Neurocomputing 151 (2015) 376–389

Contents lists available at ScienceDirect

Neurocomputing

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

Effective feature selection using feature vector graph for classification

Guodong Zhao [a,b], Yan Wu [a,*], Fuqiang Chen [a], Junming Zhang [a], Jing Bai [a]

[a] School of Electronics and Information, Tongji University, Shanghai 201804, China
[b] School of Mathematics and Physics, Shanghai Dian Ji University, Shanghai 201306, China

ARTICLE INFO

Article history:
Received 22 March 2014
Received in revised form
4 July 2014
Accepted 15 September 2014
Communicated by Haowei Liu
Available online 30 September 2014

Keywords:
Feature selection
Community modularity
Relevant independency
Feature vector graph
Classification accuracy

ABSTRACT

Optimal feature subset selection is often required as a preliminary work in machine learning and data mining. The choice of feature subset determines the classification accuracy. It is a crucial aspect to construct efficient feature selection algorithm. Here, by constructing the feature vector graph, a new feature evaluation criterion based on community modularity in complex network is proposed to select the most informative features. To eliminate the relevant redundancy among features, conditional mutual information-based criterion is used to capture information about relevant independency between features, which is the amount of information they can predict about label variable, but they do not share. The most informative features with maximum relevant independency are added to the optimal subset. Integrating these two points, a method named the community modularity Q value-based feature selection (*CMQFS*) is put forward in this paper. Furthermore, our method based on community modularity can be certified by k-means cluster theory. We compared the proposed algorithm with other state-of-the-art methods by several experiments to indicate that *CMQFS* is more efficient and accurate.

1. Introduction

The rapid development of information technology makes it easy to accumulate the data sets with high dimensionality. Nevertheless, most of the features in huge dataset are irrelevant or redundant, which typically deteriorates the performance of machine learning algorithms. An effective way to mitigate the problem is to reduce the dimensionality of feature space with efficient feature selection technique. Feature selection is to identify a subset from the original feature set of data, which will improve the quality of the data. Therefore, feature selection becomes more and more important in machine learning and data mining.

According to the way they combine the optimal feature subset search with the construction of learning models, feature selection methods can be roughly divided into three types, *i.e.*, embedded, wrapper and filter methods. Embedded and wrapper methods [1,2] are classifier-dependent, which evaluate the features using a learning algorithm. They outperform filter methods in terms of accuracy; but they suffer from poor generalization ability for other classifiers and high computational complexity in the learning process for high dimensional datasets, because they are tightly coupled with specified learning algorithms [25]. On the contrary, the filter methods, which are not specified to a learning algorithm, evaluate the discrimination capability of each feature by investigating only the intrinsic properties of the data. Hence they are not coupled with any learning algorithm. Compared to the above two methods, the filter methods [3–5] are more commonly adopted for feature selection because of their simplicity, computational efficiency and scalability for very high-dimensional datasets. The filter methods have attracted great attention and a large number of filter algorithms have been developed in the past decades. In this paper, we focus on the filter methods.

The filter methods are generally subdivided into two classes: *ranking* and *subset selection*. In the first class, the *ranking* methods evaluate the significance of discrimination for each feature based on different evaluation criterion. A weight (or score) is firstly calculated for each feature depending on a specified weighting function, and the features with top weights (or scores) are picked into optimal subset while the rest are discarded. In the second class, the solution of feature optimal subset with the highest accuracy is **NP** hard [51]. In order to avoid the combinatorial search problem to find an optimal subset, the most popular variable selection methods mainly include forward, backward and floating sequential schemes, which always use heuristic approaches to provide a sub-optimal solution. In this study, a new scoring criterion based on community modularity in complex network has been developed, which could well identify the

* Corresponding author.
E-mail addresses: zgd215@163.com (G. Zhao), yanwu@tongji.edu.cn (Y. Wu).

http://dx.doi.org/10.1016/j.neucom.2014.09.027
0925-2312/

discriminative information of each feature. The analysis on relevant independency between features can effectively deal with the relevant redundancy among selected features. Hence, through the proposed method, not only the most relevant features are selected and redundant features are eliminated, but also useful intrinsic feature groups are retained.

2. Related work

As mentioned previously, a large number of filter-based feature selection (FS) algorithms have been presented in the past few decades for mining the optimal features subset from the high-dimensional feature spaces. According to the way of utilizing label information in the feature selection process, FS methods can be divided into three classes: unsupervised FS, semi-supervised FS and supervised FS.

The feature *ranking* based selection methods have been presented to calculate the features' scoring based on constructing the different scoring functions. The *Variance Score* method [6] is to select the features with maximum variances by calculating the variance of each feature to reflect its representative power. The *Laplacian Score* method [6] is another popular unsupervised FS under the assumption that two close samples should have similar feature values and a good feature should have similar values for samples in the same class and large margin values for samples in different classes. Using normalized mutual information to measure the dependence between a pair of features, Yazhou Ren [7] proposed an improved Laplacian Score-based feature selection method based on local and global structure preserving. As a supervised FS, *Fisher Score* chose features with the best discriminate ability [8–10]. A number of supervised learning algorithms have been used to implement the filter methods, which include Relief family [11–15] and Fuzzy-Margin-based Relief (FM Relief) [16]. Battiti [17] investigated the application of mutual information criterion to evaluate candidate features and to select the top ranked features used as input data for a neural network classifier. These FS algorithms based on score function are widely used in data mining and pattern recognition. However, these methods have been criticized for their ignoring the redundancy among features, which may lead to selection over many redundant features and bring a bad influence on the performance of the following classifiers. To overcome the above problem, the optimal feature subset based methods have been affected considering the redundancy among the selected features by most researchers. Mutual Information (MI) is a measure of the amount of information between two random variables, which is symmetric and non-negative, and is zero if and only if the variables are independent. Then, the methods based on Mutual Information (MI) have been popular lately. Yu and Liu [18] introduced a novel framework that decoupled relevance analysis and redundancy analysis. They proposed a correlation-based subset selection method named FCBF for relevance and redundancy analysis, and then removed redundant features by approximate Markov Blanket technique. The *MIFS* algorithm [17] was proposed to calculate the mutual information both with respect to class variables and the already selected features for each feature and selected those features that have maximum mutual information with class labels but less redundant among the selected features. However, the MIFS algorithm ignored feature synergy, MIFS and its variants may cause a big bias when features are combined to cooperate together. To avoid the draw-back, Gang Wang [45] proposed a novel feature selection method for text categorization called conditional mutual information maximin (CMIM) where the triplet form is used to estimate conditional mutual information (CMI), aiming at greatly relieving the computation overhead and a set of individually discriminating and weakly dependent features can be selected. Based on information gain and MI, FESLP was proposed by Ye Xu [46] to address the link prediction problem, whose superior advantage is that those features with the greatest discriminative power are selected and simultaneously the correlations among features such that redundancy in the learned feature space are minimized as small as possible. The CMIF method has been proposed by Hongrong Cheng [19] based on the link between interaction information and conditional mutual information, which not only takes account of both redundancy and synergy interactions of features and identifies discriminative features, but also combines feature redundancy evaluation with classification tasks. Skwak and Choi [20] improved the *MIFS* method under the assumption of uniform distributions of information of input features, and put forward an algorithm called MIFS-U. Both MIFS and MIFS-U involved a redundancy parameter β, which is used to interpret the redundancy among input features. If $\beta=0$, the MI among input features is not taken into consideration and is deteriorated into the FCBF method. However, if β is chosen too large, the algorithms simply included the relation among the input features, and does not include the relation between the individual input features and the class [21]. Peng proposed the 'minimal redundancy maximal relevance' (*mrmr*) method [22], a special case of the *MIFS* algorithm when $\beta=1/|S|$, where $|S|$ is the number of feature in S. Then, the 'normalized mutual information feature selection' (*NMIFS*) algorithm [23] used the normalized MI by the minimum entropy of both features and the average normalized MI as a measure of redundancy of the individual feature and the subset of selected feature, where authors claimed that the *NMIFS* algorithm was an enhancement over the *MIFS*, *MIFS-U* and *mrmr* algorithms. Based on metric applied on continuous and discrete data representations, Jose´ Martı´nez Sotoca [24] built a dissimilarity space using information theoretic measure, in particular conditional mutual information between features with respect to a relevant variable that represents the class labels. Applying a hierarchical clustering, the algorithm searched for a compression of the information contained in the original set of features. Sun [25] presented a new scheme for feature relevance, interdependence and redundancy analysis using information theoretic criteria, whose primary characteristic was that the feature was weighted according to its interaction with the selected features. And the weight of features will be dynamically updated after each candidate feature has been selected. Gavin Brown [26] has pointed out that common heuristics for information based feature selection (including Markov Blanket algorithms as a special case) are approximate iterative maximizations of the conditional likelihood and presented a unifying framework for information theoretic feature selection, bringing almost two decades of researches on heuristic filter criteria under a single theoretical interpretation. From a new view point of finding the unique information, Okan Sakar [27] has proposed a method called Kernel Canonical Correlation Analysis based minimum Redundancy-Maximum Relevance (*KCCAmrmr*) which explored and used all the correlated functions (covariants) between variables to compute their unique (conditional) information about the target. However, most traditional information-theoretic based selectors will ignore some features which have strong discriminatory power as a group but are weak as individuals. To cope with this problem, Sun [28,29] introduced a cooperative game theory based frame work to evaluate the *power* of each feature, which can be served as a metric of the importance of each feature according to the intricate and intrinsic interrelation among features and provided the weighted features to feature selection algorithm, which was more stable for complex dataset, but had issue with high runtime complexity. More feature selection methods in the detailed description can be found in [30–32].

In the past few decades, a lot of complex network theories have also become extremely useful as the representation of a wide variety of systems in different areas, such as biological, social, technological, and information networks. Network analysis has become crucial to understand the features of these systems. In this paper, a new feature evaluation criterion based on community modularity in network analysis is proposed to evaluate discriminatory power of each