A Comparative Study Based on Rough Set and Classification Via Clustering Approaches to Handle Incomplete Data to Predict Learning Styles

Hemant Rana, Indira Gandhi National Open University, New Delhi, India
Manohar Lal, Indira Gandhi National Open University, New Delhi, India

ABSTRACT

Handling of missing attribute values are a big challenge for data analysis. For handling this type of problems, there are some well known approaches, including Rough Set Theory (RST) and classification via clustering. In the work reported here, RSES (Rough Set Exploration System) one of the tools based on RST approach, and WEKA (Waikato Environment for Knowledge Analysis), a data mining tool—based on classification via clustering—are used for predicting learning styles from given data, which possibly has missing values. The results of the experiments using the tools show that the problem of missing attribute values is better handled by RST approach as compared to the classification via clustering approach. Further, in respect of missing values, RSES yields better decision rules, if the missing values are simply ignored than the rules obtained by assigning some values in place of missing attribute values.

KEYWORDS
Classification Via Clustering, Data Mining, Knowledge Discovery, Missing Attribute Values, Reduct, Rough Set Theory, RSES, Rule Generation, WEKA

1. INTRODUCTION

In many real life applications, for example, in the case of large enrolments of students (as is the case with Indira Gandhi National Open University (IGNOU), New Delhi, India, which enrolls around 3 million students), missing data in respect of values of salient features, is a common occurrence. The problem necessitates the study of methods for handling information with missing attribute values. The reason for missing attribute values in the data set may be lost values and don’t care condition values (Grzymala-Busse, 2000; Grzymala-Busse, 2004). One of the reasons for missing values may be that the students may not know or may not understand the query properly. Sometimes a student forgets to answer the query or refuses to answer queries. Also, in some cases, it may be answered, but later on gets mistakenly erased by an operator (Grzymala-Busse & Zdzislaw, 2013). Such a missing value will be called lost value. For example, some irrelevant attribute values are not recorded like
age, sex of the student while designing timetable. Such missing attribute values will be called don’t care condition values.

For handling this type of problems, there are some well-known approaches, including RST and classification via clustering. In the work reported here, RSES (Rough Set Exploration System) one of the tools based on RST approach, and WEKA (Waikato Environment for Knowledge Analysis), a data mining tool—based on classification via clustering — are used for predicting learning styles from given data, which possibly has missing values.

1.1. Rough Set Theory and RSES

As RST is used as one of the approaches to handling missing attribute values to provide better decision rules, first briefly discussed RST. RST was first proposed by Z. Pawlak in early 1980’s (Pawlak, 1982). RST is a mathematical approach to deal with uncertainty, incompleteness, and vagueness in data (Pawlak, 1982). The fundamental idea of RST is that values, though distinct in data, may/do not differ in decision making. For example, for designing timetable of a school, if 40 students study the same courses of the same level, then these 40 students may be clubbed together for the purpose of designing the timetable. Then, the differences in heights, marks scored in the previous class, etc. of students do not count for this decision-making. A Decision table is an important tool in RST. In order to make decisions based on the large dataset, a decision table is used consisting of two types of attributes, viz. (i) attributes for which values are expected to be given and (ii) attributes, the values of which are to be determined. The two types of attributes are respectively classified as condition attributes and decision attributes. The each of the condition attributes is an independent variable and each of the decision attributes is a dependent variable of the decision table. The decision table with some unspecified values of condition attributes is called an incomplete decision table. In an incomplete decision table, the missing/lost values will be represented by ‘?’ and don’t care condition values will be represented by ‘*’. In RST approach, some rows/columns, containing entries having missing/don’t care values, may be ignored so that the remaining table is complete before initiating further decision-making process. To complete an incomplete decision table in this manner, some well-known methods are later discussed in detail.

The idea of RST is to find indiscernibility relations in the complete decision table. The complete decision table exhibits indiscernibility relations using some equivalence relation of condition attributes. Each of equivalence classes of indiscernibility relations is called an elementary set. For example, in the case of designing timetable for students of a school, all attributes except subjects studied and standard, may be ignored. This leads to the indiscernibility relation under which all students with the same subjects and standard become indiscernible, and hence, all of them may be treated as a sort of meta-student—student representative. On the basis of indiscernibility relation, lower and upper approximations are defined for the complete decision table. The set of objects which surely belong to the set is called Lower approximation and the set of objects which have some possibility of belonging to the set is called Upper approximation (Pawlak, 1982; Pawlak, 2002). The boundary region cases of the set are present in upper approximation and are not present in lower approximation (Pawlak, 2007). A set is said to be rough set if the boundary region is non-empty (Pawlak, 1991).

An incomplete decision table is described by a characteristic relation, in a similar way as the complete decision table is described by indiscernibility relation (Grzymala-Busse, 2004; Grzymala-Busse, 2006; Grzymala-Busse, 2007). Also, each of elementary sets is replaced by a characteristic set. The characteristic set is the intersection of blocks of attribute-value pairs for all attributes for specified cases. Similar to indiscernibility relation, defined a characteristic relation (Grzymala-Busse, 2004). The characteristic relation is reflexive. Once, the indiscernibility relation is fixed and the concept is given, lower and upper approximations are unique. But for an incomplete decision table, lower and upper approximations are different. The incomplete decision table must be completed, in the manner discussed above, before generating reduct and core set of the complete decision table. Complete decision table gives us the concept of consistency and inconsistency. A decision table is
18 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:


www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Towards Collaborative Multidimensional Query Recommendation with Triadic Association Rules

www.igi-global.com/article/towards-collaborative-multidimensional-query-recommendation-with-triadic-association-rules/133849?camid=4v1a
Intelligent DSS Under Verbal Decision Analysis

Proportional Integral Loop Shaping Control Design With Particle Swarm Optimization Tuning

Strategy, Decision Making, and Cognition: An Integrative Perspective