overly specific to the used classifier and are easy to overfitting. Embedded methods [4,29] exploit the structure of specific classes of learning classifiers to guide the feature selection process, and the defining component is a criterion derived from the fundamental knowledge of a specific class of regression or classification function [6].

* Corresponding author.
E-mail address: liuchuan@uestc.edu.cn (C. Liu).

http://dx.doi.org/10.1016/j.patrec.2017.03.018

In summary, each family of these above feature selection methods has its own advantages and disadvantages [6,20]. In general, in terms of accuracy, wrapper methods have high learning capacity. Hence, they usually obtain higher accuracy than embedded methods, which in turn are better than filter methods. However, in terms of speed, filters are the fastest among all the methods as they need not incorporate learning, while wrappers are the slowest since they typically need to evaluate the Cross-Validation (CV) [25] procedure at each iteration step [23,42]. Thus, for some high-dimensional applications such as text classification and gene analysis [22], wrapper methods could be computationally unfeasible since CV evaluation is characterized by large resource consumption [18,27,36].

In CV scheme, each feature subset obtained from wrapper methods in supervised classification problem can be evaluated using external clustering indices since we train and test the classification model, and the obtained data partition by the classification model can be compared with the original data partition provided by the class label. Alternatively, the labeled feature subset can be also seen as a special data partition returned by a clustering algorithm. Hence, we can evaluate the feature subset with an internal index [14,37,41]. In this case, we use the class label as the obtained data partition if the data partition given by the class label represents well-separated groups. Therefore, we do not need to train and test the classifier, which is time-consuming in general. If the internal index computation is more efficient than the classifier training and testing along with obtained partition evaluation process and, additionally, if the internal index is highly correlated with external indices such as F_1 measure [28], then the proposed methodology can be a good alternative to evaluate the candidate feature subsets in wrapper methods.

In this paper, we propose a new internal index, named as LW-index (LWI), in replacement of CV evaluation. LW-index is a linear approach for evaluating the feature subset in the supervised classification problem, and it will get a high value if the partition is compact and separated; which indicates respectively the variation or scattering of the data within a cluster and the isolation of the clusters from each other. Then, a new filter feature selection algorithm, i.e., Sequence Forward Search based on LW (SFS-LW), is proposed by combining LW-index with sequence search strategy in wrapper methods. The experimental results indicate that the proposed method guarantees good classification accuracy and greatly reduces the computation cost compared to the wrapper methods.

The rest of the paper is structured as follows: Section 2 reviews previous work about feature selection methods. Section 3 describes the details about LW-index. Section 4 presents the proposed filter algorithm. Section 5 demonstrates the proposed methodology with a series of experiments. Finally, the conclusions of this study are given in Section 6.

2. Related work

Filter methods — filter methods separate the classification process and feature selection components, and define a heuristic criterion, also referred to as a 'relevance index' or 'scoring', to evaluate statistics of the data independently of any particular classifier, thereby extracting features that are generic without incorporating particular assumptions. There are many hand-designed heuristic filter criteria, such as Information Gain (IG) [11], Mutual Information (MI) [29], Chi-Square (CS) [9], Cross Entropy (CE) [34], have been suggested in text categorization research. The defining component of these filter methods is based on the concept of information entropy in probability theory, quantifying the 'utility' of a particular feature in the set. Thus, the intuition behind these criterions is that a stronger correlation between the feature and the class label should imply a greater predictive ability when using the feature. To use these criterions, the features can be effectively ranked in descending order of their individual score, and then a certain number of top features are selected, where the number is decided by some other predefined stopping criterion. However, these filter methods focus on the utility of individual feature only and ignore the combination of features. Therefore, this is known to be sub-optimal in which features may be interdependent. Moreover, the optimal size of feature subset is hard to be determined.

It is widely accepted that a useful and parsimonious subset of features should not only be individually independent but also should not be redundant regarding each other—features should not be highly correlated [6]. In order to liberalize the limitation in the above methods that each feature is independent of other features, several criteria that attempt to pursue this relevancy-redundancy goal have been proposed. For instance, Battiti [3] proposed the Mutual Information Feature Selection (MIFS) criterion, which includes a term to ensure feature relevance, and introduces a penalty to enforce low correlations with features already selected. After MIFS proposed, an alternative approach proposed by Yang focuses on increasing complementary information between features with using the Joint Mutual Information (JMI) [46]. The key idea behind JMI is that the candidate feature is useful if it is complementary with existing features. Similarly, many criteria, such as Conditional Mutual Infomation Maximisation (CMIM) [17], Max-Relevance Min-Redundancy (MRMR) [38], Interaction Capping (ICAP) [5], and Double Input Symmetrical Relevance (DISR) [33], that attempted to manage the relevance-redundancy tradeoff with various heuristic terms by considering the features previously selected have been proposed in the filter field.

Wrapper methods — wrapper feature selection process includes three components, i.e., search strategy, evaluation function and performance function [1]. Firstly, search strategies are used to search through the space of features, and they can be categorized into three groups [15]: exhaustive, sequential and randomized. Though the exhaustive search could find the optimal solution, it is known to be an NP-hard problem [2]. Even there is a kind of improvement such as the branch-and-bound (Branch and Bound) method, but it still leads to huge computational cost [19]. The sequential search such as Sequence Forward Search (SFS) and Sequential Backward Search (SBS) [10,19] traverse the feature space in one direction, while the randomized search such as the Genetic Algorithm (GA) and Ant Colony Optimization algorithm (ACO) randomly generate the subset of features. Secondly, the evaluation function is used to assess the merits of the candidate feature subset. Finally, the performance function is applied to validate the selected feature subset.

To improve the efficiency of the wrapper methods, a large number of improved statistical methods and machine learning techniques have been applied to the wrapper approach as induction algorithm which include K-Nearest Neighbor (KNN) [29], Naive Bayes [16], Decision Tree [8], Neural Network [21], Support Vector Machines (SVM) [11,31,32] and etc. Though these methods may guarantee good results, they may produce feature subsets that are overly specific to the used classifier.

Meanwhile, some researchers focus on the improvement of search strategies. For instance, they use mathematical equations to imitate natural phenomena including the biological evolutionary process like Genetic Algorithm (GA) [13], animal behavior like Bat Algorithm (BA) [40] and Ant Colony Optimization (ACO) [43]. These random search strategies have been proved to be successful in some fields. However, the disadvantage of these above mentioned random strategies is that their performance is not stable. Therefore, the sequential and floating sequence search strategies have been introduced into wrapper methods to shorten the process of producing candidate feature subset. Thus, Maldonado [32] presented a sequential backward strategy to save the feature space traversing time, and the number of errors is regarded as the merit to the subset validation.