Foreword

The importance of having efficient and effective methods for data mining and knowledge discovery (DM&KD), to which the present book is devoted, grows every day and numerous such methods have been developed in recent decades. There exists a great variety of different settings for the main problem studied by data mining and knowledge discovery, and it seems that a very popular one is formulated in terms of binary attributes. In this setting, states of nature of the application area under consideration are described by Boolean vectors defined on some attributes. That is, by data points defined in the Boolean space of the attributes. It is postulated that there exists a partition of this space into two classes, which should be inferred as patterns on the attributes when only several data points are known, the so-called positive and negative training examples.

The main problem in DM&KD is defined as finding rules for recognizing (classifying) new data points of unknown class, i.e., deciding which of them are positive and which are negative. In other words, to infer the binary value of one more attribute, called the goal or class attribute. To solve this problem, some methods have been suggested which construct a Boolean function separating the two given sets of positive and negative training data points. This function can then be used as a decision function, or a classifier, for dividing the Boolean space into two classes, and so uniquely deciding for every data point the class to which it belongs. This function can be considered as the knowledge extracted from the two sets of training data points.

It was suggested in some early works to use as classifiers threshold functions defined on the set of attributes. Unfortunately, only a small part of Boolean functions can be represented in such a form. This is why the normal form, disjunctive or conjunctive (DNF or CNF), was used in subsequent developments to represent arbitrary Boolean decision functions. It was also assumed that the simpler the function is (that is, the shorter its DNF or CNF representation is), the better classifier it is. That assumption was often justified when solving different real-life problems. This book suggests a new development of this approach based on mathematical logic and, especially, on using Boolean functions for representing knowledge defined on many binary attributes.

Next, let us have a brief excursion into the history of this problem, by visiting some old and new contributions. The first known formal methods for expressing logical reasoning are due to Aristotle (384 BC–322 BC) who lived in ancient Greece, the native land of the author. It is known as his famous *syllogistics*, the first deductive system for producing new affirmations from some known ones. This can be acknowledged as being the first system of logical recognition. A long time later, in the 17th century, the notion of binary mathematics based on a two-value system was proposed by Gottfried Leibniz, as well as a combinatorial approach for solving some related problems. Later on, in the middle of the 19th century, George Boole wrote his seminal books *The mathematical analysis of logic: being an essay towards a calculus for deductive reasoning* and *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. These contributions served as the foundations of modern Boolean algebra and spawned many branches, including the theory of proofs, logical inference and especially the theory of Boolean functions. They are widely used today in computer science, especially in the area of the design of logic circuits and artificial intelligence (AI) in general.

The first real-life applications of these theories took place in the first thirty years of the 20th century. This is when Shannon, Nakashima and Shestakov independently proposed to apply Boolean algebra to the description, analysis and synthesis of relay devices which were widely used at that time in communication, transportation and industrial systems. The progress in this direction was greatly accelerated in the next fifty years due to the dawn of modern computers. This happened for two reasons. First, in order to design more sophisticated circuits for the new generation of computers, new efficient methods were needed. Second, the computers themselves could be used for the implementation of such methods, which would make it possible to realize very difficult and labor-consuming algorithms for the design and optimization of multicomponent logic circuits. Later, it became apparent that methods developed for the previous purposes were also useful for an important problem in artificial intelligence, namely, data mining and knowledge discovery, as well as for pattern recognition.

Such methods are discussed in the present book, which also contains a wide review of numerous computational results obtained by the author and other researches in this area, together with descriptions of important application areas for their use. These problems are combinatorially hard to solve, which means that their exact (optimal) solutions are inevitably connected with the requirement to check many different intermediate constructions, the number of which depends exponentially on the size of the input data. This is why good combinatorial methods are needed for their solution. Fortunately, in many cases efficient algorithms could be developed for finding some approximate solutions, which are acceptable from the practical point of view. This makes it possible to sufficiently reduce the number of intermediate solutions and hence to restrict the running time.

A classical example of the above situation is the problem of minimizing a Boolean function in disjunctive (or conjunctive) normal form. In this monograph, this task is pursued in the context of searching for a Boolean function which separates two given subsets of the Boolean space of attributes (as represented by collections

of positive and negative examples). At the same time, such a Boolean function is desired to be as simple as possible. This means that incompletely defined Boolean functions are considered. The author, Professor Evangelos Triantaphyllou, suggests a set of efficient algorithms for inferring Boolean functions from training examples, including a fast heuristic greedy algorithm (called OCAT), its combination with tree searching techniques (also known as branch-and-bound search), an incremental learning algorithm, and so on. These methods are efficient and can enable one to find good solutions in cases with many attributes and data points. Such cases are typical in many real-life situations where such problems arise. The special problem of guided learning is also investigated. The question now is which new training examples (data points) to consider, one at a time, for training such that a small number of new examples would lead to the inference of the appropriate Boolean function quickly.

Special attention is also devoted to monotone Boolean functions. This is done because such functions may provide adequate description in many practical situations. The author studied existing approaches for the search of monotone functions, and suggests a new way for inferring such functions from training examples. A key issue in this particular investigation is to consider the number of such functions for a given dimension of the input data (i.e., the number of binary attributes).

Methods of DM&KD have numerous important applications in many different domains in real life. It is enough to mention some of them, as described in this book. These are the problems of verifying software and hardware of electronic devices, locating failures in logic circuits, processing of large amounts of data which represent numerous transactions in supermarkets in order to optimize the arrangement of goods, and so on. One additional field for the application of DM&KD could also be mentioned, namely, the design of two-level (AND-OR) logic circuits implementing Boolean functions, defined on a small number of combinations of values of input variables.

One of the most important problems today is that of breast cancer diagnosis. This is a critical problem because diagnosing breast cancer early may save the lives of many women. In this book it is shown how training data sets can be formed from descriptions of malignant and benign cases, how input data can be described and analyzed in an objective and consistent manner and how the diagnostic problem can be formulated as a nested system of two smaller diagnostic problems. All these are done in the context of Boolean functions.

The author correctly observes that the problem of DM&KD is far from being fully investigated and more research within the framework of Boolean functions is needed. Moreover, he offers some possible extensions for future research in this area. This is done systematically at the end of each chapter.

The descriptions of the various methods and algorithms are accompanied with extensive experimental results confirming their efficiency. Computational results are generated as follows. First a set of test cases is generated regarding the approach to be tested. Next the proposed methods are applied on these test problems and the test results are analyzed graphically and statistically. In this way, more insights on the

problem at hand can be gained and some areas for possible future research can be identified.

The book is very well written in a way for anyone to understand with a minimum background in mathematics and computer science concepts. However, this is not done at the expense of the mathematical rigor of the algorithmic developments. I believe that this book should be recommended both to students who wish to learn about the foundations of logic-based approaches as they apply to data mining and knowledge discovery along with their many applications, and also to researchers who wish to develop new means for solving more problems effectively in this area.

Professor Arkadij Zakrevskij

Minsk, Belarus
Corresponding Member of the National Academy of
Sciences of Belarus

Summer of 2009