

Preface

Data mining is the science and technology of exploring data in order to discover previously unknown patterns. It is a part of the overall process of knowledge discovery in databases (KDD). The accessibility and abundance of information today makes data mining a matter of considerable importance and necessity.

One of the most practical approaches in data mining is to use induction algorithms for constructing a model by generalizing from given data. The induced model describes and explains phenomena which are hidden in the data. Given the recent growth of the field as well as its long history, it is not surprising that several mature approaches to induction are now available to the practitioner. However according to the “no free lunch” theorem, there is no single approach that outperforms all others in all possible domains. Evidently, in the presence of a vast repertoire of techniques and the complexity and diversity of the explored domains, the main challenge today in data mining is to know how to utilize this repertoire in order to achieve maximum reliability, comprehensibility and complexity.

Multiple classifiers methodology is considered an effective way of overcoming this challenge. The basic idea is to build a model by integrating multiple models. Researchers distinguish between two multiple classifier methodologies: ensemble methodology and decomposition methodology. Ensemble methodology combines a set of models, each of which solves the same original task. Decomposition methodology breaks down the classification task into several manageable classification tasks, enabling each inducer to solve a different task

This book focuses on decomposition in general data mining tasks and for classification tasks in particular. The book presents a complete methodology for decomposing classification problems into smaller and more man-

ageable sub-problems that are solvable by using existing tools. The various elements are then joined together to solve the initial problem.

The benefits of decomposition methodology in data mining include: increased performance (classification accuracy); conceptual simplification of the problem; enhanced feasibility for huge databases; clearer and more comprehensible results; reduced runtime by solving smaller problems and by using parallel/distributed computation; and the opportunity of using different solution techniques for individual sub-problems. These features are discussed in the book.

Obviously the most essential question that decomposition methodology should be able to answer is whether a given classification problem should be decomposed and in what manner. The main theory presented in this book is that the decomposition can be achieved by recursively performing a sequence of single, elementary decompositions. The book introduces several fundamental and elementary decomposition methods, namely: Attribute Decomposition, Space Decomposition, Sample Decomposition, Function Decomposition, and Concept Decomposition. We propose a unifying framework for using these methods in real applications.

The book shows that the decomposition methods developed here extend the envelope of problems that data mining can efficiently solve. These methods also enhance the comprehensibility of the results that emerge and suggest more efficient implementation of knowledge discovery conclusions.

In this comprehensive study of decomposition methodology, we try to answer several vital questions:

- What types of elementary decomposition methods exist in concept learning?
- Which elementary decomposition type performs best for which problem? What factors should one take into account when choosing the appropriate decomposition type?
- Given an elementary type, how should we infer the best decomposition structure automatically?
- How should the sub-problems be re-composed to represent the original concept learning?

The decomposition idea shares properties with other fields mainly ensemble methods, structured induction and distributed data mining. Numerous researches have been performed in these areas and the methodology described in this book exploits the fruits of these insightful studies. However, the book introduces a broader methodology, which results from

a rather different motivation: the desire to decompose data mining tasks and gain the benefit mentioned above.

This book was written to provide investigators in the fields of information systems, engineering, computer science, statistics and management, with a comprehensive source for decomposition techniques. In addition, those engaged in the social sciences, psychology, medicine, genetics, and other data-rich fields can very much benefit from this book.

Much of the material in this book has been developed and taught in undergraduate and graduate courses at Tel Aviv University. In particular we would like to acknowledge four distinguished graduate students that contributed to this book: Omri Arad, Lital Keshet, Inbal Lavi and Anat Okon. Therefore, the book can also serve as a text or reference book for graduate/advanced undergraduate level courses in data mining and machine learning. Practitioners among the readers may be particularly interested in the descriptions of real-world data mining projects performed with decomposition methodology.

*Oded Maimon
Lior Rokach*