

Preface

We come across a number of celebrated text books on Data Mining covering multiple aspects of the topic since its early development, such as those on databases, pattern recognition, soft computing, etc. We did not find any consolidated work on data mining in compression domain. The book took shape from this realization. Our work relates to this area of data mining with a focus on compaction. We present schemes that work in compression domain and demonstrate their working on one or more practical datasets in each case. In this process, we cover important data mining paradigms. This is intended to provide a practitioners' view point of compression schemes in data mining. The work presented is based on the authors' work on related areas over the last few years. *We organized each chapter to contain context setting, background work as part of discussion, proposed algorithm and scheme, implementation intricacies, experimentation by implementing the scheme on a large dataset, and discussion of results. At the end of each chapter, as part of bibliographic notes, we discuss relevant literature and directions for further study.*

Data Mining focuses on efficient algorithms to generate abstraction from large datasets. The objective of these algorithms is to find interesting patterns for further use by the least number of visits of entire dataset, ideal being a single visit. Similarly, since the data sizes are large, effort is made in arriving at a much smaller subset of the original dataset that is a representative of entire data and contains attributes characterizing the data. The ability to generate an abstraction from a small representative set of patterns and features that is as accurate as that can be obtained with entire dataset leads to efficiency in terms of both space and time. Important data mining paradigms include clustering, classification, association rule mining, etc. We present a discussion on data mining paradigms in Chap. 2.

In our present work, in addition to data mining paradigms discussed in Chap. 2, we also focus on another paradigm, viz., the ability to generate abstraction in the compressed domain without having to decompress. Such a compression would lead to less storage and improve the computation cost. In the book, we consider both lossy and nonlossy compression schemes. In Chap. 3, we present a nonlossy compression scheme based on run-length encoding of patterns with binary-valued features. The scheme is also applicable to floating-point-valued features that are suit-

ably quantized to binary values. The chapter presents an algorithm that computes the dissimilarity in the compressed domain directly. Theoretical notes are provided for the work. We present applications of the scheme in multiple domains.

It is interesting to explore when one is prepared to lose some part of pattern representation, whether we obtain better generalization and compaction. We examine this aspect in Chap. 4. The work in the chapter exploits the concept of minimum feature or item-support. The concept of support relates to the conventional association rule framework. We consider patterns as sequences, form subsequences of short length, and identify and eliminate repeating subsequences. We represent the pattern by those unique subsequences leading to significant compaction. Such unique subsequences are further reduced by replacing less frequent unique subsequences by more frequent subsequences, thereby achieving further compaction. We demonstrate the working of the scheme on large handwritten digit data.

Pattern clustering can be construed as compaction of data. Feature selection also reduces dimensionality, thereby resulting in pattern compression. It is interesting to explore whether they can be simultaneously achieved. We examine this in Chap. 5. We consider an efficient clustering scheme that requires a single database visit to generate prototypes. We consider a lossy compression scheme for feature reduction. We also examine whether there is preference in sequencing prototype selection and feature selection in achieving compaction, as well as good classification accuracy on unseen patterns. We examine multiple combinations of such sequencing. We demonstrate working of the scheme on handwritten digit data and intrusion detection data.

Domain knowledge forms an important input for efficient compaction. Such knowledge could either be provided by a human expert or generated through an appropriate preliminary statistical analysis. In Chap. 6, we exploit domain knowledge obtained both by expert inference and through statistical analysis and classify a 10-class data through a proposed decision tree of depth of 4. We make use of 2-class classifiers, AdaBoost and Support Vector Machine, to demonstrate working of such a scheme.

Dimensionality reduction leads to compaction. With algorithms such as run-length encoded compression, it is educative to study whether one can achieve efficiency in obtaining optimal feature set that provides high classification accuracy. In Chap. 7, we discuss concepts and methods of feature selection and extraction. We propose an efficient implementation of simple genetic algorithms by integrating compressed data classification and frequent features. We provide insightful discussion on the sensitivity of various genetic operators and frequent-item support on the final selection of optimal feature set.

Divide-and-conquer has been one important direction to deal with large datasets. With reducing cost and increasing ability to collect and store enormous amounts of data, we have massive databases at our disposal for making sense out of them and generate abstraction that could be of potential business exploitation. The term Big Data has been synonymous with streaming multisource data such as numerical data, messages, and audio and video data. There is increasing need for processing such data in real or near-real time and generate business value in this process. In Chap. 8,

we propose schemes that exploit multiagent systems to solve these problems. We discuss concepts of big data, MapReduce, PageRank, agents, and multiagent systems before proposing multiagent systems to solve big data problems.

The authors would like to express their sincere gratitude to their respective families for their cooperation.

T. Ravindra Babu and S.V. Subrahmanya are grateful to Infosys Limited for providing an excellent research environment in the Education and Research Unit (E&R) that enabled them to carry out academic and applied research resulting in articles and books.

T. Ravindra Babu likes to express his sincere thanks to his family members Padma, Ramya, Kishore, and Rahul for their encouragement and support. He dedicates his contribution of the work to the fond memory of his parents Butchiramaiah and Ramasitamma. M. Narasimha Murty likes to acknowledge support of his parents. S.V. Subrahmanya likes to thank his wife D.R. Sudha for her patient support. The authors would like to record their sincere appreciation for Springer team, Wayne Wheeler and Simon Rees, for their support and encouragement.

Bangalore, India

T. Ravindra Babu
M. Narasimha Murty
S.V. Subrahmanya

Compression Schemes for Mining Large Datasets

A Machine Learning Perspective

Ravindra Babu, T.; Murty, M.N.; Subrahmanya, S.V.

2013, XVI, 197 p. 62 illus., 3 illus. in color., Hardcover

ISBN: 978-1-4471-5606-2