Knowledge-Based Systems 89 (2015) 203–217

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Feature selection with redundancy-complementariness dispersion

Zhijun Chen [a,b,e], Chaozhong Wu [a,b], Yishi Zhang [c,f,*], Zhen Huang [d], Bin Ran [e], Ming Zhong [a,b], Nengchao Lyu [a,b]

[a] Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan 430063, China
[b] Engineering Research Center for Transportation Safety, Ministry of Education, Wuhan 430063, China
[c] School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
[d] School of Automation, Wuhan University of Technology, Wuhan 430063, China
[e] Department of Civil and Environment Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA
[f] Wisconsin School of Business, University of Wisconsin-Madison, Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 2 February 2015
Received in revised form 18 May 2015
Accepted 12 July 2015
Available online 18 July 2015

Keywords:
Classification
Feature selection
Relevance
Redundancy
Pairwise approximation
Redundancy-complementariness dispersion

ABSTRACT

Feature selection has attracted significant attention in data mining and machine learning in the past decades. Many existing feature selection methods eliminate redundancy by measuring pairwise inter-correlation of features, whereas the complementariness of features and higher inter-correlation among more than two features are ignored. In this study, a modification item concerning feature complementariness is introduced in the evaluation criterion of features. Additionally, in order to identify the interference effect of already-selected False Positives (FPs), the redundancy-complementariness dispersion is also taken into account to adjust the measurement of pairwise inter-correlation of features. To illustrate the effectiveness of proposed method, classification experiments are applied with four frequently used classifiers on ten datasets. Classification results verify the superiority of proposed method compared with seven representative feature selection methods.

1. Introduction

With the fast development of the world, the dimensional and size of data is fast-growing in most kinds of fields which challenge the data mining and machine learning techniques. Feature selection is an important and useful approach that can effectively reduce the dimensionality of feature space while retaining a relatively high accuracy in representing the original data. Thus, it plays a fundamental role in many data mining and machine learning tasks, particularly in pattern recognition, knowledge discovery, information retrieval, computer vision, bioinformatics, and so forth. The effects of feature selection have been widely recognized for its abilities in facilitating data interpretation, reducing acquisition and storage requirements, increasing learning speeds, improving generalization performance, etc. [1]. Therefore, feature selection has attracted significant attention of more and more researchers [2–8].

Generally speaking, the feature selection methods can be divided into two types: Wrapper and filter. Wrapper methods depend on specific learning algorithms. Thus the performance of wrapper methods is affected by the selected learning methods. This may makes wrapper methods computationally expensive in learning, since they must train and test classifiers for each feature subset candidate. Conversely, filter methods do not rely on any learning schemes. Instead, it is only based on some classifier-irrelevant metrics, including Fisher score [9], χ^2-test [10], mutual information [11–14], Symmetrical Uncertainty(SU) [15], etc., to estimate the discrimination power of features. Recently, new criteria and techniques such as sparse logistic regression attract increasing attention (e.g. [16]) since they have potential ability to handle very high-dimensional datasets. In this study, we only focus on filter methods.

Filter methods can also divided into feature subset selection and feature ranking ones, with regard to their search strategy. The evaluation unit for subset selection methods is a set of features, thus the set with best discrimination power is trying to be discovered [17–19]. Nevertheless, to find the best feature subset, a total of $2^m - 1$ candidate subsets (where m is # features in the original data) are possible to be traversed for feature selection task cannot be solved optimally in polynomial-time unless **P** = **NP** [20].

* Corresponding author at: Wisconsin School of Business, University of Wisconsin-Madison, 975 University Avenue, Madison, WI 53706, USA.
E-mail addresses: chenzj556@gmail.com (Z. Chen), wucz@whut.edu.cn (C. Wu), zhang685@wisc.edu (Y. Zhang), h-zhen@whut.edu.cn (Z. Huang), bran@wisc.edu (B. Ran), mzhong@whut.edu.cn (M. Zhong), lvnengchao@163.com (N. Lyu).

http://dx.doi.org/10.1016/j.knosys.2015.07.004
0950-7051/

Thus it is computationally intractable in nowadays practice, particularly in the context of big data. Unlike subset methods, feature ranking methods individually take features as the evaluation units and rank them according to their discrimination power [21,22]. These methods usually employ heuristic search strategies such as forward search, backward search, and sequential floating search.

However, whatever feature ranking or feature subsets selection methods, there are two problems possibly leading to wrong rankings or lower capacity for classification. One is that neglecting feature interaction or dependence may lead to redundancy, as some feature selection methods like MIM [23] take the assumption of independence of features. For real-world datasets, particularly those high-dimensional ones, such strong assumption may produce results far from optimal. The other problem is that group capacity of features is usually ignored, since many methods only measure the relationship between two features [11,24,22]. For example, a feature that has low individual classification capacity but is highly dependent on other features may be overlooked and even misidentified as a redundant one by only measuring its pairwise relationship with other features. However, since it is highly dependent on other features, it is also possible that it contributes largely to the discrimination power of the subset consisting of such features. Thus, it should be evaluated as a salient feature and then selected. Since the dependence among features is related to both redundancy and complementariness, it is imperative to develop more precise correlation analysis in order to distinguish them effectively. To this end, we propose a novel feature selection algorithm which tries to modify the redundancy analysis applied in prior methods by introducing a modification item and a dynamic coefficient to effectively adjust redundancy-complementariness identification. The main contributions that distinguish our work form extant studies are listed as follows:

- Complementary correlation of features is explicitly separated from redundancy.
- Redundancy-complementariness dispersion is taken into account to adjust the measurement of pairwise intercorrelation of features.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the Information theoretic metrics and evaluation criteria. A new feature selection method is included in Section 4. In Section 5, experimental study is conducted and the results are discussed. Finally, Section 6 concludes this study and proposes possible further work.

2. Related work

In recent decades, many kinds of feature selection methods have been studied. In general, there are two aims in these feature selection methods. One is to search the most class-relevant features, the other is to remove redundancy. Most feature selection algorithms can effectively find relevant features [25]. A well-known example is Relief, which is developed by Kira and Rendell [21]. The main idea of Relief is to rank features in terms of the weight corresponding to their ability to both discriminate instances with different class labels and cluster those with same class labels based on the distance between instances. However, Relief method may be ineffective since similar weights of two or more features cannot be removed by this method. In other words, this implies that redundant features cannot be identified. A typical and widely used extension of Relief is ReliefF [26], which is competent to the noisy and incomplete datasets. However, it is still unable to remove redundant features. Redundant features are considered to have negative effects on the accuracy and speed of classification methods, hence many feature selection methods are proposed to address this problem by statistic-based merics [22,27,17]. For example, Correlation based Feature Selection (CFS) algorithm proposed by Hall [27] adopts *cor* value to simultaneously measure a feature subset's correlation to the class and inter-correlation among features in it. CFS selects the subset which obtains the maximum *cor* value. However CFS does not designate specific search approaches, thus how to select feature subsets still remains to be a problem.

Minimum Redundancy and Maximum Relevance (mRMR) criterion and its variants [11,24,22] apply information theoretic metrics to separately measure class-relevance and pairwise correlation between features. A comprehensive score consisting of the two indices is applied to evaluate and select features. Fast Correlation Based Feature selection algorithm (FCBF) proposed by Yu and Liu [17] is another typical method that separately handles relevance and redundancy. FCBF utilizes Symmetrical Uncertainty (*SU*) as the metric to represent class-relevance and pairwise correlation. If the class-relevance of a feature is lower than that of another and the correlation between them, it would be identified as a redundant features and thus to be removed. Recently, an extension of FCBF, namely fast clustering-based feature selection algorithm (FAST), is proposed [28]. In this algorithm, features are firstly divided into clusters. Then for each cluster, an approximate Markov blanket based elimination strategy is applied to finally determine the selected feature subset. All of the above mentioned methods take pairwise correlation as the redundancy index and identify features with high such index to be redundant, while ignoring (1) complementary correlation between features (which we will discuss detailed in Section 3.2) and (2) correlation among more than two features, which still remain to be problems that impair the performance of feature selection.

Much effort has been made to tackle the former problem mentioned above [18,29–32,13–15,33]. Flueret [18] and Wang et al. [29] propose Conditional Mutual Information Maximization (CMIM) criterion for feature selection. CMIM harnesses Conditional Mutual Information (CMI) to measure the intensity of relevance and redundancy since CMI can implicitly identify complementary correlation between features, i.e. a large value of $CMI(F;C|\widetilde{F})$ implies (1) F is relevant to class C, and (2) F is highly complementary with $\widetilde{F}$, many information theoretic feature selection methods apply it to build up their evaluation criteria [34,31,30,35]. Algorithm based on Cumulate Conditional mutual information Minimization (CCM) criterion [13] is one of the typical algorithms that apply CMI to directly evaluate and select features. It generates candidate feature subset during the incremental step and eliminates redundancy during the shrinking step. Algorithms based on class-separability strategy extend the traditional usage of CMI in feature selection by measuring conditional mutual information between a feature and each class label [14]. Recentely, a feature selection framework basd on Data Anvelopment Analysis (DEA) is proposed [15]. Algorithm with this framework may apply MI and CMI as the evaluation indices to establish the feature evaluation system. Meanwhile, there are also several methods explicitly identifying redundancy and complementary correlation without CMI. Algorithms based on Joint Mutual Information (JMI) [32,36] take into account mutual information between a group of features and class. A typical algorithm taking JMI as metric can be found in [36], which applies JMI to measure mutual information between k features and class. Since the feature relevant to class and the one complementary to salient features will obtain high JMI values, they both will be identified as salient ones and thus is more possible to be selected. Although the above mentioned methods try to recognize complementariness from the pairwise correlation of features, measuring pairwise correlation is actually an approximation to