Applied Soft Computing 65 (2018) 33–46

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel hybrid genetic algorithm with granular information for feature selection and optimization

Hongbin Dong, Tao Li*, Rui Ding, Jing Sun

College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China

ARTICLE INFO

Article history:
Received 19 August 2017
Received in revised form
27 November 2017
Accepted 28 December 2017
Available online 7 January 2018

Keywords:
Feature selection
Granular computing
Genetic algorithm
Rough set
Parameter optimization

ABSTRACT

Feature selection has been a significant task for data mining and pattern recognition. It aims to choose the optimal feature subset with the minimum redundancy and the maximum discriminating ability. This paper analyzes the feature selection method from two aspects of data and algorithm. In order to deal with the redundant features and irrelevant features in high-dimensional & low-sample data and low-dimensional & high-sample data, the feature selection algorithm model based on the granular information is presented in this paper. Thus, our research examines experimentally how granularity level affects both the classification accuracy and the size of feature subset for feature selection. First of all, the improved binary genetic algorithm with feature granulation (IBGAFG) is used to select the significant features. Then, the improved neighborhood rough set with sample granulation (INRSG) is proposed under different granular radius, which further improves the quality of the feature subset. Finally, in order to find out the optimal granular radius, granularity λ optimization based on genetic algorithm (ROGA) is presented. The optimal granularity parameters are found adaptively according to the feedback of classification accuracy. The performance of the proposed algorithms is tested upon eleven publicly available data sets and is compared with other supervisory methods or evolutionary algorithms. Additionally, the ROGA algorithm is applied to the enterprise financial dataset, which can select the features that affect the financial status. Experiment results demonstrate that the approaches are efficient and can provide higher classification accuracy using granular information.

1. Introduction

With the rapid development of information technology, a large number of various types of data is being gathered. However, the data itself does not provide sufficient knowledge, and there are many noisy data that should be identified. Otherwise it will weaken the decision-making. Feature selection (FS) is one of the most fundamental data preprocess in data mining. The purpose of FS is to select the optimal feature subset that contains the most valuable information for making better decisions. In addition, FS may work on a better comprehension of the domain, by maintaining only the features with a good ability, according to some importance criterion, to describe the inherent patterns within the data and helps to reduce the effects of the curse of dimension [1]. The optimal feature set may be estimated according to maximize or minimize function that associates with the importance of feature.

FS is inherently a combinatorial optimization problem [1]. It is a difficult task due mainly to the large search space, where the total number of possible solutions is 2^n for a dataset with n features. There are available search strategies, such as complete search, greedy search, heuristic search, and random search. But most existing feature selection methods suffer high time consumption and low recognition accuracy due to the redundant and irrelevant features. Sometimes, it also leads to local optimum. Therefore, constructing model to reduce feature space effectively is integral. Compared to the total feature space, the smaller feature space not only improves the computational speed, but also provides a more compact model with better generalization capability. Mao and Tsang [2] proposed a two-layer cutting plane algorithm to search for the optimal feature subsets. Min et al. [3] developed a heuristic search and a backtracking algorithm to solve feature selection problems using rough set theory. The results show that the performance obtained by heuristic search techniques is similar to the backtracking algorithm. But the heuristic search cost less time. However, the papers barely mention the effect of decision structure information on the induced knowledge granules. Evolutionary computation (EC) techniques as effective methods have

* Corresponding author.
E-mail address: litao16@hrheu.edu.cn (T. Li).

https://doi.org/10.1016/j.asoc.2017.12.048
1568-4946/

been applied to search for the optimal feature subset. Compared with traditional search methods, the EC techniques do not need domain knowledge and do not make any assumption of the search space, such as whether it is linearly or nonlinearly separable, and differentiable [4]. Another significant advantage of EC techniques is that their population-based mechanism can produce multiple solutions on single run. In the literatures [5–7], ant colony optimization (ACO), particle swarm optimization (PSO) and differential evolution (DE) algorithm are adopted to optimize the feature space. Although the result may be suitable for some tasks, the issue of unstable still causes more time overhead. Therefore, with search space increasing, further research on the stability of the algorithm is particularly important.

However, no search technology will work well if the underlying evaluation measure is poor. According to learning strategies, feature selection can be divided into supervised learning and unsupervised learning. In the former case, the class information are utilized for choosing features; otherwise one uses the distribution of sample space to speculate the possible hidden patterns. Based on the feature evaluation measure, FS can be classified into three categories, namely filters, wrappers and embedded methods [8]. The main difference is that wrapper approaches consist of a classification/learning algorithm to evaluate the goodness of the feature subset. Owing to more attention is paid on the classification performance, wrapper algorithms are usually more time-consuming than filter algorithms. But for a particular classification algorithm, Wrappers usually perform better than filters. However, filters have the advantage of lower cost without considering classification accuracy. Finally, as for the embedded methods, a hybrid Fuzzy Min-Max-CART-Random Forest algorithm proposed by Seera and Lim [9] can be provided as a recent example. Chen et al. [10] proposed the CSMSVM method utilized a novel feature selection criteria, namely the margin verses cosine distance ratio, which adds the weight value of the features to maximize the margin verses cosine distance ratio. In an unsupervised framework, the normalized mutual information score is employed for computing both the similarity and the dissimilarity in literature [11]. A study by Brown et al. [12] also notes the good performance of the conditional mutual information maximization measure using the K-nearest neighbors (KNN) classifier. Although the search mechanism and evaluation criteria in the literatures have improved the performance of the feature selection algorithm to a certain extent, the distribution of the data and the convergence of the algorithm have not been considered. At the same time, there is a lack of analysis of the feature importance from the view of granularity.

In this paper, the feature selection problem is first studied from the perspective of sample granulation and feature granulation, and the novel hybrid genetic algorithm with granular information for feature selection and optimization is proposed. In view of the advantage of global search ability, the improved binary genetic algorithm with feature granulation (IBGAFG) is developed as the feature space search strategy. Each chromosome represents a solution from the perspective of the feature space, and the average combination granularity is used to achieve feature granulation. The size of the feature subset represents the size of the granulation. For a certain granularity, the evolutionary calculation of the current feature subspace is carried out according to the evaluation criterion, and finally the optimized feature subset is obtained. In the aspect of sample granulation, this paper applies the improved neighborhood rough set (INRSG) to the feature selection, and the neighborhood strategy is adopted to granulate the sample space. We measure the influence of features on classification models according to the sample distribution, and analyze the dependency degree of the feature by calculating the maximum lower approximation of the sample.

We focus on granular information about feature and sample that are useful in evaluating feature selection. Thus, we capture this idea and design a framework that the feature granular information and the sample granular information are combined for feature selection. On one hand, the basic units are the new designed chromosomes that control feature granularity. One chromosome usually represents a feasible feature subset. The individual symbol that forms the chromosome is called the key value; on the anther hand, the majority decision information of the neighborhood boundary region of the neighborhood rough set is considered to improve the capability of neighborhood dependency for feature selection. Because the best classification performance is closely related to the reasonable granularity value, so granularity λ optimization based on genetic algorithm (ROGA) is proposed. Therefore, the paper presents a feature selection model based on granular information, as shown in Fig. 1.

The rest of the paper is organized as follows: In Section 2, related works on feature selection are reviewed, and the formal description of the genetic algorithm and the knowledge granularity in the rough set theory is also discussed. Moreover, a detailed description of the proposed methodology is presented in Section 3. In Section 4, the experimental results and comparisons with other algorithms are given. Finally, the paper is summarized in Section 5.

2. Related work

2.1. Genetic algorithm

Genetic algorithm (GA), well-known for global search potential, has received much attention recently to the feature selection researchers. Aiming at feature combination optimization, the basic introduce to genetic algorithm is given in this paper.

There is a finite length of the string $A = a_1, a_2, \ldots, a_L$, and $A \Rightarrow X = (x_1, x_2, \ldots, x_n)^T$. A is called the encoding of X, denoted by $e(X)$, while X is called the decoding of A, denoted by $X = e^{-1}(A)$. The a_i can be seen as gene, and the value of a_i is binary value or floating point number, where the encoding length is L. Thus the $H_L = \{A = a_1, a_2, \ldots, a_L | a_i \in \Gamma, i = 1, 2, \ldots, L\}$ is demoted as individual space, and different individuals form the population space H_p. For one optimization problem, it is necessary to search the space H_p and to compute the fitness of the individual in H_p so that A satisfies $\max_{A \in H_p}(A)$.

The formal description of the genetic algorithm is presented in this paper. In the tuple $GAF(e, J, S, E, \Psi)$, where e, J, S, E and Ψ respectively denote coding format, fitness metric, selection operator, genetic operator and parameter set. Therefore, the basic process of genetic algorithm is as follows: (1) initialization. Randomly generate N individuals as the initial population, and set the number of evolution as well; (2) individual evaluation. Calculate or evaluate the fitness of each individual according to the evaluation criteria; (3) population evolution. Employ the selection operation, the crossover operation and the mutation operation to produce the next generation; (4) termination test. If the maximum fitness of the individual is the optimal solution or the maximum number of iterations, then terminate the calculation, otherwise return (2).

2.2. Neighborhood rough set

In the information system, the granularity is the average measure of the degree of information and knowledge. The knowledge granularity has the ability to describe knowledge differences. The smaller the knowledge granularity and the stronger its ability to distinguish; while the larger the knowledge granularity and the weaker the ability to distinguish. Given the classification data and it can be formalized as a decision system $DS = <U, A, D>$. The $U = \{x_1, x_2, \ldots, x_n\}$ is the sample set, and the $A = \{a_1, a_2, \ldots, a_N\}$ is attribute set that describe the sample, while D is the classification decision attribute.