Studies on Rough Sets in Multiple Tables

R.S. Milton[1], V. Uma Maheswari[2], and Arul Siromoney[2]

[1] Department of Computer Science,
Madras Christian College, Chennai – 600 059, India
[2] Department of Computer Science and Engineering,
College of Engineering Guindy,
Anna University, Chennai – 600 025, India
asiro@vsnl.com, umam@annauniv.edu

Abstract. Rough Set Theory is a mathematical tool to deal with vagueness and uncertainty. Rough Set Theory uses a single information table. Relational Learning is the learning from multiple relations or tables. This paper studies the use of Rough Set Theory and Variable Precision Rough Sets in a multi-table information system (MTIS). The notion of approximation regions in the MTIS is defined in terms of those of the individual tables. This is used in classifying an example in the MTIS, based on the elementary sets in the individual tables to which the example belongs. Results of classification experiments in predictive toxicology based on this approach are presented.

Keywords: Rough Set Theory, Variable Precision Rough Sets, multi-table information system, relational learning, prediction.

1 Introduction

Rough set theory [1,2,3,4], introduced by Zdzislaw Pawlak in the early 1980s, is a mathematical tool to deal with vagueness and uncertainty. Rough set theory defines an indiscernibility relation that partitions the universe of examples into elementary sets. In other words, examples in an elementary set are indistinguishable. A concept is rough when it contains at least one elementary set that contains both positive and negative examples. The indiscernibility relation is defined based on a single table.

Relational Learning is based on multiple relations or tables. Inductive Logic Programming (ILP) [5,6] is one of the approaches to Relational Learning. A brief survey of research in Rough Sets and Relational Learning is presented in [7].

The authors' work is in the intersection of Rough Sets and ILP. The gRS–ILP model [8,9] introduces a rough setting in Inductive Logic Programming. It describes the situation where any induced logic program cannot distinguish between certain positive and negative examples. Any induced logic program will either cover both the positive and the negative examples in the group, or not cover the group at all, with both the positive and the negative examples in this group being left out.

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3641, pp. 265–274, 2005.

The Variable Precision Rough Set (VPRS) model [10] allows for a controlled degree of misclassification. The Variable Precision Rough Set Inductive Logic Programming (VPRSILP) model [11] is an extension of the gRS–ILP model using features of the VPRS model. The cVPRSILP approach [12] uses clauses as the attributes. Test cases are classified based on their proximity to significant elementary sets.

In this paper, a Multi–Table Information System is defined that extends the single information table of rough set theory to multiple information tables. Notions from the Variable Precision Rough Set model are then introduced and used for prediction. An illustrative experiment in toxicology is then presented.

2 Multi–table Information System

2.1 Definitions

Tables and Information Systems: Consider a universe U of elements. A table T is defined as $T = (U, A, V, \rho)$, where A is a finite set of *attributes*; $V = \bigcup_{a \in A} V_a$ is the set of *attribute values* of all attributes, where V_a is the *domain* (the set of possible values) of attribute a; and $\rho : U \times A \to V$ is an *information function* such that for every element $x \in U$, $\rho(x, a) \in V_a$ is the value of attribute a for element x. This definition is based on the definition of Rough Set Information System in [1].

Each element $x \in U$ can be pictured as corresponding to a row in a table of rows and columns, with each column corresponding to an attribute $a \in A$. $\rho(x, a)$ is the value in the table at the intersection of the row corresponding to x and the column corresponding to a.

We define a *Multi-Table Information System (MTIS)* as a finite set of tables denoted as $T = \{T_0, T_1, \ldots, T_n\}$, where each table $T_i, 1 \leq i \leq n$ is defined as above and is denoted as $T_i = (U_i, A_i, V_i, \rho_i)$. $T_0 = \{U_0, A_0, V_0, \rho_0\}$ is a decision table with one of the attributes $d \in A_0$, as a binary valued *decision attribute*. A similar definition is also found in [13]. We denote U_0 as U, the universe of examples, since A_0 has the decision attribute.

In every $T_i, 0 \leq i \leq n$, let $L_i \subset A_i$ consist of *link attributes* that are used to link different tables. These are attributes that are common between different tables.

Let $I_i = (U_i, B_i, V_i, \rho_i), 0 \leq i \leq n$, where $B_i = A_i - L_i$. We note that I_i corresponds to the classical rough set information system.

Elementary Sets: An equivalence relation R_i, called *indiscernibility relation*, is defined on the universe U_i of an information system I_i, $0 \leq i \leq n$,as

$$R_i = \{(x, y) \in U_i \times U_i \mid \forall b \in B_i, \rho_i(x, b) = \rho_i(y, b)\}$$

In the information system I_i, the *elementary set* containing the element $x \in U_i$, with respect to the indiscernibility relation R_i, is

$$[x]_{R_i} = \{y \in U_i \mid yR_ix\}$$