Applied Soft Computing 38 (2016) 176–189

Contents lists available at ScienceDirect

Applied Soft Computing

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

A multi-attribute decision-making model for the robust classification of multiple inputs and outputs datasets with uncertainty

Kuang Yu Huang [a], I-Hui Li [b,*]

[a] Department of Information Management, Ling Tung University Taichung, Taiwan
[b] Department of Information Networking and System Administration, Ling Tung University Taichung, Taiwan

ARTICLE INFO

Article history:
Received 17 April 2014
Received in revised form 6 July 2015
Accepted 5 September 2015
Available online 1 October 2015

Keywords:
Multi-criteria decision making
Multiple inputs and outputs datasets
Variable precision rough set
Fuzzy set theory
FVM-index method

ABSTRACT

Many multiple-criteria decision-making (MCDM) methods have been proposed for decision-making environments. However, the performance of these methods is degraded by the uncertainty and inaccuracy which characterizes most practical decision-making environments as a result of the inherent prejudices and preferences of the decision-makers or experts and an insufficient volume of multiple inputs and outputs (MIO) information. Accordingly, the present study proposes an enhanced MIO classification method to address these limitations of existing MCDM methods. The proposed MIO classification method designated as the FVM-index method integrates fuzzy set theory (**F**ST), variable precision rough set (**V**PRS) theory, and a modified cluster validity index (**M**CVI) function, and is designed specifically to filter out the uncertainty and inaccuracy inherent in the surveyed MIO real-valued dataset; thereby improving the classification performance. The effectiveness of the proposed approach is first demonstrated by comparing the MIO classification results obtained for three relating UCI datasets: (1) the original dataset; (2) a dataset with a large amount of inaccurate instances; and (3) an FVM-index filtered dataset extracted from the original dataset using a statistical approach. Then, the validity of the proposed approach is illustrated by using an Augmented Reality product design and a hospital related datasets. The results confirm that the proposed FVM-index method provides a good classification performance even in the presence of inaccuracy and uncertainty. As a result, it provides a robust approach for the extraction of reliable decision-making rules.

1. Introduction

As computer technology and information-gathering methods have advanced in recent decades, the size and complexity of modern day data repositories have increased accordingly. Data mining technique is widely used to extract information or knowledge from large volumes of data and can be performed using a variety of different tools, including association, classification, clustering, prediction, sequential patterns, and similar time sequences [1–4]. Among these methodologies, classification and regression, referring to the type of predictive modeling, build a model that will permit the value of one response variable (also called the dependent, or target variable) to be predicted from the known values of other explanatory variables (also called the independent or predictor variables) [1,2]. Decision-makers typically use machine learning or data mining techniques, such as classification or regression methods, to discover the useful knowledge within a dataset. Thus, in enabling decision-makers to make reliable decisions when faced with such large quantities of data, highly efficient and robust data classification/regression methods are required. For the case where the decision involves a single criterion (e.g., the cost), the decision-making process is relatively straightforward. However, in many practical cases, the decision-making process involves a number of different criteria, e.g., price, size, power consumption, functionality, and so on; all of these have a different weighting and must be jointly considered. In such a case, more sophisticated multi-criteria decision-making methods are required to support the decision-maker in evaluating the competing alternatives. Consequently, the problem of multi-criteria decision-making (MCDM), also referred to as multi-criteria decision-analysis (MCDA) has attracted increasing attention in recent decades. It is formulated by considering a set of alternatives and a set of criteria $G = (g_1, g_2, \ldots, g_m)$, where n is the number of alternatives and m is the number of criteria. In most approaches, the multi criteria

* Corresponding author at: 1, Ling Tung Rd., Taichung 408, Taiwan. Tel.: +886 4 23892088; fax: +886 4 22895293.
E-mail addresses: kyhuang@teamail.ltu.edu.tw (K.Y. Huang), sanity@teamail.ltu.edu.tw (I-H. Li).

http://dx.doi.org/10.1016/j.asoc.2015.09.015
1568-4946/

evaluation for an alternative a_i is presented by the vector $g(a_i)=(g_1(a_i), g_2(a_i), ..., g_m(a_i))$, where $g_j(a_i)$ is the performance of the alternative $a_i \in A$ on criterion $g_j(a_i)$ [5]. For example, the use of Data Envelopment Analysis (DEA), which is a tool for Multiple Criteria Decision Making (MCDM) and is applied to handle the multiple inputs and outputs (MIO) process [6], would be investigated for defining the maximizing criteria as *outputs* and the minimizing criteria as *inputs* to measure productive efficiency of alternatives (decision making units) [5,7]. Broadly speaking, MCDM methods can be classified in terms of the nature of the problem they considered, namely continuous or discrete [8]. Methods of the former type are referred to as multiple-objective decision-making (MODM) methods, while those of the latter type are referred to as multiple-attribute decision-making (MADM) methods [9]. The basic objective of MODM methods is to obtain the alternatives of decision-making through mathematical programming model. By contrast, the aim of MADM methods [10–12] is to find the optimal solution among a set of limited alternatives which consist of multiple attributes by evaluating the corresponding importance of each attribute.

Fig. 1, reproduced from [13,14] illustrates the basic concepts and techniques pertinent to a complex MCDM problem-solving environment. As shown in the figure, the MCDM process involves three stages, namely (1) Data Processing/Statistical and Multivariate Analysis, (2) Planning/Designing, and (3) Evaluating/Choosing. Although the first two stages are relatively straightforward for most decision-making problems, the third stage, namely Evaluating/Choosing, is often far more problematic. For example, in evaluating and choosing between competing alternatives, the MCDM model must minimize the effects of the inherent preferences and prejudices of the decision-maker or experts in order to achieve an objective outcome. Moreover, the relations among the various attributes in the dataset are extremely complicated for most multi-input (conditional attributes and independent variables) and multi-output (decision attributes and dependent variables) information systems. In the end, many large real-world datasets are characterized by inaccuracy and uncertainty. (Note that the term *inaccuracy* refers to the case where the values of same multi-independent variables correspond to those of different multi-dependent variables. Similarly, the term *uncertainty* refers to a state of having limited knowledge where it is impossible to exactly describe the existing state.) Survey errors is the central focus of a study by Groves [15,16], in which the author found that non-response, sampling, interviewer effects, mode effects, various other types of measurement errors, and processing errors would affect the usefulness of surveys. Furthermore, Wolfson et al. and other researchers [17–19] demonstrated that in applications requiring interaction with the physical world, data uncertainty is an inherent property due to measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. Accordingly, Huang and Lin [20] proposed a VPRS (variable precision rough set)-based index approach to filter out the uncertainty inherent in the datasets with mixed attribute type (numerical, categorical). One major limitation of this approach is that it aimed at classifying the single-dependent-variable datasets.

Little is known about how to filter out uncertain instances from the original dataset when the existing MCDM methods are applied, even though a large number of studies have been made on proposed MCDM methods for decision-making environments. Thus, in facilitating reliable decisions, MCDM models must be capable of filtering out such inaccuracies and uncertainties. However, due to a lack of training samples, most of these MCDM models are unable to filter out the inaccurate or uncertain instances; therefore throwing doubt on the validity of the identified decision rules. It has been argued, by Triantaphyllou [21] and others, that although this MCDM is very relevant in practice, there are few methods available and their quality is hard to determine, i.e., the difficulty that always occurs when trying to compare decision methods and choose the best one is that a paradox is reached. As Lin [10] noted in his review of state-of-the-art in fuzzy decision-making methods, "none of the fuzzy decision-making methods examined in this study is perfectly effective in terms of both evaluative criteria. The methods are less accurate when the decision-making problems become more complex." In addition, the relationship among attributes becomes more complex and thus the corresponding computation cost increases as the number of attributes increases, it will be limited to find the efficient and effective resolution using the conventional MADM approaches when faced with such MIO datasets with *inaccuracy* and *uncertainty* [22–24].

Furthermore, there are several studies in literature that used discretization approach as preprocessing for most inductive learning methods [25–28], such as VPRS theory. Accordingly, the present study proposes an enhanced MIO classification method designated as the FVM-index method, which integrates fuzzy set theory (FST), VPRS theory, and a cluster validity index (CVI) function in order to discretize each real-valued attribute, to filter out the inherent uncertainty and inaccuracy in the surveyed datasets and to obtain a more reliable and robust MIO classification outcome as a result. The validity of the proposed approach is demonstrated using some datasets for illustration purposes.

The remainder of this paper is organized as follows. Section 2 presents the fundamental principles of FST theory, VPRS theory, and the proposed CVI function. Section 3 describes the integration of these concepts to form the proposed FVM-index method. Section 4 evaluates and discusses the MIO classification performance of the proposed method. Finally, Section 5 provides some brief concluding remarks and indicates the intended direction of future research.

2. Review of related methodologies

2.1. VPRS theory

VPRS theory is a generalized form of rough set (RS) theory which inherits all of the basic mathematical properties of the original RS model. The RS model assumes that the universe under consideration is known and that all of the conclusions derived from the model are applicable only to this universe. However, in practice, there is evidence to suggest that only a smaller set of examples suffices to generalize the conclusions drawn from a larger population [29,30]. In contrast to the original RS model, the VPRS model provides a more robust classification performance in the case of uncertainty. For example, provided that the majority of the available data can be correctly classified, even partially correct classification rules still provide valuable trend information regarding future test cases. VPRS operates on what may be represented as a knowledge-represented system or an information system [31]. The basic principles and notations of information systems (S), and the application of VPRS theory to the processing of such systems, are described in the following sub-sections.

2.1.1. β-lower and β-upper approximate sets

For a given dataset, the records which are indistinguishable from one another when evaluated using a specific subset of all the system attributes can be defined using an equivalence or indiscernibility relationship. In VPRS theory, this indiscernibility concept is operated using approximate sets. A typical information system has the form $S=(U, A, V_q, f_q)$, where U is a non-empty finite set of records, A is a non-empty finite set of attributes describing these records, $X \subseteq U$, and $R \subseteq A$. Generally speaking, the attributes in set A can be partitioned into a set of conditional attributes $C \neq \phi$ and