Pattern Recognition 45 (2012) 2992–3002

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Feature evaluation and selection with cooperative game theory

Xin Sun [a,b], Yanheng Liu [a,b,]*, Jin Li [c], Jianqi Zhu [a,b], Huiling Chen [a,b], Xuejie Liu [a,b]

[a] College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China
[c] School of Philosophy and Society, Jilin University, Changchun, Jilin 130012, China

ARTICLE INFO

Article history:
Received 3 June 2011
Received in revised form
12 January 2012
Accepted 2 February 2012
Available online 10 February 2012

Keywords:
Machine learning
Feature selection
Cooperative game theory
Filter method

ABSTRACT

Recent years, various information theoretic based measurements have been proposed to remove redundant features from high-dimensional data set as many as possible. 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, this paper introduces a cooperative game theory based framework to evaluate the *power* of each feature. The *power* can be served as a metric of the importance of each feature according to the intricate and intrinsic interrelation among features. Then a general filter feature selection scheme is presented based on the introduced framework to handle the feature selection problem. To verify the effectiveness of our method, experimental comparisons with several other existing feature selection methods on fifteen UCI data sets are carried out using four typical classifiers. The results show that the proposed algorithm achieves better results than other methods in most cases.

1. Introduction

Feature selection, also known as variable selection, is one of the fundamental problems in the fields of machine learning, pattern recognition and statistics. With the new emergences in computer applications, such as social networks clustering, gene expression array analysis and combinatorial chemistry, datasets with tens or hundreds of thousands of features are available. Nevertheless, most of the features in huge dataset are irrelevant or redundant, which lead learning algorithms to low efficiency and over-fitting. Thus, feature selection becomes one of the most active research areas to address this problem. The essential idea of feature selection is to eliminate the irrelevant and redundant features from data set as many as possible. Feature selection in machine learning has been well studied, aiming at finding a good feature subset which produces higher classification accuracy [1]. Also, it is helpful to acquire a better understanding of relationships among the features. Recently several researches have combined feature selection and classification together in various application area to improve the performance of machine learning, e.g., video semantic detection [2], text categorization [3], bioinformatics [4, 5] and intrusion detection [6].

Up to present, several different approaches are employed in feature selection, such as genetic algorithm [7], simulated annealing [8], SVM [9] and boosting method [10]. Furthermore, all of these feature selection methods typically fall into three categories: embedded, wrapper and filter methods. Embedded methods are embedded in and specific to a given machine learning algorithm, and select the features through the process of generating the classifier. Wrappers, evaluating each subset by specified learning algorithms which were treated as a black box, can choose optimal features to yield high prediction performance. One drawback of the wrapper methods, however, is their less generalization of the selected features on other classifiers and high computational complexity in learning, because they are tightly coupled with specified learning algorithms. What's more, they may have a risk of over fitting to the algorithm. Consequently, wrapper methods can hardly deal with large scale problems.

Filter methods are independent of learning algorithms. Instead, they rely on statistical tests over the original features of the training data. In practice, the filter methods have much lower computational complexity than the wrappers, meanwhile, they achieve comparable classification accuracy for most classifiers. Thus, the filter methods are very popular to high-dimension data set. To date, a modest number of efficient filter selection algorithms have been proposed in the literature. It is noteworthy that among various evaluation criterions, Information-theoretic based measurements achieve excellent performance and have drawn more and more attention. This is due to that such measurements capture both linear and nonlinear dependencies without requiring a theoretical probability

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China, Tel.: +86 0431 85159419; fax: +86 0431 85168337.
E-mail addresses: sunxin1984@yahoo.com.cn (X. Sun), lyh_lb_lk@yahoo.com.cn (Y. Liu).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.001

distribution or specific model of dependency. However, most of these selectors discard features which are highly correlated to the selected ones although relevant to the target class, which is likely to ignore features which as a group have strong discriminatory power but are weak as individuals [11]. To untie this knot, this paper introduces a cooperative game theory based framework to evaluate the power of each feature. Then a general filter feature selection scheme is presented based on the introduced framework to handle the feature selection problem using any Information-theoretic criteria.

The rest of this paper is structured as follows: In Section 2, related works are briefly reviewed. Section 3 introduces some basic concepts of information theory in feature selection and the necessary background of cooperative game theory. Section 4 provides a cooperative game theory-based framework to evaluate the *power* of each feature, and presents a general filter feature selection algorithm. Section 5 gives experimental results on UCI data sets to evaluate the effectiveness of our approach and some discussions. Conclusions and future work are presented in Section 6.

2. Related work

So far, researchers have proposed lots of selection algorithms to find the optimal feature subset from high-dimension features space [12]. Wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain [13], e.g., Kabir et al. [14] proposed a wrapper method using neural networks, Inza et al. [15] presented a wrapper method by Estimation of Bayesian Network Algorithm. Embedded methods have better computational complexity than wrapper methods [4]. Guyon et al. [16] propose a embedded method(SVM-RFE) utilizing Support Vector Machine methods based on Recursive Feature Elimination. It has been successfully applied in the area of gene expression analysis. For more detailed reviews on embedded and wrapper methods, readers can refer to the previous literature [11, 13, 17–20] for more information. Since our proposed selection method is independent of any learning algorithms, we focus our attention only on filters. In the following, the state-of-the-art filter methods are briefly reviewed.

For feature selection, one of the most critical challenges is to measure the goodness of a feature subset in determining an optimal one [20]. Unlike wrappers, filters do not employ a learning algorithm to evaluate the selected attribute subsets. Instead, they evaluate the significance of features according to some measurements, such as distance [21–23], rough set theory [24], χ^2 [25], information theory [26] and others. Among the distance based measures, Relief, which is firstly proposed by Kira [21] and later enhanced to support multi-class datasets [22], is one of the most successful ones and adopted Euclidean distance to assign a relevance weight to each feature. The key idea of Relief is to iteratively estimate feature weights according to their ability to discriminate between instances that are near to each other. Having seen abroad spectrum of successful uses of Relief algorithm, Robnik-Šikonja and Kononenko [27] theoretically and empirically investigated and discussed several variations of Relief. Since Relief randomly picks out an instance from training dataset, the optimal results of Relief are not guaranteed. Liu et al. [28] applied selective sampling to Relief in order to obtain results that are better than using random sampling and similar to the results using all the instances. To overcome the disadvantage that Relief lacks a mechanism to deal with outlier data, Sun [29] proposed an iterative Relief algorithm to alleviate the deficiencies of Relief by exploring the framework of the Expectation-Maximization algorithm. Other distance based measures, such as Kolmogorov Distance and Normalized Compression Distance, are also popularly used in feature selection [23]. Hu et al. [30] introduced a concept of neighborhood margin and neighborhood soft margin to measure the minimal distance between different classes. They utilized the criterion of neighborhood soft margin to evaluate the quality of candidate features and construct a forward greedy algorithm for feature selection.

Rough set theory has been proven to be an efficient tool for modeling and reasoning with uncertainty information. Feature selection under rough set theory is a consistency-based method [31], which attempts to retain the discriminatory power of original features for the objects from the universe [32]. Recent years, researchers have focused their attention on feature selection algorithms based on rough sets [32, 33]. However, algorithms based on rough sets are often computationally time consuming. Qian et al. [34] introduced a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. Based on this framework, they also presented a general heuristic incomplete feature selection algorithm as an application of the proposed accelerator. Recently, another consistency-based method [35] have been proposed to use pairwise constraints for feature selection by Zhang et al. They devised two novel score functions based on pairwise constraints to evaluate the feature goodness and named the corresponding algorithms as Constraint Score.

The prediction capability of individual feature and the intercorrelation of feature subset are two important aspects in feature selection. There exist broadly two approaches to measure the correlation among features. One is based on classical linear correlation and the other is based on information theory [11]. Recent years a large amount of literatures on information theoretic ranking criteria have been proposed. A major advantage of information theoretic criteria is that they capture higher order statistics of the data [26]. Battiti et al. [36] investigated the application of the mutual information criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a neural network classifier. Then, an algorithm MIFS was proposed that takes both the mutual information with respect to the output class and with respect to the already selected features into account. However, the MIFS algorithm may fail when redundant features have much information about the output. Nojun et al. [37] proposed an improved algorithm of feature selection that makes more careful use of the mutual information between input attributes and others than the original MIFS. Kwak and Choi [38] proposed a new method of calculating mutual information between input and class variables based on the Parzen window, and applied this to a feature selection algorithm for several classification problems. Novovicova et al. [39] proposed a new sequential forward selection algorithm mMIFS-U that uses novel estimation of the conditional mutual information between candidate feature and classes given a subset of already selected features. Because of the difficulty in directly implementing the maximal dependency condition, Peng et al. [40] first derived an equivalent form, called minimal-redundancy-maximal-relevance criterion (mRMR), for first-order incremental feature selection. Then they presented a two-stage feature selection algorithm by combining mRMR and other more sophisticated feature selectors. Yu and Liu [41] introduced a new framework that decouples relevance analysis and redundancy analysis. They developed a correlation-based method named symmetrical uncertainty(SU) for relevance and redundancy analysis, and then removed redundant features by approximate Markov Blanket technique. In traditional selectors, mutual information is estimated on the whole sampling space. This, however, cannot exactly represent the relevance among features. To cope with this