Information Sciences 453 (2018) 66–79

Contents lists available at ScienceDirect

Information Sciences

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

Characteristic sets and generalized maximal consistent blocks in mining incomplete data

Patrick G. Clark[a], Cheng Gao[a], Jerzy W. Grzymala-Busse[a,b,*], Teresa Mroczek[b]

[a] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
[b] Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Rzeszow 35-225, Poland

ARTICLE INFO

Article history:
Received 8 December 2017
Revised 4 March 2018
Accepted 3 April 2018
Available online 4 April 2018

Keywords:
Incomplete data mining
Characteristic sets
Maximal consistent blocks
Rough set theory
Probabilistic approximations

ABSTRACT

Mining incomplete data using approximations based on characteristic sets is a well-established technique. It is applicable to incomplete data sets with a few interpretations of missing attribute values, e.g., lost values and "do not care" conditions. On the other hand, maximal consistent blocks were introduced for incomplete data sets with only "do not care" conditions, using only lower and upper approximations. In this paper we introduce an extension of the maximal consistent blocks to incomplete data sets with any interpretation of missing attribute values and with probabilistic approximations. We prove new results on probabilistic approximations based on generalized maximal consistent blocks. Additionally, we present results of experiments on mining incomplete data using both characteristic sets and maximal consistent blocks and using two interpretations of missing attribute values: lost values and "do not care" conditions. We show that there is some evidence that the best approach is using middle probabilistic approximations based on characteristic sets or on maximal consistent blocks.

1. Introduction

We consider incomplete data sets, using two interpretations of missing attribute values: lost values and "do not care" conditions [1]. A lost value is interpreted as a value that we do not know since it was erased or not inserted into the data set. Rules are induced from existing, specified attribute values. A "do not care" condition is interpreted as an arbitrary value from the attribute domain. For example, if an attribute is the hair color, and possible values are blond, dark and red, a "do not care" condition is interpreted as any of these three colors.

For incomplete data sets special kinds of approximations: Singleton, subset and concept should be used [1]. In this paper we consider probabilistic approximations, an extension of lower and upper approximations. Such approximations are defined using a probability denoted by α. If $\alpha = 1$, the probabilistic approximation is lower; if α is a positive number, slightly greater than 0, the probabilistic approximation is upper. Such approximations were usually used for completely specified data sets [2–10]. Probabilistic approximations were extended to incomplete data sets in [11]. First experimental results on such approximations were reported in [12,13].

Maximal consistent blocks were introduced for incomplete data sets with only "do not care" conditions, using only lower and upper approximations [14]. Algorithms for computing maximal consistent blocks for data sets with "do not care" con-

* Corresponding author at: Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA.
E-mail addresses: cheng.gao@ku.edu (C. Gao), jerzy@ku.edu (J.W. Grzymala-Busse), tmroczek@wsiz.rzeszow.pl (T. Mroczek).

https://doi.org/10.1016/j.ins.2018.04.025
0020-0255/

Table 1
An incomplete data set.

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	High	?	yes	yes
2	*	yes	no	yes
3	*	no	*	yes
4	Normal	?	yes	yes
5	Normal	no	no	no
6	*	no	yes	no
7	Normal	*	?	no
8	High	*	no	no

ditions were presented in [15–17]. The main objective of this paper is to extend the theory of maximal consistent blocks to arbitrary interpretation of missing attribute values using probabilistic approximations, an extension of ordinary lower and upper approximations.

In [14] a new kind of lower and upper approximations, based on maximal consistent blocks, was defined. In this paper we define three kinds of probabilistic approximations: singleton, subset and concept, all based on generalized maximal consistent blocks. We prove that all three kinds of probabilistic approximations, based on maximal consistent blocks, are equal to each other.

Given the family $\mathscr{C}(B)$ of maximal consistent blocks, we may define a relation in a similar way as a characteristic relation is defined from the characteristic sets. This relation, denoted by $S(B)$, will be called a relation implied by the family $\mathscr{C}(B)$ of maximal consistent blocks. We show that for data sets with all missing attribute values interpreted as lost $S(B)$ is an equivalence relation.

Additionally, an obvious question is what the better choice for data mining is: characteristic sets or maximal consistent blocks, in terms of an error rate computed as the result of ten-fold cross validation. We conducted experiments on data sets with two interpretations of missing attribute values, lost values and "do not care" conditions. For our experiments we used three kinds of probabilistic approximations: lower, middle (with $\alpha = 0.5$) and upper. Surprisingly, for many data sets quality of rule sets based on either characteristic sets or maximal consistent blocks do not differ significantly. However, when the difference is significant, using middle approximations based on characteristic sets or on maximal consistent blocks is the best approach. Some preliminary results of these experiments were presented in [18].

2. Incomplete data sets

An example of an incomplete data set is presented in Table 1. Lost values and "do not care" conditions are denoted by symbols of "?" and "*", respectively. In Table 1 there are nine missing attribute values, the total number of potential attribute values is $8 \times 3 = 24$, so the percentage of missing attribute values is 37.5%. A *concept* is a set of all cases with the same decision value. In Table 1 there are two concepts, the set {1, 2, 3, 4} of all cases with flu and the other set {5, 6, 7, 8}.

We use notation $a(x) = v$ if an attribute a has the value v for the case x. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

For complete data sets, for an attribute-value pair (a, v), a *block* of (a, v), denoted by $[(a, v)]$, is the following set

$$[(a,\ v)] = \{x | x \in U, a(x) = v\}.$$

For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way [1,19]:

- If for an attribute a and a case x, $a(x) = ?$, then the case x should not be included in any blocks $[(a,\ v)]$ for all values v of attribute a,
- If for an attribute a and a case x, $a(x) = *$, then the case x should be included in blocks $[(a,\ v)]$ for all specified values v of attribute a.

For the data set from Table 1, all blocks of attribute-value pairs are

$$[(Temperature, normal)] = \{2, 3, 4, 5, 6, 7\},$$
$$[(Temperature, high)] = \{1, 2, 3, 6, 8\},$$
$$[(Headache, no)] = \{3, 5, 6, 7, 8\},$$
$$[(Headache, yes)] = \{2, 7, 8\},$$
$$[(Cough, no)] = \{2, 3, 5, 8\},$$
$$[(Cough, yes)] = \{1, 3, 4, 6\}.$$