Pattern Recognition 77 (2018) 20–29

Contents lists available at ScienceDirect

Pattern Recognition

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

Feature selection method with joint maximal information entropy between features and class

Kangfeng Zheng[a], Xiujuan Wang[b,*]

[a] School of Cyberspace Security, Beijing University of Posts and Telecommunications, 100876, Beijing, China
[b] Faculty of Information Technology, Beijing University of Technology, 100124, Beijing, China

ARTICLE INFO

Article history:
Received 16 May 2017
Revised 3 December 2017
Accepted 9 December 2017
Available online 12 December 2017

Keywords:
BPSO
Entropy
Feature selection
Maximal information coefficient

ABSTRACT

Feature selection remains a popular method for quantity reduction of attributes of high-dimensional data, to reduce computational costs in classifications. A new feature selection method based on the joint maximal information entropy between features and class (FS-JMIE) is proposed in this paper. Firstly, the joint maximal information entropy (JMIE) is defined to measure a feature subset. Next, a binary particle swarm optimization (BPSO) algorithm is introduced to search the optimal feature subset. Finally, classification is performed on UCI corpora to verify the performance of our proposed method compared to the traditional mutual information (MI) method, CHI method, as well as a binary version of particle swarm optimization-support vector machines (BPSO-SVMs) feature selection. Experiments show that FS-JMIE achieves an equal or better performance than MI, CHI, and BPSO-SVM. Further, FS-JMIE manifests relatively better robustness to the number of classes. Moreover, the method shows higher consistency and better time-efficiency than BPSO-SVM.

1. Introduction

Classification has been widely applied in many fields, e.g., image processing, text classification, and attack detection. Many applications such as gene expression analysis in bioinformatics involve a huge number of features and a comparatively smaller sample number [1,2], which leads to the "curse of dimensionality" [3]. Therefore, dimensionality reduction is developed to solve the aforementioned problem. Feature selection and feature transformation are two main forms of dimensionality reduction [4], wherein feature selection aims at finding the most informative feature subset from the original feature space.

Feature selection can be divided into Wrappers, Filters, and Embedded methods based on the evaluation criteria [5–8]. Wrapper methods can enable the most effective and optimal approach while suffering from heavy computational burden [9]. Filter methods are usually preferred in real applications due to their simple and light computations. Embedded techniques aim at evaluating the optimal subset of features in the learning phase [10].

Moreover, feature selection algorithms can be categorized into two families, supervised and unsupervised, according to whether the data label is involved or not. Supervised feature selection selects a feature subset based on labeled data, while unsupervised feature selection has no prior knowledge on the true functional classes [11]. Four primary steps are involved in the supervised method [6]: evaluation criteria; search approaches; stopping criterion; validation methods. The evaluation criteria will decide what kind of features can be kept and further influence the final performance. Search approaches will decide whether a suitable feature can be found and how fast the selection is executed.

Feature selection remains an active research focus with the increasing application of classification, which inspires our research. A new feature selection method based on the joint maximal information entropy between features and class (FS-JMIE) is proposed in this paper. Firstly, joint maximal information entropy (JMIE) is defined to evaluate a feature subset. Secondly, a binary particle swarm optimization (BPSO) algorithm is introduced to fly within the possible subset space in order to determine a solution. The paper is organized as follows: Section 2 summarizes related work; Details are provided in Section 3; In Section 4, we present the experimental results and a conclusion is presented at the end.

2. Related work

Evaluation criteria play an important role in feature selection and many relevant studies have been conducted. Traditional evaluation criteria involve, e.g., statistical correlation methods, consistency-based methods, and mutual information (MI). MI mea-

[*] Corresponding author.
E-mail addresses: kfzheng@bupt.edu.cn (K. Zheng), xjwang@bjut.edu.cn (X. Wang).

https://doi.org/10.1016/j.patcog.2017.12.008
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

sures the mutual dependence between two variables in the probability and information theory. It is introduced to select features by quantifying the "amount of information" obtained on the class variable through the attribute [12–15]. CHI-square is a generally applied nonparametric hypothesis test method that analyzes the correlation between two or more samples. It has been frequently adopted in feature selection [16]. Besides, Varshavsky et al. [17] developed a novel unsupervised feature-filtering method based on SVD-entropy for biological data. A feature was selected according to its contribution to the entropy (CE) calculated on a leave-one-out basis. Banerjee et al. [1] modified the definition of CE in [17] and extended the SVD-entropy-based approach to a supervised framework. Monirul et al. [14] presented a feature selection algorithm using neural networks where the neural networks architectures were determined automatically during the feature selection process. The correlation information in this algorithm was used to select less correlated features if they enhanced the Acc of the neural networks. Zhao et al. [15] put forward a method named the community modularity Q value-based feature selection. First, a feature vector graph is constructed. Second, the community modularity Q value is calculated for each feature and used as metric to determine the most informative features. The maximal information coefficient (MIC) is a novel metric identifying the correlations between pairs of variables [18]. The basic idea: if a relationship exists between variable pairs, then a grid can be drawn on their scatter plot that partitions the data encapsulating that relationship. The author claims that this new method possesses the two properties of generality and equitability and it could identify the widely correlations between two variables. Later, MIC was introduced in feature selection. In [19], MIC was used to measure the correlation strength between each feature and the label. With a threshold determined in the experiment, features that had an MIC larger than the threshold were selected. In [20], the features that had small MICs with classes were firstly removed. Then, the best first search strategy was employed to further reduce the feature number.

Moreover, to determine the optimal feature subset within the subset space, many studies have been conducted. The top-ranking method [21,22] selects the "top-*K* best features" according to the evaluation criteria between a feature and the class label. This method is simple and needs light computations. However, the choice of *K* will influence the feature selection result. Exhaustive search and evaluation of all the possible feature subsets is an NP-hard problem [23]. Therefore, researchers have developed many heuristic subset search strategies [24–26]. Huang et al. [24] proposed a feature selection method combining a genetic algorithm in the global search for the best subset of features in a wrapper manner with the local search in a filter manner based on conditional mutual information. Gheyas et al. [25] proposed a hybrid search algorithm, named SAGA, based on simulated annealing, a genetic algorithm, a generalized regression neural network, and a greedy search algorithm without including filter steps. Furthermore, swarm intelligence algorithms have been combined with feature selection in recent years, e.g., the ant colony algorithm and particle swarm optimization algorithm (PSO). Zorarpacı et al. [26] developed a hybrid feature selection method combining the artificial bee colony optimization technique with the differential evolution algorithm for classification tasks. Due to its briefness and global capability [27], the BPSO [28] algorithm has gained more attention from researchers [29–34]. In [29], the bare-bones particle swarm optimization (BBPSO) was proposed to find the optimal feature. Here, a reinforced memory strategy was designed to update the local leaders of particles to avoid the degradation of outstanding genes in the particles. Further, a uniform combination was proposed to balance the local exploitation and the global exploration of the algorithm. In [30], a fast Hamming distance—and BPSO—based algorithm is proposed to select important features from gene expression data. Zhang et al. [31] proposed a multi-objective feature selection algorithm based on BBPSO with two new operators, i.e., a reinforced memory strategy and a hybrid mutation. Qiao et al. [32] proposed a discrete binary particle swarm optimization-support vector machines (BPSO-SVMs) wrapper-mode feature selection algorithm that searches the feature space guided by the result of an SVM 10-fold crossover validation. In [33], a probability-based encoding technology and an effective hybrid operator—together with the ideas of the crowding distance, external archive, and Pareto domination relation—were applied to PSO. However, the fitness of a particle in most studies has been defined as the classification performance based on the corresponding feature subset of the regarding particle and the whole feature selection process needs more training time.

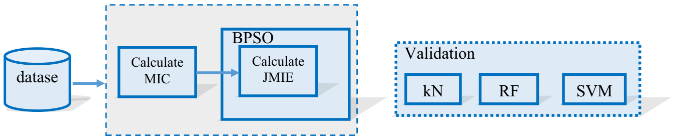

Fig. 1. Framework of our work.

The development of a highly accurate and fast search algorithm to select the optimal feature subset is an open problem that impels our work. Two problems will be stressed, i.e., the definition of new evaluation criteria and application of the search algorithm.

3. FS-JMIE algorithm

This paper presents a new supervised feature selection scheme named FS-JMIE that defines a metric to measure the performance of the feature subset and then introduces the BPSO algorithm to achieve subset searching, as illustrated in Fig. 1.

3.1. JMIE

MIC has drawn much attention because of its generality, equitability, and especially its effectiveness in identifying the widely correlations between two variables [18]. The existing feature selection algorithms-based MIC [19,20] can only measure the relativity between a feature and the class. However, it is also required to evaluate the relativity among a set of features. In this paper, a new metric named FS-JMIE is proposed, which takes into account both the relativity among features and that between features and the class.

Let us assume that corpus D comprises M data points characterized as an N-dimensional vector $\boldsymbol{d_i} = (d_{i,1}, d_{i,2}, \cdots, d_{i,N})$. Each dimension corresponds to a defined attribute A_i and D is labeled with L categories C, $C = \{c_j\}$, $j = 1, 2, \cdots, L$. Based on the labeled data, each feature/attribute A_i can be quantified with an M-dimensional vector denoted as $\boldsymbol{a_i} = (a_{i,1}, a_{i,2}, \cdots, a_{i,M})$ with $a_{i,k} = d_{k,i}$. According to the definition of MIC [18], we can obtain the MIC matrix between two random features from the feature set as follows:

$$\tilde{\boldsymbol{R}} = \left\{\tilde{r}_{ij}\right\}_{1\le i,j\le N} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1N} \\ r_{21} & 1 & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & 1 \end{bmatrix}, \quad r_{ij} = MIC\left(\boldsymbol{a}_i, \boldsymbol{a}_j\right), \tag{1}$$

where $\tilde{\boldsymbol{R}}$ carries all relativity information of features. However, the influence of each feature on the class is excluded in $\tilde{\boldsymbol{R}}$. Therefore, a joint MIC matrix $\boldsymbol{R}$ for N features and the class is constructed as