Chapter 4

Discovery of Data Patterns with Applications to Decomposition and Classification Problems

Sinh Hoa Nguyen¹, Andrzej Skowron², Piotr Synak³

¹ Institute of Computer Science, Warsaw University, Banacha 2, 02-097 Warsaw, Poland  
e-mail: hoa@mimuw.edu.pl

² Institute of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland  
e-mail: skowron@mimuw.edu.pl

³ Polish-Japanese Institute of Computer Techniques Koszykowa 86, 02-018 Warsaw, Poland  
e-mail: synak@pjwstk.waw.pl

1 Introduction

Data mining community is searching for efficient methods of extracting patterns from data [20],[22],[39],[46],[45]. We study problems of extracting several kinds of patterns from data. The simplest ones are called templates. We consider also more sophisticated relational patterns extracted automatically from data.

We present several strategies searching for patterns represented by so called templates. In the simplest case the template is a ”long enough” value vector of some features (attributes) supported by ”sufficiently many” objects. The high computational complexity of the searching problem for optimal templates shows that it is necessary to develop efficient heuristics for extracting efficiently semi-optimal templates from large data sets. Among the discussed heuristics there are some using information about the distribution of attribute values in data tables [26] easily computable from data. We also mention some more advanced techniques based on application of genetic algorithms [25], [51]. An important class of the methods for pattern discovery from data is based on relational patterns [37]. These patterns are defined in a given data table by the optimal similarity (tolerance) relations in some preassumed classes of tolerance relations [28]. A tolerance relation is optimal if the set of parameters (with respect to the assumed quality measure) specifying this relation allows to construct the relevant data patterns for a given data table.

There are different possible applications of patterns extracted from data.

Some of them can be used to decompose large data tables [25], [28]. The set of objects supporting e.g. a given template can be treated as regular, in a sense, sub-domain of object universe because it consists of many objects sharing many common features. The large data tables can be decomposed into a binary tree of templates or patterns. Each node of the tree is related to one step of decom-
position. The process of the decomposition stops when the sub-tables attached to leaves have a feasible size with respect to the existing methods of decision rules generation. We apply previously developed methods based on rough set approach (see e.g. [5], [30],[22],[30],[35]) for decision rules generation from the decision tables attached to leaves. In the process of new cases classification for any new object a path in the tree is selected by matched templates. Next the object is classified on the basis of decision rules generated from the sub-table attached to the leaf of that path.

We also discuss strategies searching for patterns (almost) included in decision classes. This process can be treated as searching for strong approximate default decision rules [22].

Our methods can also be used to search for approximate decision rule synthesis from data tables. The approximate nature of these rules is specified by some constraints. The strong decision rule can be understood like in the case of associations (see e.g. [1], [2]) but can also be characterized by some additional constraints e.g. assuming a high specificity (see e.g. [11], [21]) of the synthesized approximate decision rules guaranteed by the discovered templates or patterns. It is important to observe that relational patterns are expressed in a higher level language than templates so the former ones can lead to better generalization than the latter.

In the paper we concentrate on some efficient methods for patterns generation from data and their application to decomposition of data tables and object classification. We discuss the results of the performed computer experiments. We also investigate the complexity of the searching problem for the optimal template.

The paper consists of five parts. Introduction as well as general remarks related to the pattern discovery problem are presented in the first part. In the second part we introduce rough set preliminaries used in the paper. Methods for template generation from data tables are investigated in the third part. We also present some applications of discovered templates. In the fourth part we describe the relational pattern problem and methods for relational patterns extraction from data and their applications. The conclusions are included in the last part.

The third part of the paper is organized as follows:

In Section 3.1 we recall the template definition. We investigate the computational complexity of the template problem in Section 3.2. In Section 3.3 we show some searching methods for semi-optimal templates. The applications of templates for classification and decomposition are discussed in Section 3.4.

The fourth part of the paper is organized as follows:

In Section 4.1 we introduce some basic definitions related to patterns defined by tolerance relations. In Section 4.2 we propose a classification of methods searching for tolerance relation from data. A geometrical illustration of some tolerance relation families used for discovery of relational patterns is discussed in Section 4.3. In Section 4.4 we show some heuristics for semi-optimal tolerance relation generation. The applications of discovered from data tolerance relations are discussed in Section 4.5. The experimental results of methods based on tolerance relation are presented in the last section.