Information Sciences 307 (2015) 73–88

Contents lists available at ScienceDirect

Information Sciences

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

A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure

Zhichun Wang [a,b], Minqiang Li [a,*], Juanzi Li [c]

[a] College of Management and Economics, Tianjin University, Tianjin 300072, PR China
[b] College of Information Science and Technology, Beijing Normal University, Beijing 200875, PR China
[c] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 27 June 2013
Received in revised form 3 February 2015
Accepted 15 February 2015
Available online 21 February 2015

Keywords:
Feature selection
Data mining
Mutual information
Multi-objective evolutionary algorithm

ABSTRACT

Feature selection is an important task in data mining and pattern recognition, especially for high-dimensional data. It aims to select a compact feature subset with the maximal discriminative capability. The discriminability of a feature subset requires that selected features have a high relevance to class labels, whereas the compactness demands a low redundancy within the selected feature subset. This paper defines a new feature redundancy measurement capable of accurately estimating mutual information between features with respect to the target class (MIFS-CR). Based on a relevance measure and this new redundancy measure, a multi-objective evolutionary algorithm with class-dependent redundancy for feature selection (MECY-FS) is presented. The MECY-FS algorithm employs the Pareto optimality to evaluate candidate feature subsets and finds compact feature subsets with both the maximal relevance and the minimal redundancy. Experiments on benchmark datasets are conducted to validate the effectiveness of the new redundancy measure, and the MECY-FS algorithm is verified to be able to generate compact feature subsets with a high predictive capability.

1. Introduction

High-dimensional data with irrelevant and redundant features present a challenge to many data-mining algorithms. Dimension reduction is the process of reducing the number of features that describe data, and it mainly includes two approaches: *feature extraction* and *feature selection* [15]. Feature extraction refers to algorithms that create new features based on transformations or combinations of the original feature set. Principal component analysis [18] and independent component analysis [26] are two widely used feature extraction algorithms. New features generated by feature extraction may provide a better discriminative ability than original features, but these new features cannot retain an exact physical meaning [15]. Feature selection is the process of selecting the best subset of the input feature set to maximise discrimination capability. Unlike feature extraction, the results obtained by feature selection maintain the original form of the selected features. This property has led to the widespread study of feature selection. The potential benefits of feature selection include the facilitation of data visualisation and understanding, a reduction on measurement and storage requirements, the increased efficiency in training and utilisation, and the alleviation of dimensionality to improve prediction performance

* Corresponding author.
E-mail addresses: zcwang@bnu.edu.cn (Z. Wang), mqli@tju.edu.cn (M. Li), lijuanzi@tsinghua.edu.cn (J. Li).

http://dx.doi.org/10.1016/j.ins.2015.02.031
0020-0255/

[14]. Many successful, practical applications of feature selection have also been reported, such as web page classification [30], software fault prediction [4], and bioinformatics [29].

Dash and Liu [6] presented a framework for typical feature selection algorithms consisting of four basic steps: a generation procedure, an evaluation function, a stopping criterion, and a validation procedure. Among them, the generation procedure and evaluation function are the two major steps. The generation procedure is a search process that generates feature subsets for evaluation. Various search strategies in this procedure include *complete*, *heuristic*, and *random* strategies. The evaluation function aims to measure the discriminating ability of a feature subset to distinguish different class labels. Dash and Liu divided evaluation functions into five categories: *distance*, *information*, *dependence*, *consistency*, and *classifier error rate* functions. Algorithms using these four evaluation functions are known as the filter feature selection approach, while algorithms using classifiers are called the wrapper feature selection approach. In general, wrapper methods achieve better results, but filter methods run more efficiently. Among all of the evaluation criteria used in filter methods, the information metric (i.e., mutual information) has attracted the most attention because it can well quantify the correlation between features, and it is not sensitive to noise or outlier data [16]. Many mutual information-based feature selection algorithms have been proposed, such as MIFS [2], NMIFS [8], MIFS-U [21] and mRMR [25], but these algorithms have two limitations. First, they adopt greedy searching to incrementally select features, which usually generates local optimal solutions. Second, the feature selection criteria of these algorithms combine feature relevance and redundancy measures and use parameters to control the tradeoff between relevance and redundancy, which should be dynamically changed based on different datasets.

Genetic algorithms have long been used in feature selection algorithms [13,16,38,40] because their genetic search capabilities can help these algorithms effectively avoid being trapped in local optima. Most recently, some feature selection approaches based on multi-objective evolutionary algorithms have also been proposed [11,12,22,32,33,37,38]. Most of these approaches (e.g. [11,12,22,33,37,38]) employ multi-objective optimisation techniques to simultaneously minimise the classification error and the size of selected features, and these approaches are all wrapper methods and often consume very long running time when dealing with large datasets. The approach proposed by Spolaôr et al. [32] is a filter one, which uses the compromise programming technique to select a single solution from the Pareto optimal solutions returned by NSGA-II; this approach costs less time than wrapper approaches, but it does not take specific classifiers into account when selecting features.

This paper investigates several greedy feature selection algorithms based on mutual information and presents a new class-dependent redundancy measure for feature selection. The new measure can accurately estimate the correlation between features by taking the class variable into account. This paper also presents a multi-objective evolutionary algorithm using class-dependent redundancy for feature selection (MECY-FS). The MECY-FS algorithm takes advantage of genetic searching and multi-objective optimisation to overcome the limitations of greedy feature selection algorithms. MECY-FS is a hybrid approach that combines both filter and wrapper methods to evaluate the feature subsets. During the evolution process, relevance and redundancy metrics based on mutual information are used to evaluate feature subsets; after a set of Pareto optimal feature subsets is obtained, specific classifiers are trained on each feature subset to test the classification error rates; for a certain classifier, the feature subset that produces the minimum classification error rate is selected as the final output of MECY-FS. Therefore, MECY-FS tends to select compact and predictive feature subsets for specified classifiers.

The rest of this paper is organised as follows. Section 2 presents a background on entropy and mutual information. Section 3 reviews related work, and Section 4 presents the proposed redundancy measure. Section 5 describes the proposed algorithm MECY-FS, and Section 6 reports the experimental results. Section 7 concludes the paper and points out future research directions.

2. Preliminaries

In Shannon's information theory [31], *entropy* is a key information measure indicating the uncertainty of random variables. If X is a random variable with discrete values and $p(x) = \Pr(X = x)$ is the probability density function of X, then the *entropy* of X can be defined as

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

For two discrete random variables X and Y with joint probability density $p(x,y)$, the *joint entropy* of X and Y is defined as

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

When some variables are known and others are not, the remaining uncertainty is measured by *conditional entropy*. Let variable Y be given. Then the *conditional entropy* $H(X|Y)$ of X with respect to Y is

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

where $p(x|y)$ is the posterior probability of X given Y. According to this definition, if $H(X|Y)$ is zero, then X completely depends on Y; otherwise, $H(X|Y) = H(X)$ indicates that Y offers no useful information about X. *Joint entropy* and *conditional entropy* are related as follows: