Expert Systems With Applications 53 (2016) 27–42

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unsupervised probabilistic feature selection using ant colony optimization

Behrouz Zamani Dadaneh*, Hossein Yeganeh Markid, Ali Zakerolhosseini

Faculty of Computer Science and Engineering, Shahid Beheshti University, G.C, Tehran, Iran

ARTICLE INFO

Keywords:
Feature selection
Unsupervised methods
Filter approaches
Ant colony optimization
Classification accuracy

ABSTRACT

Feature selection (FS) is one of the most important fields in pattern recognition, which aims to pick a subset of relevant and informative features from an original feature set. There are two kinds of FS algorithms depending on the presence of information about dataset class labels: supervised and unsupervised algorithms. Supervised approaches utilize class labels of dataset in the process of feature selection. On the other hand, unsupervised algorithms act in the absence of class labels, which makes their process more difficult. In this paper, we propose unsupervised probabilistic feature selection using ant colony optimization (UPFS). The algorithm looks for the optimal feature subset in an iterative process. In this algorithm, we utilize inter-feature information which shows the similarity between the features that leads the algorithm to decreased redundancy in the final set. In each step of the ACO algorithm, to select the next potential feature, we calculate the amount of redundancy between current feature and all those which have been selected thus far. In addition, we utilize a matrix to hold ant related pheromone which shows the rate of the co-presence of every pair of features in solutions. Afterwards, features are ranked based on a probability function extracted from the matrix; then, their m-top is returned as the final solution. We compare the performance of UPFS with 15 well-known supervised and unsupervised feature selection methods using different classifiers (support vector machine, naive Bayes, and k-nearest neighbor) on 10 well-known datasets. The experimental results show the efficiency of the proposed method compared to the previous related methods.

1. Introduction

In pattern recognition, feature selection (FS) is a procedure to select the informative features from an original set by eliminating irrelevant and redundant features. The availability of large amounts of features represents a challenge to classification problems. Using all features requires the estimation of a considerable number of parameters during the classification process. Therefore, each feature used in the classification process should add extra information (Liu, Motoda, & Yu, 2004; Song, Ni, & Wang, 2013; Uysal & Gunal, 2012). On the other hand, feature selection aims to reduce the number of features, thus directly targeting the curse of dimensionality problem and often allowing learning algorithms to obtain better performing classifiers (Farmer, Bapna, & Jain, 2004). Feature selection has many benefits such as facilitating data visualization and data understanding, reducing measurement and storage requirements, reducing training and utilization times, and defying the curse of dimensionality to improve prediction performance (Guyon & Elisseeff, 2003). Feature selection is widely applied in different areas such as bioinformatics, face recognition, text categorization, data mining, gene microarray analysis, etc. (Ding & Peng, 2005; Gheyas & Smith, 2010; Guyon, Weston, Barnhill, & Vapnik, 2002; Lazar et al., 2012; Liu & Motoda, 2007; Liu & Yu, 2005; Saeys, Inza, & Larrañaga, 2007; Song et al., 2014; Sotoca & Pla, 2010; Wei et al., 2014).

Features fall into four categories: (i) irrelevant, (ii) weakly relevant and redundant, (iii) weakly relevant but non-redundant, and (iv) strongly relevant (Yu & Liu, 2004). Relevant features are those which have a main role in the desired classification problem and represent the highest information about the problem. On the other hand, redundant features are modeled as the features which provide no more information than relevant features, but are correlated to the relevant features. Irrelevant features may have a negative impact on the accuracy of classifiers (Yu & Liu, 2004). Therefore, the purpose of feature selection methods is to eliminate redundant and irrelevant features in order to extract a subset of features that gives as much information as the whole feature set does. As a result, after feature reduction, the classifier encounters a small amount of information which helps save significant computation time and improves performance (Liu & Motoda, 2007). Feature

* Corresponding author. Tel.: +98 21 29904194; fax: +98 21 22431804.
E-mail addresses: b.zamani@mail.sbu.ac.ir (B.Z. Dadaneh), h.yeganeh@mail.sbu.ac.ir (H.Y. Markid), a-zaker@sbu.ac.ir (A. Zakerolhosseini).

http://dx.doi.org/10.1016/j.eswa.2016.01.021
0957-4174/

selection approaches are generally divided into three categories: filter, wrapper, and embedded methods (Liu & Yu, 2005).

Filter methods use statistical characteristics of data as the principal criteria for selecting the subset of features. The proper criteria are applied to rank features and a threshold is applied to select their best subset. Filter methods evaluate features without considering any learning algorithms; therefore, they are very popular for high dimensional data (Unler, Murat, & Chinnam, 2011). Some popular filter methods are mutual information based (Cover & Thomas, 1991; Fleuret, 2004; Lewis, 1992; Song et al., 2014; Tesmer & Estevez, 2004; Wei et al., 2014), fast correlation based filter (FCBF) (Yu & Liu, 2004), max-relevance and min-redundancy (mRMR) (Peng, Long, & Ding, 2005), feature selection based on interaction capping (ICAP) (Jakulin, 2005), conditional infomax feature extraction (CIFE) (Lin & Tang, 2006), and relevant feature selection (Relief-F) (Liu & Motoda, 2007).

Filter methods can be divided into univariate and multivariate methods (Lai, Reinders, & Wessels, 2006; Saeys et al., 2007). In univariate methods, the importance of feature is measured individually using an evaluation criterion, while in multivariate methods, the dependencies between features are also regarded as the importance of features (Saeys et al., 2007). The well-known univariate filter methods include information gain (Raileanu & Stoffel, 2004), gain ratio (Mitchell, 1997; Quinlan, 1986), symmetrical uncertainty (Biesiada & Duch, 2007), Gini index (Dodge, 2008), Fisher score (Gu, Li, & Han, 2012), Laplacian score (He, Cai, & Niyogi, 2005), and term variance (TV) (He et al., 2005). Also, there are well-known multivariate filter methods such as max-relevance and min-redundancy (mRMR) (Peng et al., 2005), mutual correlation (Haindl, Somol, Ververidis, & Kotropoulos, 2006), random subspace method (RSM) (Lai et al., 2006), and relevance-redundancy feature selection (RRFS) (Ferreira & Figueiredo, 2012).

Wrapper methods evaluate the subset of selected features using learning algorithms (Kohavi & John, 1997). These methods train a model to score feature subsets. In each step, a model will be trained by new features; then, the model is tested on a specific set which is called test set and the error rate of the model gives the related subset scores. Since wrapper methods train a new model for each subset, they are more computationally intensive than filter methods, but usually provide a much better feature subset for that particular type of model (Kohavi & John, 1997). The search strategies used in wrapper methods fall into two categories: sequential and random (Kabir, Shahjahan, & Murase, 2011). Sequential search methods select features sequentially and tend to become trapped in a local optimum (Theodoridis & Koutroumbas, 2008), whereas random search strategies apply randomness in their search procedures to escape local optimum solutions (Aghdam, Ghasem-Aghaee, & Basiri, 2009; Farmer et al., 2004; Meiri & Zahavi, 2006; Sikora & Piramuthu, 2007).

In embedded approaches, feature selection involves learning process; therefore, search process will be performed by a learning algorithm (Saeys et al., 2007). These methods use all the dataset and do not divide the dataset into train and test sets. In addition, the optimum subset of features is obtained earlier than the selected features in wrapper methods, because embedded methods do not evaluate each of the solutions in the same manner as wrapper methods do. Support vector machine (SVM) (Pal & Foody, 2010) and decision tree algorithm (Sugumaran, Muralidharan, & Ramachandran, 2007) are the well-known algorithms which are utilized in the construction of embedded algorithms.

From another point of view, feature selection approaches can be divided into two categories: supervised and unsupervised methods. Unsupervised methods utilize inter-feature relations to determine the relevancy of features, whereas supervised methods select features with maximum representative and discriminant power (Wang, Nie, & Huang, 2014). Recently, more unsupervised feature selection approaches have been proposed such as Laplacian score (He et al., 2005), term variance (TV) (He et al., 2005), mutual correlation (Haindl et al., 2006), random subspace method (RSM) (Lai et al., 2006), relevance-redundancy feature selection (RRFS) (Ferreira & Figueiredo, 2012), feature selection based on spectral graph theory (SPEC) (Zhao & Liu, 2007), unified trace ratio formulation and k-means clustering based feature selection (TRACK) (Wang et al., 2014), and unsupervised feature selection using ant colony optimization (Tabakhi, Moradi, & Akhlaghian, 2014).

For high-dimensional data, evaluating all states is computationally non-feasible and requires heuristic search methods (Chuang, Tsai, & Yang, 2011). Recently, nature inspired metaheuristic algorithms have been employed to select features such as genetic algorithm (De Stefano, Fontanella, Marrocco, & di Freca, 2014; Oh, Lee, & Moon, 2004; Raymer, Punch, Goodman, Kuhn, & Jain, 2000; Sikora & Piramuthu, 2007), particle swarm optimization (Chuang et al., 2011; Zhang, Gong, Hu, & Zhang, 2015), and ant colony optimization (Aghdam et al., 2009; Al-Ani, 2005; Chen, Chen, & Chen, 2013; Chen, Miao, & Wang, 2010; Vieira, Sousa, & Runkler, 2010; Kashef & Nezamabadi-pour, 2015; Markid, Dadaneh, & Moghaddam, 2015; Tabakhi et al., 2014).

Ant colony optimization (Dorigo & Caro, 1999) is a method that has been widely applied in feature selection (Al-Ani, 2005; Nemati et al., 2009). It was initially used for solving traveling salesman problem (Dorigo & Gambardella, 1997b) and has been successfully applied for a different number of problems such as classification (Dorigo & Stützle, 2010; Parpinelli, Lopes, & Freitas, 2002), image processing (Tian, Yu, & Xie, 2008) and fuzzy control design (Castillo, Lizárraga, Soria, Melin, & Valdez, 2015; Castillo, Neyoy, Soria, Melin, & Valdez, 2015). In recent years, some ACO-based methods for feature selection have been reported, most of which have used a complete graph with n nodes, where n corresponds to the number of features; however, these approaches have some variations in detail. According to these approaches, complexity of graph edges will be $O(n^2)$. Al-Ani (2005) utilized local importance features and overall performance of subsets to search through the feature space for optimal solutions in the ACO-based feature selection method. Basiri, Ghasem-Aghaee, & Aghdam (2008) proposed an ACO-based feature selection method for predicting postsynaptic activity of proteins. Also, Nemati, Basiri, Ghasem-Aghaee, and Aghdam (2009) hybridized ACO with genetic algorithm for feature selection in protein function prediction. Aghdam et al. (2009) proposed a text feature selection algorithm using ant colony optimization. The algorithm used classifier performance and length of the selected feature subset as heuristic information for ACO (Aghdam et al., 2009). Huang (2009) proposed a hybrid classification system with feature subset selection and model parameter optimization based on ACO. Nemati, Boostani, and Jazi (2008) applied an ACO algorithm to reduce size of features in automatic speaker verification. Vieira et al. (2010) proposed an algorithm for feature selection based on two cooperative ant colonies, which minimized two objectives: number of features and classification error. Chen et al. (2010) proposed a new rough set approach to feature selection based on ACO and adopt mutual information based feature significance as heuristic information. The method started from the feature core, which changed the complete graph to a smaller one (Chen et al., 2010). Xiong, Wang, and Lin (2010) presented a hybrid feature selection algorithm based on dynamic ant colony algorithm which used mutual information as heuristic function.

There is another approach to model features in a graph based on ACO which is called binary method. In this approach, features as the graph nodes stand in a sequential order one after another. There are two directed arcs between every node and its subsequent one. One of them shows the next node is selected and another shows it is not. By applying this policy, graph edge complexity (the number of edges to model the problem as a graph) is