Information Sciences 322 (2015) 197–222

Contents lists available at ScienceDirect

Information Sciences

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

Consistency of incomplete data

Patrick G. Clark [a], Jerzy W. Grzymala-Busse [a,b,*], Wojciech Rzasa [c]

[a] *Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence 66045-7621, USA*
[b] *Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszow, Poland*
[c] *Interdisciplinary Centre for Computational Modeling, Rzeszow University, 35-310 Rzeszow, Poland*

ARTICLE INFO

Article history:
Received 7 December 2014
Received in revised form 20 May 2015
Accepted 11 June 2015
Available online 22 June 2015

Keywords:
Consistency
Incomplete data
Missing attribute value
Rough set theory
Probabilistic approximation

ABSTRACT

Consistency is well-known for completely specified data sets. A specified data set is defined as consistent when any pair of cases with the same attribute values belongs to the same concept. In this paper we generalize the definition of consistency for incomplete data sets using rough set theory. We discuss two types of missing attribute values: lost values and "do not care" conditions. For incomplete data sets there exist three definitions of approximations: singleton, subset and concept. Any approximation is lower or upper, so we may define six types of consistencies. We show that two pairs of such consistencies are equivalent, hence there are only four distinct consistencies of incomplete data. Additionally, we discuss probabilistic approximations and study properties of corresponding consistencies. We illustrate the idea of consistency for incomplete data sets using experiments on many incomplete data sets derived from eight benchmark data sets.

1. Introduction

It is important to recognize if a complete, fully specified data set is consistent. A consistent data set is one that does not contain conflicting cases. Conflicting cases are those cases for which all attribute values are the same yet the decision values are different. Such cases belong to different concepts (or classes). Consistency of data sets is well described using rough set theory. A complete data set is consistent if for any concept X its lower approximation is equal to X or, equivalently, if its upper approximation is equal to X.

Consistency of a data set is one of the main ideas of machine learning in general [6,20,31] and one of the fundamental ideas of rough set theory in particular [22,24,25,28,34]. Some preliminary ideas related to consistency of incomplete data were discussed in [19,27]. In both papers only incomplete data with one interpretation of missing attribute values ("do not care" conditions) were considered. The main objective was to find attribute reductions. Additionally, in [19] only singleton approximations were concerned, in [27] approximations were defined using maximal consistent blocks. In [16] an incomplete data set was defined as inconsistent the same way as in [19,27]. In [5] consistent and inconsistent covering decision systems were presented. For incomplete data sets, an idea of consistent objects was introduced, and a non-invasive imputation method was presented in [7]. Similarly, some imputation methods, taking into account inconsistency, were discussed in [2]. Consistency of data based on fuzzy set theory was presented in [35]. An idea of variable consistency, for complete data sets, was introduced in [37]. A variable consistency model of dominance-based rough set approach was discussed in many papers, see, e.g., [1,8,29].

[*] Corresponding author at: Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence 66045-7621, USA.
E-mail addresses: patrick.g.clark@gmail.com (P.G. Clark), jerzy@ku.edu (J.W. Grzymala-Busse), wrzasa@ur.edu.pl (W. Rzasa).

http://dx.doi.org/10.1016/j.ins.2015.06.011

An important problem is to decide when incomplete data sets are consistent. Methods for computing reducts of incomplete data are simpler for consistent data [19,27].

In this paper, for the first time, we study properties of consistency for incomplete data sets including two interpretations of missing attribute values, three definitions of non-probabilistic approximations and three definitions of probabilistic approximations as well. For incomplete data sets there exist many definitions of approximations. In this paper, like in [4,12], we use three types of lower and upper approximations (singleton, subset and concept). The results of [4] demonstrate that when measuring performance in terms of a ten-fold cross validation error rate, all three kinds of approximations do not differ significantly (5% significance level, two-tail test). However, for a specific data set, error rates for different kinds of approximation used in data mining may differ drastically [4]. Therefore, for any data set, all three approximations should be considered and the best one should be selected for data mining. It justifies use of all three kinds of approximations in our experiments.

Theoretically, there exist six types of consistencies for incomplete data sets: singleton lower and upper consistent, subset lower and upper consistent, and concept lower and upper consistent. In this paper we show that a concept X is singleton lower consistent if and only if it is concept upper consistent and that X is subset lower consistent if and only if it is concept lower consistent. Thus, we show that there exist four distinct types of consistencies of incomplete data, represented by singleton lower consistency, singleton upper consistency, subset lower consistency and subset upper consistency.

Additionally, we distinguish between two interpretations of missing attribute values: lost values and "do not care" conditions. Lost values are defined as those attribute values that were erased from or not included in the data set. During data mining we induce rule set from given (specified) attribute values. "Do not care" conditions are the result of a refusal to answer a question. This may happen because the question was embarrassing or inconvenient: for example, some people refuse to reveal their salary in response to a questionnaire. Another possibility is that the question appears irrelevant: for example, patient responders doubt that hair color is relevant to a particular disease.

Lower and upper approximations of all three types: singleton, subset and concept may be generalized to probabilistic approximations. Both lower and upper approximations are special cases of probabilistic approximations. Lower approximations are probabilistic approximations with the probability equal to one, upper approximations are associated with the positive smallest possible probability.

Two possible types of consistencies are defined using probabilistic approximations. The first type occurs if there exists some probability for which the corresponding approximation of the type singleton, subset or concept is equal to the concept. In the second type of consistency, for any probability the corresponding approximations are all equal to the concept. The second type of consistency is called strong. We show that a concept X is singleton strongly consistent if and only if it is subset strongly consistent. Additionally, if the concept is subset strongly consistent then it is concept strongly consistent. For special kinds of data sets, with only "do not care" conditions, singleton, subset and concept strong consistencies are equivalent.

We conducted many novel experiments on incomplete data sets. The only similar experiments were reported in [19,27]. In [19,27] the some kinds of reducts were reported, using only one type of approximation and one type of missing attribute values. In [2] the number of inconsistent meta-objects after imputation of missing attribute values was identified, however, no approximations were used at all. As follows from our experiments, some benchmark data sets from the Repository of the University of California at Irvine are singleton, subset or concept consistent. Some preliminary results of these experiments were presented in [3].

2. Complete data sets

Our basic assumption is that the data sets are presented in the form of a *decision table*. An example of a decision table is shown in Table 1. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases is denoted by U. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7\}$. Some variables are called *attributes* while one selected variable is called a *decision* and is denoted by d. The set of all attributes will be denoted by A. In Table 1, $A = \{Temperature, Headache, Cough\}$ and $d = Flu$. For an attribute a and case $x, a(x)$ denotes the value of the attribute a for case x. For example, *Temperature* $(1) =$ *normal*.

Table 1
A complete decision table.

Case	Attributes			Decision
	Temperature	Headache	Cough	Flu
1	normal	yes	yes	no
2	normal	yes	no	no
3	high	yes	yes	no
4	normal	no	no	yes
5	normal	no	no	yes
6	high	yes	yes	yes
7	very-high	no	yes	yes