Data Mining via Discretization, Generalization and Rough Set Feature Selection

Xiaohua Hu¹ and Nick Cercone²

¹ Knowledge Stream Partner, 148 State St., Boston, MA 02109, USA
² Dept. of Computer Science, University of Waterloo, Waterloo, Canada

Received 10 July 1998
Revised 9 September 1998
Accepted 8 October 1998

Abstract. We present a data mining method which integrates discretization, generalization and rough set feature selection. Our method reduces the data horizontally and vertically. In the first phase, discretization and generalization are integrated. Numeric attributes are discretized into a few intervals. The primitive values of symbolic attributes are replaced by high level concepts and some obvious superfluous or irrelevant symbolic attributes are also eliminated. The horizontal reduction is done by merging identical tuples after substituting an attribute value by its higher level value in a pre-defined concept hierarchy for symbolic attributes, or the discretization of continuous (or numeric) attributes. This phase greatly decreases the number of tuples we consider further in the database(s). In the second phase, a novel context-sensitive feature merit measure is used to rank features, a subset of relevant attributes is chosen, based on rough set theory and the merit values of the features. A reduced table is obtained by removing those attributes which are not in the relevant attributes subset and the data set is further reduced vertically without changing the interdependence relationships between the classes and the attributes. Finally, the tuples in the reduced relation are transformed into different knowledge rules based on different knowledge discovery algorithms. Based on these principles, a prototype knowledge discovery system DBROUGH-II has been constructed by integrating discretization, generalization, rough set feature selection and a variety of data mining algorithms. Tests on a telecommunication customer data warehouse demonstrates that different kinds of knowledge rules, such as characteristic rules, discriminant rules, maximal generalized classification rules, and data evolution regularities, can be discovered efficiently and effectively.
Keywords: Data mining, knowledge rules, data warehouse, discretization, generalization, feature selection, rough set.

1 Introduction

One goal of the knowledge discovery from databases (KDD) process is the extraction of rules and understandable patterns from raw data [5]. The huge and growing amount of data in databases presents one of the most challenging problems and presents one of the best opportunities for KDD system designers. Before applying data mining algorithms and techniques which permit the effective and efficient extraction of rules and understandable forms (patterns) from data in large databases, we must reduce the data first. Data driven techniques either rely on heuristics to guide their search through the large space of possible relations between combinations of attribute values and classes hidden in the data or reduce the search space.

We present a method to mine different kinds of rules from databases by integrating discretization, generalization and rough set feature selection. Our method reduces the data horizontally and vertically. In the first phase, discretization and generalization are integrated. Numeric attributes are discretized into a few intervals. If discretization results in one interval, then the attribute is removed. The primitive values of symbolic attributes are replaced by high level concepts and some obvious superfluous or irrelevant symbolic attributes are also eliminated. The horizontal reduction is done by merging identical tuples after substituting an attribute value by its higher level value in a pre-defined concept hierarchy for symbolic attributes, or the discretization of continuous (or numeric) attributes. This phase greatly decreases the number of tuples we consider further in the database(s). In the second phase, a novel context-sensitive feature merit measure is used to rank features, a subset of relevant attributes is chosen, based on rough set theory and the merit values of the features. A reduced table is obtained by removing those attributes which are not in the relevant attributes subset and the data set is further reduced vertically without changing the interdependence relationships between the classes and the attributes. Rough set-based value reduction is further performed on the reduced table and all redundant condition values are dropped. Finally, the tuples in the reduced relation are transformed into different knowledge rules based on different knowledge discovery algorithms. Based on these principles, a prototype knowledge discovery system, DBROUGH-II coded in Java, has been constructed by integrating discretization, generalization, rough set feature selection and a variety of data mining algorithms.

The rest of our paper consists of the following. Using an example for illustration, we describe the principle of discretization and generalization in Section 2. In Section 3, we introduce a novel feature merit measure, which we use to rank features, and we present a rough set based feature selection algorithm. This algorithm is used to choose a relevant and important feature subset for the learning task. The architecture of DBROUGH-II is discussed in Section 4. Tests on a telecommunication customer data warehouse demonstrates that dif-