Applied Soft Computing 29 (2015) 371–378

Contents lists available at ScienceDirect

Applied Soft Computing

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

Using group genetic algorithm to improve performance of attribute clustering

Tzung-Pei Hong [a,b,*], Chun-Hao Chen [c], Feng-Shih Lin [b]

[a] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC
[b] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan, ROC
[c] Department of Computer Science and Information Engineering, Tamkang University, Taipei 251, Taiwan, ROC

ARTICLE INFO

Article history:
Received 14 December 2012
Received in revised form 4 September 2014
Accepted 6 January 2015
Available online 21 January 2015

Keywords:
Attribute clustering
Feature selection
Genetic algorithm
Grouping genetic algorithm
Data mining

ABSTRACT

Feature selection is a pre-processing step in data mining and machine learning, and is very important in analyzing high-dimensional data. Attribute clustering has been proposed for feature selection. If similar attributes can be clustered into groups, they can then be easily replaced by others in the same group when some attribute values are missing. Hong et al. proposed a genetic algorithm (GA) to find appropriate attribute clusters. However, in their approaches, multiple chromosomes represent the same attribute clustering result (feasible solution) due to the combinatorial property, and thus the search space is larger than necessary. This study improves the performance of the GA-based attribute clustering process based on the grouping genetic algorithm (GGA). In the proposed approach, the general GGA representation and operators are used to reduce redundancy in the chromosome representation for attribute clustering. Experiments are also conducted to compare the efficiency of the proposed approach with that of an existing approach. The results indicate that the proposed approach can derive attribute grouping results in an effective way.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Feature selection is an important pre-processing step in data mining and machine learning [8]. An appropriate subset of features not only reduces the execution time required for deriving rules [2], but also improves classification accuracy. Feature selection is also critical to data classification and data retrieval, and has been widely used in many research fields, such as pattern recognition, statistics, and data mining. Since feature selection is an optimization problem, there are many techniques could be utilized. Some well-known approaches like genetic algorithms [15,17], particle swarm optimization [20,21], ant colony optimization [13], and other bio-inspired optimization algorithms [9,10,22].

There are many GA-based approaches which have been proposed for feature selection [15,17]. In addition, some PSO- or ACO-based algorithms have also been proposed for the feature selection problems [13,21]. For example, in [21], a multi-objective particle swarm optimization (PSO) approach was presented for feature selection. The goal of that approach was to generate a set of Pareto solutions (feature subsets) for classification. In [13], a hybrid metaheuristic named ant-cuckoo colony optimization was proposed, which was a hybrid of ant colony optimization and cuckoo search for feature selection in digital mammogram. In other words, the main purpose of those approaches was to provide a set of selected attributes for classification.

However, in order to overcome the problem of high dimensionality, various feature selection techniques have been proposed [1,23]. A good feature subset can help the adopted learning algorithm get better results in less time. Finding an optimal feature subset has been shown to be an NP-hard problem [3]. In 2007, Hong and Liou proposed a feature selection approach based on the concept of feature clustering [12]. In their approach, the problem of selecting K features from input features could be considered as a K-clustering problem, with each cluster providing one feature as a member of the final feature subset. This approach can find an approximate feature subset for classification. In addition, the algorithm can find another feature to replace the selected feature if the feature value of the object is missing. Based on the same idea, Hong and Wang [16,17] proposed genetic algorithm (GA)-based clustering methods for attribute clustering to find an approximate feature subset for classification. As Falkenauer pointed out, the

* Corresponding author at: Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC. Tel.: +886 75919191.
E-mail addresses: tphong@nuk.edu.tw (T.-P. Hong), chchen@mail.tku.edu.tw (C.-H. Chen), m983040076@student.nsysu.edu.tw (F.-S. Lin).

http://dx.doi.org/10.1016/j.asoc.2015.01.001
1568-4946/© 2015 Elsevier B.V. All rights reserved.

general GA has some weakness when solving grouping problems [6]. Because of the encoding scheme, multiple chromosomes would map to the same attribute clustering result (feasible solution) due to the combinatorial property, and thus the search space is larger than necessary.

Furthermore, the GA operations cannot distinguish chromosomes that are mapped to the same result. In this case, two chromosomes with the same result may generate a new chromosome that is quite different from its parents, meaning that GA-based methods require more time to converge. Falkenauer [6] thus proposed the group genetic algorithm (GGA) to address this problem. The GGA has the same workflow as that of the GA, but uses different encoding schema and different operators. It has been demonstrated that the efficiency of the GGA is superior to that of the GA in some areas, especially with regard to grouping problems [4].

The present study thus proposes a GGA-based attribute clustering approach. In other words, the main goal of the proposed GGA-based approach is to divide attributes into K groups for classification. Attributes in the same group mean they have similar properties. Then, we can select attributes from each group, and gather them together for classification. In addition, in case a selected attribute A has too many missing values, we can select attribute B from the same group to replace A. Thus, based on GGA, we define chromosome representation, genetic operations, and fitness function (please see Section 3) for the proposed approach for solving the attribute clustering problem. Besides, the contributions of this paper are stated as follows: (1) a GGA-based attribute grouping approach is proposed for solving attribute clustering problem for classification; (2) by utilizing the derived attribute clustering results, when an attribute A has many missing values or less data instances, instead of the attribute A, an attribute B in same group can be selected for classification.

The rest of this paper is organized as follows. Some related studies are reviewed in Section 2. The proposed method based on the GGA for attribute clustering with some examples are also given to illustrate its use is described in Section 3. The experimental results are given and discussed in Section 4. Finally, the conclusions and suggestions for future work are stated in Section 5.

2. Review of related work

This chapter reviews some related research, covering topics such as feature selection, attribute clustering, attribute dependency measurements, and the GA used to address grouping problems.

2.1. Feature selection

Feature selection is an important pre-processing procedure in machine learning and data mining, especially when the learning process is executed on high-dimensional datasets. Dash et al. [5] defined feature selection for classification as finding the minimally sized subsets of features that could maintain the classification accuracy and the resulting class distribution. A good feature subset not only reduces the training time and I/O requirement, but also leads to a better understanding of the data and more accurate predictions. An input feature set usually has some features that are redundant or irrelevant to the objective. These features may cause the classifier to infer in the wrong direction for finding the result, and thus decrease the speed of the process. Irrelevant features cannot improve, or may even decrease, the performance of the target objective. Redundant features are relevant to the target objective, but can be replaced by other features. The purpose of feature selection is thus to find an appropriate subset of features that are relevant to the target concept. In other words, it is a procedure that removes irrelevant or redundant features to improve the efficiency of applications working with high-dimensional datasets. There are many GA-based approaches which have been proposed for feature selection [15,17]. In addition, some PSO- or ACO-based algorithms have also been proposed for the feature selection problems [13,21]. The main purpose of those approaches is to provide a set of selected attributes for classification. In our previous work, a GA-based algorithm has been proposed for attribute grouping [17], not only for feature selection.

2.2. Attribute dependency measurements

Dependency measurements are used to estimate the similarity between attributes. They were proposed by Han et al. [11] and Li et al. [14]. Hong and Liou used a dependency measure in their attribute clustering method based on feature selection approaches [10]. Attributes that provide a similar contribution to the classification have a high dependency on each other. Formally, given two attributes A_i and A_j, the relative degree of dependency of A_i with regard to A_j is denoted as $Dep(A_i, A_j)$, defined as formula (1):

$$Dep(A_i, A_j) = \frac{\left|\prod_{A_i}(U)\right|}{\left|\prod_{A_i \cup A_j}(U)\right|}, \tag{1}$$

where U is a set of training examples and $\prod_{A_i}(U)$ is the projection of U on attribute A_i. The degree of dependency degree is not symmetrical. The average of $Dep(A_i, A_j)$ and $Dep(A_j, A_i)$ is thus used to represent the similarity of attributes A_i and A_j.

2.3. Attribute clustering based on genetic algorithms

Hong and Wang [16] presented a GA-based clustering method for attribute clustering to find approximate feature subsets for classification. They first proposed an approach which considers the average classification accuracy and the cluster balance of the attribute clusters, represented by chromosomes, as the fitness evaluation criteria. The average classification accuracy is used to calculate all the possible feature subset combinations from the chromosome clustering results, and this is then used to evaluate the classification ability of the selected feature subset with regard to the given dataset. The other measure used in this approach is the cluster balance, which is used to help the GA clustering algorithm to find clusters with similar numbers of attributes. If the clustering result represented by a chromosome is more balanced, then the value is larger. The fitness function then combines the two measures together to get a good trade-off between accuracy and cluster balance.

2.4. Genetic algorithms and grouping problems

Many methods based on GA have been proposed for solving grouping problems [12], although some challenges remain for standard GA. The two main weaknesses of GA in this respect are described below.

First, the standard encoding scheme of GA is highly redundant with regard to grouping problems. Assume that there are N objects to be clustered into K clusters. Each chromosome may be represented as an N-gene sequence, with each gene being one of the K group symbols and representing the object of the gene belonging to that group. For example, assume that there is a chromosome, *ABBAC*, which represents five objects in three groups. The first and the fourth objects are in the group A, the second and third objects are in another group B, and the fifth object is in a singleton group C. This encoding scheme has $K!$ distinct chromosomes to represent the same grouping result, and the genetic operators cannot identify them. Hence the search space increases significantly because of

Download English Version:

https://daneshyari.com/en/article/495252

Download Persian Version:

https://daneshyari.com/article/495252

Daneshyari.com