Engineering Applications of Artificial Intelligence 44 (2015) 33–45

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Brief paper

A graph theoretic approach for unsupervised feature selection

Parham Moradi*, Mehrdad Rostami

Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

ARTICLE INFO

Article history:
Received 21 November 2014
Received in revised form
28 April 2015
Accepted 6 May 2015
Available online 27 May 2015

Keywords:
Unsupervised feature selection
Filter method
Graph clustering
Node centrality

ABSTRACT

Feature subset selection is a major problem in data mining which can help to reduce computation time, improve prediction performance, and build understandable models. Specifically, feature selection realized in the absence of class labels, namely unsupervised feature selection, is challenging and interesting. In this paper a novel graph-theoretic approach for unsupervised feature selection has been proposed. The proposed method works in three steps. In the first step, the entire feature set is represented as a weighted graph. In the second step, the features are divided into several clusters using a community detection algorithm and finally in the third step, a novel iterative search strategy based on node centrality is developed to select the final subset of features. The proposed feature selection method offers two major advantages: first, our method groups features into different clusters based on their similarities, in which the features in the same cluster are similar to each other, and to obtain the reduced redundancy set, the final subset of features is selected from different clusters. Second, the node centrality measure and term variance are used to identify the most representative and informative feature subset; hence, the optimal size of the feature subset can be automatically determined. The performance of the proposed method has been compared to those of the state-of-the-art unsupervised and supervised feature selection methods on eight benchmark classification problems. The results show that our method has produced consistently better classification accuracies.

1. Introduction

In recent years, with advancement of science and technology, datasets have grown hugely and now include large numbers of features. Accordingly machine learning methods often deal with samples consisting of thousands of features. In high-dimensional data, typically many features are irrelevant and/or redundant for a given learning task, having harmful consequences in terms of performance and/or computational cost (Cadenas et al., 2013; MonirulKabir et al., 2011; Unler et al., 2011). Moreover, a large number of features require a large storage space. To deal with such datasets, several dimensionality reduction methods have been proposed in literature with the goals of reducing the computational cost and improving the general abilities of the learning models (Liu and Zheng, 2006; Monirul Kabir et al., 2010). Clearly, computation time to build models with smaller numbers of features will be lower than that for large ones. Moreover, low-dimensional representation of the problem reduces the risk of "overfitting." Furthermore, dimensionality reduction methods provide us with a way to better understanding of the data in machine learning or pattern recognition applications.

Dimensionality reduction techniques can be categorized mainly into feature extraction and feature selection (Farahat et al., 2013; Liu and Zheng, 2006; Zhang et al., 2014). The feature extraction methods usually transform the data from the original space into a new space with lower dimension. On the other hand, the feature selection methods directly reduce the number of original features by selecting a subset of features that still retains sufficient information for classification. The goal of the feature selection methods is to seek the relevant features with the most predictive information from the original feature set. Feature selection has been established as an important technique in many practical applications such as text processing (Aghdam et al., 2009; Shamsinejadbabki and Saraee, 2011; Uğuz, 2011), face recognition (Chakraborti and Chatterjee, 2014; Kanan and Faez, 2008; Vignolo et al., 2013), image retrieval (da Silva et al., 2011; Rashedi et al., 2013), medical diagnosis (Inbarani et al., 2014), case-base reasoning (Zhu et al., 2015), collaborative filtering basd recommender systems (Ramezani et al., 2013) and bioinformatics (Jaganathan and Kuppuchamy, 2013).

The feature selection methods can be classified into four categories including filter, wrapper, embedded, and hybrid models (Cadenas et al., 2013; Hu et al., 2015; Saeys et al., 2007; Song et al., 2013). In the filter-based methods each feature is ranked without

* Corresponding author. Tel.: +98 8733660073; fax: +98 8733668513.
E-mail addresses: p.moradi@uok.ac.ir (P. Moradi), me.rostami@gmail.com (M. Rostami).

http://dx.doi.org/10.1016/j.engappai.2015.05.005
0952-1976/

consideration of any learning algorithms based on its discriminating power between different classes. The filter model can be broadly classified into univariate and multivariate approaches (Lai et al., 2006; Moradi and Rostami, 2015; Saeys et al., 2007; Tabakhi and Moradi, 2015; Tabakhi et al., 2014). In the univariate filter approach each feature is considered separately, thereby ignoring feature dependencies. These kinds of methods can effectively identify and remove irrelevant features independently of any learning algorithms, but they are unable of removing redundant features. Due to lack of consideration of possible dependency between features, these methods build weak learning models. On the other hand, some of the filter-based methods, called multivariate filter approaches, can handle both irrelevant and redundant features, which improves the accuracy of the learning model compared to the univariate filter-based feature selection methods (Ferreira and Figueiredo, 2014; Peng et al., 2005; Moradi and Rostami, 2015; Tabakhi and Moradi, 2015).

The wrapper-based feature selection methods apply a learning algorithm to evaluate the quality of feature subsets in the search space iteratively. These methods can effectively identify and remove irrelevant and redundant features. Since the wrapper approach uses learning algorithms to evaluate the selected feature subsets, it requires a high computational cost for high-dimensional datasets. In the embedded model the feature selection procedure is considered as a part of the model building process. This model can handle both irrelevant and redundant features; consequently training learning algorithms with large numbers of features will be time-consuming. Furthermore, the goal of the hybrid-based methods is to use computational efficiency of the filter model and proper performance of the wrapper model. However, the hybrid model may suffer in terms of accuracy, because the filter and wrapper models are considered as two separate steps.

According to whether the class labels of training data are available, the feature selection methods can be roughly grouped into two categories, i.e., supervised feature selection and unsupervised feature selection (Huang et al., 2012; Inbarani et al., 2014; Saeys et al., 2007). In the supervised methods, training patterns are described by the vector of feature values with a class label. The class labels are used to guide the search process for relevant information, while the unsupervised feature selection is a difficult problem. Consequently, the interesting topic of feature selection for unsupervised learning is a more complex issue, and research into this field is recently getting more attention in several communities. Term variance (TV) (Theodoridis and Koutroumbas, 2009), Laplacian score for feature selection (LS) (He et al., 2005), relevance-redundancy feature selection (RRFS) (Ferreira and Figueiredo, 2012), and unsupervised feature selection based on ant colony optimization (UFSACO) (Tabakhi et al., 2014) are some existing methods in this domain. Moreover, the unsupervised wrapper feature selection methods utilize a clustering algorithm to evaluate the quality of selected features. On the one hand, the main drawback of these methods is higher computational complexity in learning due to use of specified learning algorithms. Therefore, they are inefficient on the datasets with large numbers of features. On the other hand, the unsupervised filter approach requires the statistical analysis of the feature set only for solving the feature selection task without utilizing any learning models. A feature selection method may be evaluated according to efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. These issues are in conflict with each other: generally improving one of them reduces the other one. In other words, the filter-based feature selection methods have paid much attention to the computational time and typically are faster, while the unsupervised wrapper methods usually consider the quality of selected features. Therefore, a trade-off between these two issues has become an important and necessary goal for providing a good search method. Keeping these in mind, in this paper we propose a novel unsupervised feature selection method by integrating the concept of graph clustering with the node centrality measures. The proposed method, called the Graph Clustering with Node Centrality for unsupervised feature selection, in short GCNC, works in three steps: in the first step, the problem space is represented by a graph in which each node denotes a feature and edge weights show the feature similarities. In the second step, features are divided into several clusters using an efficient community detection algorithm. Finally in the third step, a novel search strategy is used for selecting the most relevant and influential feature from each cluster. To identify the influential feature we used Laplacian centrality (Qi et al., 2012) which is a centrality measure for weighted networks. To the best of our knowledge, this is the first paper to apply the node centrality measure to the problem of feature selection. The proposed method does not need any learning algorithms or class labels to select feature subsets; therefore, it can be classified as an unsupervised filter-based approach and will be computationally efficient for high-dimensional datasets.

The rest of the paper is organized as follows: Section 2 gives a brief review of previous works. Section 3 presents the preliminary concepts. Section 4 presents the proposed feature selection method based on a graph theoretic approach. Section 5 reports the experimental results on well-known datasets using different classifiers. Finally, Section 6 presents the conclusion.

2. Related works

Feature selection has been a fertile field of research and development since 1970s in statistical pattern recognition, machine learning, data mining, and there have been a number of attempts to review the feature selection methods (Chandrashekar and Sahin, 2014; Liu and Yu, 2005; Saeys et al., 2007). In this section, we briefly review various feature selection methods that can be classified into four categories including filter, wrapper, embedded, and hybrid approaches. Moreover, graph based feature selection methods are also reviewed.

The filter approach requires only a statistical analysis on a feature set for solving the feature selection task without utilizing any learning algorithms. Thus, the methods in this approach are typically fast. The filter-based feature selection methods can be classified into univariate and multivariate methods. In the univariate methods, informativeness of each feature is evaluated individually, according to a specific criterion, such as the Information gain (Yu and Liu., 2003), gain ratio (Mitchell, 1997), term variance (TV) (Theodoridis and Koutroumbas, 2009), Laplacian score (LS) (He et al., 2005), and Fisher score (FS) (Gu et al., 2011). This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to reduction of classification performance compared to other types of feature selection methods. In order to overcome this problem, multivariate filter based methods were introduced, which consider feature dependencies in their processes. Therefore, the multivariate approach evaluates the relevance of the features considering how they function as a group, taking into account their dependencies (Ferreira and Figueiredo, 2012; Tabakhi et al., 2014).

In the wrapper approach, each subset is evaluated by a specified learning model which is treated as a black box and is able to choose optimal or near optimal features to yield high prediction performance (Chandrashekar and Sahin, 2014; Liu and Yu, 2005). Although the wrapper models may produce better results, they are expensive to run and can break down with very large numbers of features. This is due to the use of learning