Foundations of Classification

J. T. Yao Y. Y. Yao and Y. Zhao

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{jtyao, yyao, yanzhao}@cs.uregina.ca

Summary. Classification is one of the main tasks in machine learning, data mining, and pattern recognition. A granular computing model is suggested for learning two basic issues of concept formation and concept relationship identification. A classification problem can be considered as a search for suitable granules organized under a partial order. The structures of search space, solutions to a consistent classification problem, and the structures of solution space are discussed. A classification rule induction method is proposed. Instead of searching for a suitable partition, we concentrate on the search for a suitable covering of the given universe. This method is more general than partition-based methods. For the design of covering granule selection heuristics, several measures on granules are suggested.

1 Introduction

Classification is one of the main tasks in machine learning, data mining, and pattern recognition [3, 10, 12]. It deals with classifying labelled objects. Knowledge for classification can be expressed in different forms, such as classification rules, discriminant functions, and decision trees. Extensive research has been done on the construction of classification models.

Mainstream research in classification focus on classification algorithms and their experimental evaluations. By comparison, less attention has been paid to the study of fundamental concepts such as structures of search space, solution to a consistent classification problem, as well as the structures of a solution space. For this reason, we present a granular computing based framework for a systematic study of these fundamental issues.

Granular computing is an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules in problem solving [25, 27, 33, 34]. A *granule* is a subset of the universe. A family of granules that contains every object in the universe is called a *granulation* of the universe. The granulation of a given universe involves dividing the universe into subsets or grouping individual objects into clusters. There are many fundamental issues in granular computing, such as the granulation of a given

universe, the descriptions of granules, the relationships between granules, and the computation of granules.

Data mining, especially rule-based mining, can be molded in two steps, namely, the formation of concepts and the identification of relationship between concepts. Formal concept analysis may be considered as a concrete model of granular computing. It deals with the characterization of a concept by a unit of thoughts consisting the intension and the extension of the concept [4, 23]. From the standing point of granular computing, the concept of a granule may be exemplified by a set of instances, i.e., the extension; the concept of a granule may be described or labelled by a name, i.e., the intension. Once concepts are constructed and described, one can develop computational methods using granules [27]. In particular, one may study relationships between concepts in terms of their intensions and extensions, such as sub-concepts and super-concepts, disjoint and overlap concepts, and partial sub-concepts. These relationships can be conveniently expressed in the form of rules and associated quantitative measures indicating the strength of rules. By combining the results from formal concept analysis and granular computing, knowledge discovery and data mining, especially rule mining, can be viewed as a process of forming concepts and finding relationships between concepts in terms of intensions and extensions [28, 30, 32].

The organization of this chapter is as follows. In Section 2, we first present the fundamental concepts of granular computing which serve as the basis of classification problems. Measures associated with granules for classification will be studied in Section 3. In Section 4, we will examine the search spaces of classification rules. In Section 5, we remodel the ID3 and PRISM classification algorithms from the viewpoint of granular computing. We also propose the kLR algorithm and a granule network algorithm to complete the study of the methodology in granular computing model.

2 Fundamentals of a Granular Computing Model for Classification

This section provides an overview of the granular computing model [28, 30].

2.1 Information tables

Information tables are used in granular computing models. An information table provides a convenient way to describe a finite set of objects called a universe by a finite set of attributes [14, 33]. It represents all available information and knowledge. That is, objects are only perceived, observed, or measured by using a finite number of properties.