Information Sciences 280 (2014) 368–384

Contents lists available at ScienceDirect

Information Sciences

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

Mining incomplete data with singleton, subset and concept probabilistic approximations

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, KS 66045-7621, USA
[b] Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszow, Poland
[c] Department of Computer Science, Rzeszow University, 35–310 Rzeszow, Poland

ARTICLE INFO

Article history:
Received 23 January 2013
Received in revised form 28 March 2014
Accepted 2 May 2014
Available online 14 May 2014

Keywords:
Probabilistic approximations
Extensions of probabilistic approximations
Singleton, subset and concept probabilistic approximations
Incomplete data

ABSTRACT

Rough set theory provides a very useful idea of lower and upper approximations for inconsistent data. For incomplete data these approximations are not unique. In this paper we investigate properties of three well-known generalizations of approximations: singleton, subset and concept. These approximations were recently further generalized as to include an additional parameter α, interpreted as a probability. In this paper we report novel properties of singleton, subset and concept probabilistic approximations. Additionally, we validated such approximations experimentally. Our main objective was to test which of the singleton, subset and concept probabilistic approximations are the most useful for data mining. Our conclusion is that, for a given incomplete data set, all three approaches should be applied and the best approach should be selected as a result of ten-fold cross validation. Finally, we conducted experiments on complexity of rule sets and the total number of singleton, subset and concept approximations.

1. Introduction

Lower and upper approximations, defined for complete data (without any missing attribute values), are fundamental concepts of rough set theory [41–43]. In the real world many data sets are incomplete. In this paper we will discuss two interpretations of missing attribute values: lost values and "do not care" conditions. The former interpretation is used when the original attribute value was erased or was—mistakenly—not entered. In this case we should reason from existing data. In the later interpretation, all attribute values may be used to replace a missing attribute value. This interpretation corresponds to a refusal to answer a question, while all possible answers may apply. For example, one of the attributes is hair color and the concept is a set of patients sick with the flu. A patient may refuse to tell hair color since it seems to be irrelevant. If we want to use a "do not care" interpretation of a missing attribute value, all possible hair colors will be used for further analysis.

Until recently, rough set theory was enhanced by probabilistic reasoning in two different ways. The first possibility was studying probabilistic approximations, depending on an additional parameter α interpreted as a probability. These approximations were based on an equivalence relation. Typical representatives of such research are: Variable Precision Rough Sets, Bayesian Models and Decision-Theoretic Rough Set Models [6,29,44–47,57,58,63,64]. An approach based on dominance relation and called Variable Consistency Rough Set was presented in [1,7]. On the other hand, a generalization of the rough

* Corresponding author at: Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045-7621, USA. Tel.: +1 785 864 4488; fax: +1 785 864 3226.
E-mail addresses: patrick.g.clark@gmail.com (P.G. Clark), jerzy@ku.edu (J.W. Grzymala-Busse), wrzasa@univ.rzeszow.pl (W. Rzasa).

http://dx.doi.org/10.1016/j.ins.2014.05.007
0020-0255/

set based on applying an arbitrary binary relation (or covering) instead of the equivalence relation (or partition) was studied in [38,48,54–56,61,62]. In this approach approximations under consideration were standard, i.e., lower or upper.

Research on probabilistic approximations based on an arbitrary binary relation and not restricted only to lower and upper approximations, i.e., defined for any value of the parameter $0 < \alpha \leqslant 1$, was initiated in [17]. First results on practical usefulness of such probabilistic approximations were published in [2].

A probabilistic approximation, depending on a parameter (probability) α, represents the entire spectrum of approximations, with the parameter value between 0 and 1. Standard lower and upper approximations, extremes of the spectrum, are included in the spectrum. Any probabilistic approximation, with $0 < \alpha \leqslant 1$, potentially may be more useful for mining data than standard approximations. As follows from our experiments, with changing the parameter alpha, the error rate, evaluated by ten-fold cross validation, varies with an unpredictable rate. Experiments are necessary to tell what the optimal value of the parameter alpha is.

Note that an idea of a probabilistic rule is well-known. Usually it is based on supplying a rule with an additional information about the conditional probability of a concept described by the rule given the rule domain, see, e.g., [5,30,52,59,60]. A number of measures were proposed to be associated with a rule [5,30,40,49], e.g., weighted information gain [40]. In [19] changing a rule strength and in [59,60] calibrating probabilities were discussed. A review of methods handling missing attribute values was presented in [53].

Our approach to rule induction from data sets with missing attribute values is unique due to computing the singleton, subset or concept probabilistic approximations, depending on the parameter α, for every concept. The best value of α should be selected by some kind of validation techniques, e.g., by ten-fold cross validation. Similar approaches to rule induction were applied in [4,34,35,39]. However, in all of these papers only standard approximations were used.

The idea of non-probabilistic extensions of approximations for incomplete data, i.e., singleton, subset and concept approximations, together with a characteristic relation, was introduced in [10–12]. An experimental comparison of these three types of approximations was presented in [20,26,27]. As it was shown in [22], some probabilistic approaches to missing attribute values, such as Most Common Value for symbolic attributes and Average Value for numerical attributes and Concept Most Common Value for symbolic attributes and Concept Average for numerical attributes, are highly successful. These methods and rough-set approaches to missing attribute values, using standard lower and upper approximations, were published in [15,16,21]. Probabilistic approaches were either worse or not better than rough set approaches. Additionally, the same probabilistic approaches were compared with probabilistic approximations for six data sets with many missing attribute values in [3]. Rough set approaches were better for five data sets, for one data set probabilistic approach was more successful. Singleton, subset and concept approximations were generalized to probabilistic approximations in [17]. In this paper we present novel theoretical properties of singleton, subset and concept probabilistic approximations, and we present results of experimental validation of usefulness of such approximations.

With three different probabilistic approximations: singleton, subset and concept, an intriguing question is which option should be used in the practice of mining incomplete data. Thus, our main objective was to test which of the singleton, subset and concept probabilistic approximations are the most useful for data mining. Our conclusion is that, for a given incomplete data set, all three approaches should be applied and the best approach should be selected as a result of ten-fold cross validation.

Additionally, we conducted experiments on complexity of rule sets: the total number of rules and conditions. In general, rule sets induced from the data sets with "do not care" conditions are simpler than rule sets induced from the data sets with lost values.

In yet another series of experiments we recorded the total number of singleton, subset and concept approximations for data sets with lost values and "do not care" conditions. The total number of any type of approximations is always smaller for data sets with lost values.

In the next section we discuss the main tools of our approach: attribute–value blocks, characteristic sets, and a characteristic relation. Then we present an idea of definability. In Section 4 we study fundamental theoretical properties of non-probabilistic approximations, probabilistic approximations, and probabilistic approximations for complete data. In Section 5 we describe how our experiments were conducted. In Conclusions we summarize theoretical properties and results of experiments.

2. Attribute–value pair blocks

We assume that the input 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 will be denoted by U. In Table 1, $U = \{1,2,3,4,5,6,7,8\}$. 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, $A = \{Wind, Humidity, Temperature\}$ and $d = Trip$.

An important tool to analyze data sets is a *block of an attribute–value pair*. Let (a, v) be an attribute–value pair. For *complete* decision tables, i.e., decision tables in which every attribute value is specified, a block of (a, v), denoted by $[(a, v)]$, is the set of all cases x for which $a(x) = v$, where $a(x)$ denotes the value of the attribute a for the case x. For incomplete decision tables the definition of a block of an attribute–value pair is modified [10–12].