Knowledge-Based Systems 64 (2014) 70–80

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Feature selection using data envelopment analysis

Yishi Zhang [a], Anrong Yang [a], Chan Xiong [a], Teng Wang [b], Zigang Zhang [a,*]

[a] School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department of Computer Science and Software Engineering, Auburn University, Auburn 36849, USA

ARTICLE INFO

Article history:
Received 7 October 2013
Received in revised form 8 March 2014
Accepted 25 March 2014
Available online 5 April 2014

Keywords:
Feature selection
Data envelopment analysis
Super-efficiency
Relevance
Redundancy

ABSTRACT

Feature selection has been attracting increasing attention in recent years for its advantages in improving the predictive efficiency and reducing the cost of feature acquisition. In this paper, we regard feature selection as an efficiency evaluation process with multiple evaluation indices, and propose a novel feature selection framework based on Data Envelopment Analysis (DEA). The most significant advantages of this framework are that it can make a trade-off among several feature properties or evaluation criteria and evaluate features from a perspective of "efficient frontier" without parameter setting. We then propose a simple feature selection method based on the framework to effectively search "efficient" features with high class-relevance and low conditional independence. Super-efficiency DEA is employed in our method to fully rank features according to their efficiency scores. Experimental results for twelve well-known datasets indicate that proposed method is effective and outperforms several representative feature selection methods in most cases. The results also show the feasibility of proposed DEA-based feature selection framework.

1. Introduction

In many data mining applications, identifying the most characterizing features (or attributes, variables, hereafter they will be used interchangeably) of the observed data is critical to optimize the classification result. Tremendous new computer and internet applications, e.g. the prevalent use of social media, generate large amounts of data at an exponential rate in the world. Massive irrelevant and redundant features existing in the feature space deteriorate the performance of machine learning algorithms, and thus present challenges to feature selection.

Feature selection is desirable and essential for a number of reasons, such as reducing the complexity of training a classifier and the cost of collecting features, improving the quality of the data, and even resulting in an improvement in classification accuracy [1,2]. Roughly, there are three types of feature selection methods [3,4]: Embedded methods, filters and wrappers. As for embedded methods in C4.5 [5] or SVM-RFE [6], the process of selecting features is integrated into the learning algorithm [5]. Wrappers rely on performance estimated by a specific learning method to evaluate and select features. The drawbacks of embedded methods and wrappers are their expensive computational complexity in learning and poor generalization to other learning methods as they are tightly coupled with specified ones. In contrast, filters assess features based on some classifier-agnostic criteria (e.g. Fisher score [7], χ^2-test [3,8], mutual information [9–11], symmetrical uncertainty [1], Hilbert–Schmidt operator [12], etc.) and select features by focusing only on the intrinsic properties of the data. Developing efficient and effective filter methods has attracted great attention during past years [1,9,13,14].

Feature ranking and feature subset selection are two typical categories of feature selection regarding the output style. The former outputs ranked features weighted by their predictive power [15–17] while the latter evaluates feature subsets and searches for the best one [18,1,19–21]. Since finding an optimal subset is usually intractable and many problems related to feature selection have been shown to be NP-hard [22,23], a trade-off between result optimality and computational efficiency has been taken under consideration in literature. Heuristic methods with various feature evaluation criteria have thus been proposed [17,24,19,1]. These criteria mainly focus on the measurement of feature relevance, redundancy, conditional independence, inter-dependence, etc., and the combination of such criteria (e.g. relevance analysis with mutual information + redundancy analysis with conditional mutual information) brings about diversity of feature selection methods. Nevertheless, most of the combinations evaluate features

[*] Corresponding author. Address: School of Management, Huazhong University of Science and Technology, 1037 Luoyu Road, Hongshan District, Wuhan 430074, China. Tel./fax: +86 27 87556463.

E-mail addresses: z14@hust.edu.cn (Y. Zhang), anrong_yang@126.com (A. Yang), chanxiongcola@gmail.com (C. Xiong), tzw0019@auburn.edu (T. Wang), zigang_zhang@hust.edu.cn (Z. Zhang).

http://dx.doi.org/10.1016/j.knosys.2014.03.022
0950-7051/© 2014 Elsevier B.V. All rights reserved.

with either prior arguments or constant coefficients, and the relative importance (weight) of each feature property such as relevance or redundancy usually cannot be identified. For example, MIFS [17] applies two information-theoretic metrics to respectively measure feature dependence (D) and redundancy (R), and uses $max(D - \beta R)$ to evaluate the quality of selected features. Parameter β plays a role of mediating the weight of measured redundancy and thus any changes to it may influence the quality of the finally-selected features. Owing to more than one feature property or criterion to be considered and utilized, feature selection can thus be categorized as a multi-index evaluation process.

Data Envelopment Analysis (DEA) is an effective nonparametric method for efficiency evaluation and has been widely applied in many industries. It employs linear programming to evaluate and rank the Decision Making Units (DMUs) when the production process presents a structure of multiple inputs and outputs. Inspired by this, we regard feature selection as evaluation process with multiple inputs and outputs, and introduce in this paper a novel DEA-based feature selection framework. An effective feature selection method based on this framework is then proposed and evaluated. The remainder of the paper is organized as follows: Section 2 briefly reviews related works. Section 3 introduces some related information-theoretic metrics and Section 4 introduces a novel feature selection framework based on DEA. Then Section 5 proposes a feature selection method based on this framework. In Section 6, experimental results are given to evaluate the effectiveness of proposed method comparing with the representative feature selection methods on twelve well-known datasets, and some discussions are presented. Section 7 finally summarizes the concluding remarks and points out the future work.

2. Related work

Various aspects of feature selection have been studied for years. One of the key aspects is to measure the quality of selected features. John, Kohavi, and Pfleger classified features into three categories, i.e. strongly relevant, weakly relevant, and irrelevant ones [25], and feature selection research at that time mainly focused on searching for relevant features [26]. However, since the existence and effect of feature redundancy were pointed out [27,28,9,10], how to effectively select more relevant and less redundant features has been a hot issue in literature [18,29,1,19,30,17,24,10,31]. Although other characteristics like conditional independence [32,20,21,33] are revealed and studied, they are all variants of the basic concepts of feature relevance and redundancy. In the following text, we mainly review and analyze related work from the viewpoint of redundancy analysis.

Regarding the relationship between the class and the features, feature redundancy analysis can be divided into two categories: One only measures the correlation among features without considering the effect of the class [18,17,29,9,24,1,10,34] while the other considers such effect [27,19,30,35,32,20,21,33,12]. In other words, the former considers redundancy in an unsupervised manner, while the latter is more consistent with supervised learning scheme (Hereafter we call them unsupervised redundancy analysis and supervised redundancy analysis, respectively.).[1] Correlation-based Feature Selection (CFS) method [18] is a typical algorithm that handles redundancy by unsupervised redundancy analysis. A correlation-based metric $cor = kr_{ic}/\sqrt{k + k(k-1)r_{ij}}$ (where k is the number of the currently-selected features, r_{ic} is the average correlation between the features and the class, r_{ij} is the average inter-correlation of the features) is applied by CFS and the subset maximizing its cor value will be chosen as the final selected one. A series of representative feature selection methods with minimal Redundancy and Maximal Relevance (mRMR) criterion [17,29,9,24] also apply unsupervised redundancy analysis to measure redundancy.[2] They generally apply $D = \alpha \cdot \sum_{F_i \in \mathbf{S}} I(F_i, C)$ to measure relevance and $R = \beta \cdot \sum_{F_i, F_j \in \mathbf{S}} I(F_i, F_j)$ to measure redundancy, where I denotes mutual information and α and β are importance weights corresponding to D and R, respectively, and apply $\max(D - R)$ or $\max(D/R)$ to evaluate and select features. Since unsupervised redundancy analysis only refers to the inter-dependence among features, it is not enough to effectively identify redundancy when the class concept is considered: When the redundancy score between two features is large, it is even unable to determine which one is redundant. Most of the feature selection methods with unsupervised redundancy analysis implicitly eliminate redundancy with the help of their relevance analysis in order to get more relevant features while keeping them independent to each other, whereas exceptions also exist in literature. For example, FCBF [1] and QMIFS-p [10] handle redundancy with unsupervised redundancy analysis in a more explicit manner. FCBF sorts features in a descending order according to their relevance scores and measures the inner-correlation between any pairs of features. Then it removes redundant features via an approximate Markov blanket criterion: If $SU(F_i, C) > SU(F_j, C)$ and $SU(F_j, C) < SU(F_i, F_j)$ (where SU is symmetrical uncertainty and F_i, F_j and C are candidate features and the class, respectively), then F_j is determined as a redundant feature which should be removed. Recently, this method is extended to remove redundancy from a viewpoint of feature subsets with a cluster-based search strategy [34]. QMIFS-p applies the so-called MISF criterion to explicitly measure the similarity between the candidate feature (F) and the currently-selected features ($F_j \in \mathbf{S}$). Since criterion MISF $(F) = \text{argmax}_{F_j \in \mathbf{S}}(I(F, F_j)/H(F_j))$ estimates redundancy without considering the effect of the class, it is an unsupervised redundancy metric. Dissimilar from FCBF, QMIFS-p identifies relevance and redundancy during every iteration, which enhances the ability for identifying potential redundant features during the search process.

All of the above representative methods with unsupervised redundancy analysis measure the inter-dependence between pairs of features instead of that among the feature subsets, and hence neglect the complementarity which may provide a significant performance improvement among two or more features [3]. In addition, inter-independence may no longer exist when the class is considered. Features with less inter-dependence may be conditional dependent on each other given the distribution of the class, and hence redundancy may still exist under this circumstance. To untie the knots, a series of feature selection methods with supervised redundancy analysis are proposed [27,19,30,35,32,20,21,33,12]. Conditional Mutual Information (CMI) is an important information-theoretic metric referring to supervised redundancy analysis. It measures conditional dependence between two variables with respect to the class concept: A very small value of $I(F; C|\mathbf{S})$ (which denotes the CMI between feature F and the class C given the feature subset $\mathbf{S}$) implies that F carries little additional information about C given the subset $\mathbf{S}$, namely (a) F is redundant to $\mathbf{S}$ or (b) F is irrelevant to C. The advantage of paying attention to both relevance and redundancy makes CMI a frequently used metric in related work [19,30,35,33,34]. The algorithms with Conditional Mutual Information Maximization (CMIM) criterion [19,30] harness CMI to conduct supervised redundancy analysis. To avoid the difficulty of joint distribution estimation on insufficient samples, CMIM only uses $\widetilde{F} = \arg\min_{\widetilde{F} \in \mathbf{S}} I(F; C|\widetilde{F})$ as the conditioning feature on behalf of $\mathbf{S}$, and selects F satisfying $\max_{F \in \mathbf{F}} I(F; C|\widetilde{F})$, where $\mathbf{F}$ is the original feature set. The algorithm with Joint Mutual

[1] In [36], they are also called "redundancy" and "conditional redundancy", respectively.

[2] For the sake of convenience, here we regard MIFS [17] as a general form of mRMR.