Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 03, March : 2022

Two-Phase Rule Induction fromIncomplete Data

1*Biswanath Biswal
Nalanda Institute of Technology, Bhubaneswar
Dept. of Basic Science & Humanities
E-mail ID: biswanathbiswal@thenalanda.com

E-mail: bipinbihari05@gmail.com

Abstract. A framework for learning new rules from incomplete data isintroduced so that the user can easily identify attributes with or without rule values. The rule defines two measurement levels. A two-step rule induction algorithm is presented. Instead of filling in missing attribute values before or during a rule call, we split rule induction into two steps. In the first step, rules and subrules are induced based on non-missing values. In the second step, partial rules are modified and refined by filling in some missing values. Such rules faithfully reflect knowledge embedded in incomplete data. The study not only provides new insights into rule inference from incomplete data, but also offers a practical solution.

Keywords: Missing attribute values, Filled-in values, Two-phase rule induction.

1 Introduction

A major focus of machine learning and data mining research is to extract useful knowledge from a large amount of data. For such a purpose, the integrality of data is very important. However, real-world data sets frequently contain missing values, i.e., attribute values of objects may be unknown or missing [14]. To deal with missing values, many methods have been proposed [2,3,4,5,6,8,1,10]. They may be broadly classified into three categories. The first category mainly focuses on transforming incomplete data into complete data by filling in the missing values. Rules are induced from the completed data. The

second category fills in the missing values during the process of rule induction. The third category considers tolerance relations or similarity relations defined based on missing values [5,6,10].

The third category may be considered as a special case of the first category. In fact, one may first fill in the missing values and then derive a similarity rela- tion. One disadvantage of filling in attribute values before the learning process is that the learning goals and algorithms are not directly considered. Another disadvantage is that all missing values are filled in although some of them are not necessary. Since rules normally contain only a subset of all (*attribute, value*) pairs, we do not need to fill in all missing values. To avoid those problems, one can combine the processes of filling in the missing values and learning together. Algorithms like C4.5 [1] fill in missing attribute values according to some special learning goals. In addition, missing values are filled in only when the demand arises in the learning process.

Two fundamental important problems still remain in the existing algorithms for inducing rules from incomplete data. One is the use of the filled-in values. Any method of filling in missing values is based on certain assumption about the data, which may not be valid. However, rule learning algorithms treat filled-in values as if they are the original values. This may result in rules having more number of filled-in values and less number of the original values. That is, we may obtain rules that more fit the filled-in values. Although these rules may have good statistical characteristics, they are in fact not reliable. The other issue is the use of induced rules by users. Without a clear distinction between the filled-in values and the original values, a user may find it difficult to interpret and apply rules. In fact, a user may misuse a rule by putting more weights on filled-in values. Solutions for those two problems require new ideas and methodologies.

The objective of this paper is to propose a new framework of

rule induction by separating filled-in and the original values in both the learning process and the induced rules. A two-phase model is suggested. The first phase induces partial rules based only on the original values. Rules induced in the first phase will be associated with a quantitative measures such as confidence, coverage and generality [11,13]. In the second phase of rule induction, filled-in values are taken into account so that the performance of the rules may improve. A new form of rules is introduced, in which known attribute values and filled-in attribute values are used. A user can easily identify attributes with or without missing values in rules.

2 A Framework of Two-Phase Rule Induction

The two-phase framework of rule induction from incomplete data uses a new form of rule involving both the known attribute values and filled-in values. In the first phase of rule induction, rules is induced only based on known attribute values. Many approaches of machine learning and data mining methods, such as concept learning, decision tree and rough set-based learning, can be used. One may associate certain quantitative measures to express the strength of a rule. In the second phase of rule induction, missing attribute values are filled in according to a certain method, and rules are induced based on the repaired data. The performance of the new rule induced in the second phase should be superior (higher) than that of rule induced in the first phase.

The main ideas of two-phase rule induction can be illustrated by a simple example. Suppose r is a rule induced in the first phase based on known attribute values with a measure α_1: