Applied Soft Computing 12 (2012) 46–63

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

An enhanced classification method comprising a genetic algorithm, rough set theory and a modified PBMF-index function

Kuang Yu Huang*

Department of Information Management, Ling Tung University, #1 Ling Tung Road, Taichung City 408, Taiwan

ARTICLE INFO

Article history:
Received 4 January 2010
Received in revised form
22 September 2010
Accepted 18 September 2011
Available online 24 September 2011

Keywords:
Fuzzy C-means
Genetic algorithm
Rough set
GRP-index method
PBMF-index function
Pseudo-supervised classification method
Discretization

ABSTRACT

This study proposes a method, designated as the GRP-index method, for the classification of continuous value datasets in which the instances do not provide any class information and may be imprecise and uncertain. The proposed method discretizes the values of the individual attributes within the dataset and achieves both the optimal number of clusters and the optimal classification accuracy. The proposed method consists of a genetic algorithm (GA) and an FRP-index method. In the FRP-index method, the conditional and decision attribute values of the instances in the dataset are fuzzified and discretized using the Fuzzy C-means (FCM) method in accordance with the cluster vectors given by the GA specifying the number of clusters per attribute. Rough set (RS) theory is then applied to determine the lower and upper approximate sets associated with each cluster of the decision attribute. The accuracy of approximation of each cluster of the decision attribute is then computed as the cardinality ratio of the lower approximate sets to the upper approximate sets. Finally, the centroids of the lower approximate sets associated with each cluster of the decision attribute are determined by computing the mean conditional and decision attribute values of all the instances within the corresponding sets. The cluster centroids and accuracy of approximation are then processed by a modified form of the PBMF-index function, designated as the RP-index function, in order to determine the optimality of the discretization/classification results. In the event that the termination criteria are not satisfied, the GA modifies the initial population of cluster vectors and the FCM, RS and RP-index function procedures are repeated. The entire process is repeated iteratively until the termination criteria are satisfied. The maximum value of the RP cluster validity index is then identified, and the corresponding cluster vector is taken as the optimal classification result. The validity of the proposed approach is confirmed by cross validation, and by comparing the classification results obtained for a typical stock market dataset with those obtained by non-supervised and pseudo-supervised classification methods. The results show that the proposed GRP-index method not only has a better discretization performance than the considered methods, but also achieves a better accuracy of approximation, and therefore provides a more reliable basis for the extraction of decision-making rules.

1. Introduction

The method proposed in this study, designated as the GRP-index method, provides the means to solve the following problems for a dataset characterized by imprecision and uncertainty: (1) discretizing the continuous values of all the individual attributes within a dataset; (2) evaluating the optimality of the discretization results; (3) determining the optimal number of clusters per attribute; and (4) improving the classification accuracy of datasets characterized by uncertainty.

Classification is a supervised process in which new data instances with multiple attributes are grouped into relevant categories based on their class information. Classification provides an invaluable means of uncovering the implicit knowledge within a dataset. However, in implementing classification schemes, a problem commonly arises in clustering the data instances within the dataset, that is in "partitioning" the real-value attributes into discrete partitions. Therefore, to classify continuous value datasets lacking class information in a meaningful way, it is first necessary to accurately discretize the instances within the dataset.

Clustering techniques are of fundamental importance in many fields, and have therefore received extensive attention. Since the original introduction of fuzzy sets by Zadeh [1], numerous fuzzy set-based approaches have been proposed for modeling systems with uncertainties. For example, the Fuzzy C-means (FCM) method [2] assigns each instance within the dataset to one or more clusters with a certain degree of belonging such that the instances within each cluster are similar to one another. However, while this method has found widespread use in pattern recognition applications, it

* Tel.: +886 9 22621030.
E-mail address: kyhuang@mail.ltu.edu.tw

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.09.009

requires the number of clusters within the dataset to be known in advance. Unfortunately, in constructing many information systems, this information is not available *a priori*, and thus FCM is of only limited use in most practical partitioning applications. Thus, a method is required for accurately determining the optimal number of clusters per individual attribute within the dataset.

In general, the problem of evaluating the optimality of the clustering results obtained for a particular dataset is referred to as the cluster validity problem [3,4]. Many methods have been proposed for assessing the validity of the clustering results obtained using fuzzy clustering schemes [5]. Amongst these methods, traditional indices such as the partition coefficient [6,7] and classification entropy coefficient [8,9] are based simply on the membership values of the items within the dataset and are therefore computationally straightforward. However, recent studies have shown that the accuracy of a cluster validity index can be improved by considering not only the dataset itself, but also the matrix U used to partition the data [10–17]. In general, existing clustering methods cluster the dataset in accordance with the norms of the instances rather than the values of the individual attributes of the instances. However, in most real-world datasets, the instances are characterized by multiple attributes. That is, each attribute is an independent parameter of the instances within the dataset. As a result, the clustering results obtained using traditional clustering methods fail to take sufficient account of the complex interrelationships between the various attributes of the dataset. Thus, the GRP-index classification method proposed in this study incorporates an enhanced discretizing method, designated as the FRP-index method, which not only discretizes the values of all the individual attributes within the dataset, but also considers the accuracy of the corresponding classification results. In developing the proposed method, five major questions are addressed: (1) "does the optimal number of clusters per attribute within a dataset necessarily coincide with the number of clusters which ensures the optimal accuracy of approximation [18]?"; (2) "to what extent do discretizing schemes based on the data values of complex datasets capture the true partitions within the dataset"?; (3) "to what extent do the classification results obtained using attribute-based discretization schemes enable reliable decision-making rules to be derived?"; (4) "in adopting an attribute-based discretization approach, how can one determine the optimal number of clusters for each conditional and decision attribute?"; and (5) "in searching for the optimal number of clusters per attribute, how can one specify an appropriate search range?".

The literature contains many classifiers for automatic classification purposes, including decision-tree algorithms such as ID3 [19], rule-based algorithms such as CN2 [20], back-propagation networks [21], support vector machines [22,23], conformal predictors [24], Bayesian classifiers [25], and so forth. All of these schemes have their respective merits and have found widespread use in a diverse range of applications, including weather prediction, manufacturing process planning, medical diagnosis, and so on. However, such methods cannot deal effectively with datasets with no output values specified or datasets characterized by uncertain or missing information. Furthermore, to the best of the author's knowledge, no attempt has ever been made to integrate such classification algorithms with any form of cluster validity index function. Thus, there is no guarantee that the classification results obtained using these algorithms represent the optimal solution in terms of the number of clusters within the dataset or the accuracy of the classification results. Therefore, when classifying datasets with uncertain or missing information, it is preferable to utilize rough set (RS) theory for classification purposes, and to integrate the RS model with some form of iterative cluster generation/cluster index evaluation procedure such that the optimal discretizing solution can be obtained.

RS theory was first introduced more than twenty years ago [18] and has emerged as a powerful technique for the automatic classification of datasets [26] in a diverse range of fields, including machine learning, forecasting, knowledge acquisition, decision analysis, knowledge discovery, and pattern recognition. As the scale of business and scientific databases continues to increase, the use of RS theory to "mine" data repositories in order to extract reliable classification rules has become increasingly common [27–37]. In RS theory, the uncertain nature of the information within the system of interest is handled using the concept of lower and upper approximate sets. The lower approximate set contains all the instances within the system which can be unambiguously ascribed to a particular target set, and provides the information required to extract decision rules with which to classify new data arrivals. Meanwhile, the upper approximate set contains all the instances within the system which may possibly belong to a particular target set. Thus, the accuracy of approximation [18] can be quantified in terms of the cardinality ratio of the lower approximation set to the upper approximation set. However, the performance of RS models is fundamentally dependent upon the quality of the discretizing results to which they are applied. Therefore, in meeting the needs of decision-makers in analyzing complex datasets and deriving appropriate decision rules, it is necessary to integrate the RS method with some form of optimized discretization scheme. In the present study, this is achieved using an enhanced discretizing method, designated as the FRP-index method, based on FCM, RS theory and a modified form of the PBMF function. To discretize the values of the individual attributes within the dataset, some form of discretization method is required. Although many fuzzy clustering methods have been proposed to achieve this, the FRP-index method deliberately utilizes FCM since it is well established and widely used. Having processed the discretizing results obtained from FCM using a RS classification model, a modified form of the PBMF-index function, designated as the RP-index function, is used to measure the fuzzy distances between the instances and the centroids of the lower approximate sets associated with each cluster of the decision attribute.

In order to optimize the number of clusters per conditional and decision attribute, it is necessary to integrate the clustering mechanism with some form of optimization technique. Genetic algorithms (GAs) [38–40] have been successfully applied in a wide variety of optimization and classification-type problems in recent years. Accordingly, in the GRP-index method proposed in this study, the FRP-index method is integrated with a GA. Some researchers have proposed an indirect approach in which the elements of the GA chromosome represent the centers of the clusters, and the objective of the optimization process is to assign each instance to the closest cluster center [41,42]. However, in the GA used in the GRP-index method, the elements of the chromosome represent the number of clusters per corresponding conditional or decision attribute. Having initialized the GA population, each chromosome (cluster vector) is processed using the FRP-index method. Specifically, for each cluster vector, FCM is applied to discretize the continuous values of the individual conditional and decision attributes within the dataset. The partitioning results are then processed using a RS model in order to determine the upper and lower approximate sets, the accuracy of approximation, and the centroids of the lower approximate sets associated with each cluster of the decision attribute. The cluster centroids and accuracy of approximation are then input to the RP-index function in order to obtain the cluster validity index for the discretization/classification solution. The GA population is then evolved, and the FCM, RS and RP-index function procedures are repeated once again with the new set of cluster vectors. The entire process is repeated iteratively until the GA termination criteria are satisfied. The cluster validity indices associated with all of the cluster vectors