Approximate Boolean Reasoning: Foundations and Applications in Data Mining

Hung Son Nguyen

Institute of Mathematics,
Warsaw University Banacha 2, 02-097
Warsaw, Poland
son@mimuw.edu.pl

Abstract. Since its introduction by George Boole during the mid-1800s, Boolean algebra has become an important part of the *lingua franca* of mathematics, science, engineering, and research in artificial intelligence, machine learning and data mining. The Boolean reasoning approach has manifestly become a powerful tool for designing effective and accurate solutions for many problems in decision-making and approximate reasoning optimization. In recent years, Boolean reasoning has become a recognized technique for developing many interesting concept approximation methods in rough set theory. The problem considered in this paper is the creation of a general framework for concept approximation. The need for such a general framework arises in machine learning and data mining. This paper presents a solution to this problem by introducing a general framework for concept approximation which combines rough set theory, Boolean reasoning methodology and data mining. This general framework for approximate reasoning is called *Rough Sets and Approximate Boolean Reasoning* (RSABR). The contribution of this paper is the presentation of the theoretical foundation of RSABR as well as its application in solving many data mining problems and knowledge discovery in databases (KDD) such as feature selection, feature extraction, data preprocessing, classification of decision rules and decision trees, association analysis.

Keywords: Rough sets, data mining, boolean reasoning, feature selection and extraction, decision rule construction, discretization, decision tree induction, association rules, large data tables.

1 Introduction

The rapidly growing volume and complexity of modern databases make the need for technologies to describe and summarize the information they contain increasingly important. Knowledge Discovery in Databases (KDD) and data mining are new research areas that try to overcome this problem. In [32], KDD was characterized as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data, while data mining is a process of extracting implicit, previously unknown and potentially useful patterns and relationships from data, and it is widely used in industry and business applications.

J.F. Peters and A. Skowron (Eds.): Transactions on Rough Sets V, LNCS 4100, pp. 334–506, 2006.

As the main step in KDD, data mining methods are required to be not only accurate but also to deliver understandable and interpretable results for users, e.g., through visualization. The other important issue of data mining methods is their complexity and scalability. Presently, data mining is a collection of methods from various disciplines such as mathematics, statistics, logics, pattern recognition, machine learning, non-conventional models and heuristics for computing [43], [45], [155] [67].

Concept approximation is one of the most fundamental issues in machine learning and data mining. The problem considered in this paper is the creation of a general framework for concept approximation. The need for such a general framework arises in machine learning and data mining. Classification, clustering, association analysis or regression are examples of well-known problems in data mining that can be considered in the context of concept approximation. A great effort by many researchers has led to the design of newer, faster and more efficient methods for solving the concept approximation problem [100].

Rough set theory has been introduced by Zdzisław Pawlak [109] as a tool for concept approximation relative to uncertainty. Basically, the idea is to approximate a concept by three description sets, namely, *lower approximation, upper approximation* and *boundary region*. These three sets have been fundamental to the basic approach of rough set theory, since its introduction by Zdzisław Pawlak during the early 1980s (see, e.g., [107], [108], [109], [110]). The approximation process begins by partitioning a given set of objects into equivalence classes called blocks, where the objects in each block are indiscernible from each other relative to their attribute values. The approximation and boundary region sets are derived from the blocks of a partition of the available objects. The boundary region is constituted by the difference between the lower approximation and upper approximation, and provides a basis for measuring the "roughness" of an approximation. Central to the philosophy of the rough set approach to concept approximation is minimization of the boundary region. This simple but brilliant idea leads to many efficient applications of rough sets in machine learning and data mining such as feature selection, rule induction, discretization or classifier construction [57], [58], [143], [137], [142], [79].

Boolean algebra has become part of the *lingua franca* of mathematics, science, engineering, and research in artificial intelligence, machine learning and data mining ever since its introduction by George Boole during the 19th century [13]. In recent years, the combination of Boolean reasoning approach and rough set methods have provided powerful tools for designing effective as well as accurate solutions for many machine learning and data mining problems [141], [97], [91], [162], [142], [139], [164], [61], [38].

The problem considered in this paper is the creation of a general framework for concept approximation. The need for such a general framework arises in machine learning and data mining. This paper presents a solution to this problem by introducing a general framework for concept approximation which combines rough set theory, Boolean reasoning methodology and data mining. This general framework for approximate reasoning is called *Rough Sets and Approximate*