Pattern Recognition 48 (2015) 2798–2811

Contents lists available at ScienceDirect

Pattern Recognition

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

Relevance–redundancy feature selection based on ant colony optimization

Sina Tabakhi, Parham Moradi*

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

ARTICLE INFO

Article history:
Received 13 July 2014
Received in revised form
5 February 2015
Accepted 25 March 2015
Available online 8 April 2015

Keywords:
Pattern recognition
Curse of dimensionality
Feature selection
Multivariate technique
Filter model
Ant colony optimization

ABSTRACT

The curse of dimensionality is a well-known problem in pattern recognition in which the number of patterns is smaller than the number of features in the datasets. Often, many of the features are irrelevant and redundant for the classification tasks. Therefore, the feature selection becomes an essential technique to reduce the dimensionality of the datasets. In this paper, unsupervised and multivariate filter-based feature selection methods are proposed by analyzing the relevance and redundancy of features. In the methods, the search space is represented as a graph and then the ant colony optimization is used to rank the features. Furthermore, a novel heuristic information measure is proposed to improve the accuracy of the methods by considering the similarity between subsets of features. The performance of the proposed methods was compared to the well-known univariate and multivariate methods using different classifiers. The results indicated that the proposed methods outperform the existing methods.

1. Introduction

Pattern recognition is a branch of artificial intelligence whose aim is to seek to learn a model with the purpose of automatic classification of new patterns into a number of predefined classes [1]. The rapid development of information technologies in the past several decades has lead to production of datasets with large numbers of features and relatively few patterns. This presents a well-known challenge, called the *curse of dimensionality*, to pattern recognition methods and increases the computational time complexity of building the model. On the other hand, many of the features in the datasets are irrelevant and redundant for the model and may have a negative effect on the prediction accuracy [2–4].

A common way to deal with such problems is the feature selection technique. Feature selection is an important step in data preprocessing for designing many pattern recognition systems, especially in high-dimensional datasets. The goal of the feature selection technique is to seek the relevant features with the most predictive information from the original feature set. This technique reduces the dimensionality of datasets by eliminating many irrelevant and redundant features which improves the performance of the learnt model and avoids overfitting. On the other hand, this reduction helps to speed up the learning process and leads to a simple and understandable predictor model [2,5,6].

Feature selection has been established as an important technique in many practical applications of pattern recognition such as text processing [7,8], face recognition [9,10], bioinformatics [11,12], speaker verification [13], medical diagnosis [14,15], and financial domains [16,17].

The feature selection procedure needs a search strategy to explore the search space and find the optimal subset of features. This strategy requires a measure to evaluate the quality of the feature subsets. Finding the optimal subset requires exhaustive search over all possible combinations of features, meaning that its size is 2^n, where n denotes the number of features. In practical applications, the computational complexity of this approach is impractical even on moderate datasets. Therefore, it has been shown that finding the optimal feature subset is a NP-hard problem [4,18,19]. One approach for dealing with this problem is applying classical search methods such as branch and bound [20] and best first search [21] that avoid exhaustive enumeration of all subsets of features. These methods find the optimal subset, but they rely on the assumption of monotonicity and perform poorly in real-world datasets.

Thus, the other approach is proposed for finding a near-optimal feature subset with less computational effort. This approach seeks to identify and remove irrelevant and redundant features in high-dimensional datasets instead of the exhaustive search over the feature subsets. The feature selection methods in this approach can be classified into four categories including filter, wrapper, embedded, and hybrid models [2,4,6,12,22]. Some of the filter based methods use a specific criterion to evaluate the relevance of features. These kinds of methods which are called the univariate

* Corresponding author. Tel.: +98 8733660073; fax: +98 8733668513.
E-mail addresses: sina.tabakhi@ieee.org (S. Tabakhi), p.moradi@uok.ac.ir (P. Moradi).

http://dx.doi.org/10.1016/j.patcog.2015.03.020
0031-3203/

filter model can effectively identify and remove the irrelevant features independently of any learning algorithms, but they are unable of removing redundant features. Since the possible dependency between features is disregarded, these methods lead to a weak learning model. On the other hand, some of the filter based methods, called the multivariate filter model, can handle both irrelevant and redundant features which improve the accuracy of the learning model compared to that of the univariate filter based feature selection methods. The search strategy of the multivariate filter model involves only a single iteration and can easily be trapped into local optimum.

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. Due to the frequent use of the learning algorithm in the search process, this model requires high computational time, especially for high-dimensional datasets. In the embedded model the feature selection procedure is considered as a part of the process of building the model. Although this model can handle both irrelevant and redundant features, training the learning algorithms with a large number of features will be time-consuming. On the other hand, the goal of the hybrid based methods is to use the computational efficiency of the filter model and the 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.

Recently, swarm intelligence based methods have attracted a lot of attention due to their good performance in solving feature selection problems. Among the swarm intelligence based methods, ant colony optimization (ACO) has been successfully used in the feature selection area of research [6,23–25]. ACO is a multi-agent system and it has some advantages such as positive feedback, the use of a distributed long-term memory, nature implementation in a parallel way, similar function to reinforcement learning schema, and a good global and local search capability due to stochastic and greedy components in the algorithm [6,26–30]. Most swarm intelligence-based methods use a learning algorithm in their search strategies to evaluate a feature subset, and they are classified as a type of the wrapper model. Therefore, they suffer the problem of high computational time and inefficiency on the datasets with large number of features.

Since the presentation of a method to handle both irrelevant and redundant features in an acceptable time is an important issue, a major purpose of the current study is to attempt to select a high-quality feature subset within a reasonable time. In this paper, we present novel unsupervised filter based feature selection methods using ACO. They bring together the computational efficiency of the filter model and the acceptable performance of the ACO algorithm. Moreover, the methods use criteria to analyze the relevance and redundancy of the features which are used as prior knowledge in the ACO algorithm to guide the search process. In the proposed methods, each feature is ranked in the iterative improvement process of the ACO algorithm without using any learning algorithms and class labels. Also, we have proposed a new heuristic information measure which considers the similarity between subsets of features to enhance the redundancy reduction process in the proposed methods.

The rest of the paper is organized as follows. Section 2 gives a brief review of previous work. Section 3 presents the proposed feature selection methods based on the ACO. Section 4 reports the experimental results on well-known datasets using different classifiers. Finally, Section 5 presents the conclusion and future work.

2. Review of feature selection algorithms

Feature selection is a fundamental research topic in pattern recognition with a long history since the 1970s, and there have been a number of attempts to review the feature selection methods [2,12,18,31]. In this section, we briefly review various feature selection methods that can be classified into four categories including filter, wrapper, embedded, and hybrid models.

In the filter model, each feature is ranked without considering any learning algorithms based on its discriminating power between different classes. Then a subset of features with the highest ranks is selected [5]. The filter model can broadly be classified into univariate and multivariate approaches [6,12,32]. The univariate filter model uses a specific statistical criterion to evaluate the relevance of each feature individually. To this end, a number of criteria have been proposed in the literature including information gain [33,34], Gini index [33,35], gain ratio [36,37], symmetrical uncertainty [34,38], chi-square test [8], Fisher score [6,39], Laplacian score [40], Relief [41], and term variance [1,6]. The univariate filter model is computationally very efficient due to independence from any learning algorithms. Although this model removes the relevant features, it does not consider the relation between features and cannot identify the redundant features. Moreover, both theoretical and empirical studies show that redundant features also affect the accuracy and computational time of the predictor model and should be removed as well [42].

On the other hand, the multivariate filter model has been developed for solving the problem of ignoring the dependency between features. Minimal-redundancy–maximal-relevance (mRMR) [19] is a well-known multivariate filter-based method which uses an incremental search process to select a subset of features with the highest relevance to the target class based on the mutual information criterion. Moreover, this criterion is used to determine the dependency between pairs of features. Random subspace method (RSM) [32] employed a multivariate search strategy on a randomly selected subset of features to better handle the noise in high dimensional datasets. Mitra et al. [43] presented a two-stage unsupervised feature selection method based on a clustering technique. In the first stage, the original feature set is divided into a number of clusters and then in the second stage, a representative feature is selected from each cluster. Haindl et al. [44] introduced a feature selection method based on mutual correlation to identify the redundancy between features. This method iteratively removes features with the largest average mutual correlations. Fast correlation-based filter (FCBF) [34,45] is an approximation filter-based method which uses the symmetric uncertainty criterion to analyze the relevance and redundancy of the features. In this method a subset of the relevance features is selected and then the final subset is created by identifying and removing the redundant features. Relevance–redundancy feature selection (RRFS) [3] is another multivariate feature selection method based on relevance and redundancy analyses. RRFS starts the selection process with most relevant features based on a given criterion and iteratively adds the next most relevant features to the selected feature subset in a greedy way. Most of the mentioned methods are greedy sequential feature selection ones based on a single iteration search process that can easily be trapped into local optimum [6].

Recently, Tabakhi et al. [6] proposed a multivariate filter method based on the ant colony optimization, called UFSACO. The method is an iterative improvement process where each feature has a chance of being selected in all iterations. The UFSACO performs an explicit redundancy analysis and implicit relevance analysis. However, one of its main limitations is that it cannot determine the relevance of features in datasets without redundancy between features and is thus incapable of eliminating irrelevant features.

In the wrapper model, a given learning algorithm is used to select a subset of features in the search space by maximizing the accuracy of the learning algorithm. In other words, the wrapper model is an iterative search process such that the results of the learning algorithm at each iteration are used to guide the search process [5]. Generally, wrapper-based methods can be classified into greedy and random search approaches [4,12]. The greedy search approach is based on the