Knowledge-Based Systems 37 (2013) 541–549

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Feature selection using dynamic weights for classification

Xin Sun [a,b], Yanheng Liu [a,c,*], Mantao Xu [b,d], Huiling Chen [a], Jiawei Han [a], Kunhao Wang [a]

[a] *College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China*
[b] *School of Computing, University of Eastern Finland, Joensuu FIN-80101, Finland*
[c] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China*
[d] *School of Electrical Engineering, Shanghai Dianji University, Shanghai 200240, China*

ARTICLE INFO

Article history:
Received 10 February 2012
Received in revised form 1 October 2012
Accepted 2 October 2012
Available online 13 October 2012

Keywords:
Machine learning
Feature selection
Information criterion
Filter method
Classification

ABSTRACT

Feature selection aims at finding a feature subset that has the most discriminative information from the original feature set. In this paper, we firstly present a new scheme for feature relevance, interdependence and redundancy analysis using information theoretic criteria. Then, a dynamic weighting-based feature selection algorithm is proposed, which not only selects the most relevant features and eliminates redundant features, but also tries to retain useful intrinsic groups of interdependent features. The primary characteristic of the method is that the feature is 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. To verify the effectiveness of our method, experimental comparisons on six UCI data sets and four gene microarray datasets are carried out using three typical classifiers. The results indicate that our proposed method achieves promising improvement on feature selection and classification accuracy.

1. Introduction

Data mining is the process of analyzing data from different perspectives and extracting it into useful information [1]. Along with new computer applications, e.g. in social networks, gene expression and combinatorial chemistry, amount of data to be analyzed is ever-increasing. Nevertheless, most of the features in huge dataset are irrelevant or redundant, which typically deteriorates the performance of machine learning algorithms. To mitigate this problem, one effective way is to reduce the dimensionality of feature space with feature selection technique [2]. The main goal of feature selection is to find the minimum subset which is optimized for the performance of machine learning algorithm. Feature selection can bring lots of benefits to machine learning algorithms [3], such as reducing the measurement cost and storage requirements, coping with the degradation of the classification performance due to the finiteness of training sample sets, reducing training and utilization time, and facilitating data visualization and understanding. Great attention has been attracted and many selection algorithms have been developed during past years. Generally, there are three kinds of feature selection methods, i.e., embedded, wrapper and filter methods. Embedded and wrapper methods are specific to a given learning algorithm. For example, Guyon et al. [4] proposed a embedded method (SVM-RFE) utilizing Support Vector Machine methods based on Recursive Feature Elimination. One drawback of these two methods is their poor generalization to other classifiers and high computational complexity in learning, because they are tightly coupled with specified learning algorithms. Filter methods are independent of learning algorithms and assess the relevance of features by looking only at the intrinsic properties of the data. In practice, filter methods have much lower computational complexity than others, meanwhile, they can achieve comparable classification accuracy for most classifiers. So far, a modest number of efficient filter selection algorithms have been proposed in literature. Among various evaluation criteria, those based on information theoretic measurements have drawn more attention because of their excellent performance (e.g. [5–8]). However, one common problem of these selectors is that they often ignore some features that have strong discriminative power as a group but are weak individually [3]. The main reason for this disadvantage is that the existing information theoretic measurements disregard the intrinsic structure [3,9] among features.

To address this problem, the present study focuses on proposing a new feature selection method that not only selects the most relevant features and eliminates redundant features, but also tries to retain useful intrinsic feature groups. The remainder of this paper is organized as follows: In Section 2, related works are briefly reviewed. Section 3 presents a new scheme for feature relevance, interdependence and redundancy analysis. Section 4 proposes a dynamic weighting-based feature selection algorithm. In Section 5, experimental results on real datasets are given to evaluate the

[*] Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China. Tel.: +86 043185159419; fax: +86 043185168337.

E-mail addresses: sunxin1984@yahoo.com.cn (X. Sun), yhliu@jlu.edu.cn (Y. Liu).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.10.001

effectiveness of our method, and some discussions are presented. 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 features from high-dimensional feature spaces [10–13]. These feature selection algorithms typically fall into two categories: feature ranking and subset selection.

Subset selection algorithms search the set of possible features for the optimal subset in which features are relevant in the given model. One critical problem for feature subset selection methods is that exhaustive search and evaluation of all the possible feature subsets usually results in a considerably high computational complexity [14]. Thus, many heuristic subset search strategies have been introduced [15,16], such as sequential forward/backward selection, random selection [17], and branch and bound search [18]. A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other [19]. Bailly and Milgram [20] presented a regression method by combining a new feature selection scheme with a specific radial basis function network. With a boosting strategy, features were evaluated by a fuzzy functional criterion using weights on examples computed from the error produced by the neural network trained at the previous step. Based on information theoretic criteria, Yu and Liu [21] introduced a new framework that decoupled relevance analysis and redundancy analysis. They developed a correlation-based subset selection method named FCBF for relevance and redundancy analysis, and then removed redundant features by approximate Markov Blanket technique.

Feature ranking algorithms rank the features by a metric and results in a rank of importance. Feature ranking based selection methods evaluate the significance of features according to some measurements, such as distance [22,23], χ^2, and information theory [24,25]. Among the distance based measures, Relief, which is firstly proposed by Kira and Rendell [22] is one of the most successful ones and adopt 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. However the optimal results of Relief are not guaranteed because Relief randomly picks out an instance from training dataset. Liu et al. [26] applied selective sampling to Relief in order to obtain results that were better than using random sampling and similar to the results using all the instances. Other distance based measures, such as Kolmogorov distance and normalized compression distance, are also popular in feature selection [23].

The prediction capability of individual feature and the inter-correlation of feature subset are two important aspects in feature ranking. There exist broadly two approaches to measure the correlation among features [3]. One is based on classical linear correlation and the other is based on information theory. Recent years have seen a large amount of literatures on information theoretic ranking criteria. A major advantage of information theoretic criteria is that they capture higher order statistics of the data. Battiti [27] investigated the application of mutual information criterion to evaluate candidate features and to select the top ranked features to be used as input data for a neural network classifier. Then, an algorithm MIFS was proposed that took 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. Novovicova et al. [28] proposed a new sequential forward selection algorithm mMIFS-U that used 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. [7] first derived an equivalent form, called Minimal Redundancy Maximal Relevance criterion (mRMR), for first-order incremental feature ranking. Then they presented a two-stage feature selection algorithm to choose salient features by wrapping a learning algorithm. To calculate the entropy and mutual information, several estimation methods are proposed in order to improve the efficiency of the information based feature selection. For example, Kwak and Choi [29] proposed a method of calculating mutual information between input and class variables based on the Parzen window. Huang and Chow [30] developed a supervised data compression algorithm to prune Gaussian probability density function estimator, then employed the estimator to estimate MI. We can find more entropy estimation methods from Ref. [31]. Sun et al. [3] proposed a Banzhaf power index method to evaluate the power of each feature, in order to select the features with high interdependence. However the Banzhaf power index method is instability when intrinsic interrelation among features is complexity. Thus Sun et al. [32] proposed another optimization algorithm based on Shapley value in order to favor the features in the smaller winning coalitions, and adopted an approximation joint mutual information metric to evaluate the relevance. The optimization algorithm method is more stability for complexity dataset, but has issue with high runtime complexity.

The aim of this section was to provide motivation and justification for the present work, not a thorough review of the feature selection methods. Readers interested in a detailed description of the methods should refer to [15,17,33].

3. Relevance, interdependence and redundancy analysis

3.1. Information theory

The fundamental concepts of information theory [34]—entropy and mutual information—provide intuitive tools to measure the uncertainty of random variables and the information shared by them. Let X be a discrete random variable and probability density function $p(x) = Pr\{X = x\}$. The entropy $H(X)$ of a discrete random variable X is defined by:

$$H(X) = -\sum_{x\in X} p(x) \log p(x). \tag{1}$$

Note that entropy is a function of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities. Furthermore, joint entropy $H(X, Y)$ extend the definition of entropy $H(X)$ to a pair of random variables and is defined as:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) \log p(x,y). \tag{2}$$

Conditional entropy $H(X|Y)$ is defined as the entropy of a random variable X conditional on the knowledge of another random variable Y. The conditional entropy $H(X|Y)$ is

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) \log p(x|y). \tag{3}$$

Mutual information (*MI*) is a measure of the amount of information shared by two variables X and Y. Consider two random variables X and Y, the mutual information $I(X; Y)$ is defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

Mutual information $I(X; Y)$ can be rewritten as $I(X; Y) = H(X) - H(X|Y)$. Thus, *MI* is the reduction in the uncertainty of one random variable due to the knowledge of the other.