
Preface

Data Mining has been identified as one of the ten emergent technologies of the 21st century (MIT Technology Review, 2001). This discipline aims at discovering knowledge relevant to decision making from large amounts of data. After some knowledge has been discovered, the final user (a decision-maker or a data-analyst) is unfortunately confronted with a major difficulty in the validation stage: he/she must cope with the typically numerous extracted pieces of knowledge in order to select the most interesting ones according to his/her preferences. For this reason, during the last decade, the designing of quality measures (or interestingness measures) has become an important challenge in Data Mining.

The purpose of this book is to present the state of the art concerning quality/interestingness measures for data mining. The book summarizes recent developments and presents original research on this topic. The chapters include reviews, comparative studies of existing measures, proposals of new measures, simulations, and case studies. Both theoretical and applied chapters are included.

Structure of the book

The book is structured in three parts. The first part gathers four overviews of quality measures. The second part contains four chapters dealing with data quality, data linkage, contrast sets and association rule clustering. Lastly, in the third part, four chapters describe new quality measures and rule validation.

PART I: OVERVIEWS OF QUALITY MEASURES

- **Chapter 1: Choosing the Right Lens: Finding What is Interesting in Data Mining**, by Geng and Hamilton, gives a broad overview

of the use of interestingness measures in data mining. This survey reviews interestingness measures for rules and summaries, classifies them from several perspectives, compares their properties, identifies their roles in the data mining process, describes methods of analyzing the measures, reviews principles for selecting appropriate measures for applications, and predicts trends for research in this area.

- **Chapter 2: A Graph-based Clustering Approach to Evaluate Interestingness Measures: A Tool and a Comparative Study**, by Hiep *et al.*, is concerned with the study of interestingness measures. As interestingness depends both on the structure of the data and on the decision-maker's goals, this chapter introduces a new contextual approach implemented in ARQAT, an exploratory data analysis tool, in order to help the decision-maker select the most suitable interestingness measures. The tool, which embeds a graph-based clustering approach, is used to compare and contrast the behavior of thirty-six interestingness measures on two typical but quite different datasets. This experiment leads to the discovery of five stable clusters of measures.
- **Chapter 3: Association Rule Interestingness Measures: Experimental and Theoretical Studies**, by Lenca *et al.*, discusses the selection of the most appropriate interestingness measures, according to a variety of criteria. It presents a formal and an experimental study of 20 measures. The experimental studies carried out on 10 data sets lead to an experimental classification of the measures. This studies leads to the design of a multi-criteria decision analysis in order to select the measures that best take into account the user's needs.
- **Chapter 4: On the Discovery of Exception Rules: A Survey**, by Duval *et al.*, presents a survey of approaches developed for mining exception rules. They distinguish two approaches to using an expert's knowledge: using it as syntactic constraints and using it to form as commonsense rules. Works that rely on either of these approaches, along with their particular quality evaluation, are presented in this survey. Moreover, this chapter also gives ideas on how numerical criteria can be intertwined with user-centered approaches.

PART II: FROM DATA TO RULE QUALITY

- **Chapter 5: Measuring and Modelling Data Quality for Quality-Awareness in Data Mining**, by Berti-Équille. This chapter offers an overview of data quality management, data linkage and data cleaning techniques that can be advantageously employed for improving quality awareness during the knowledge discovery process. It also details the steps of a

pragmatic framework for data quality awareness and enhancement. Each step may use, combine and exploit the data quality characterization, measurement and management methods, and the related techniques proposed in the literature.

- **Chapter 6: Quality and Complexity Measures for Data Linkage and Deduplication**, by Christen and Goiser, proposes a survey of different measures that have been used to characterize the quality and complexity of data linkage algorithms. It is shown that measures in the space of record pair comparisons can produce deceptive quality results. Various measures are discussed and recommendations are given on how to assess data linkage and deduplication quality and complexity.
- **Chapter 7: Statistical Methodologies for Mining Potentially Interesting Contrast Sets**, by Hilderman and Peckham, focuses on contrast sets that aim at identifying the significant differences between classes or groups. They compare two contrast set mining methodologies, STUCCO and CIGAR, and discuss the underlying statistical measures. Experimental results show that both methodologies are statistically sound, and thus represent valid alternative solutions to the problem of identifying potentially interesting contrast sets.
- **Chapter 8: Understandability of Association Rules: A Heuristic Measure to Enhance Rule Quality**, by Natarajan and Shekar, deals with the clustering of association rules in order to facilitate easy exploration of connections between rules, and introduces the *Weakness* measure dedicated to this goal. The average linkage method is used to cluster rules obtained from a small artificial data set. Clusters are compared with those obtained by applying a commonly used method.

PART III: RULE QUALITY AND VALIDATION

- **Chapter 9: A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link**, by Lerman and Azé, presents the foundations and the construction of a probabilistic interestingness measure called the likelihood of the link index. They discuss two facets, symmetrical and asymmetrical, of this measure and the two stages needed to build this index. Finally, they report the results of experiments to estimate the relevance of their statistical approach.
- **Chapter 10: Towards a Unifying Probabilistic Implicative Normalized Quality Measure for Association Rules**, by Diatta *et al.*, defines the so-called normalized probabilistic quality measures (PQM) for association rules. Then, they consider a normalized and implicative PQM

called M_{GK} , and discuss its properties.

- **Chapter 11: Association Rule Interestingness: Measure and Statistical Validation**, by Lallich *et al.*, is concerned with association rule validation. After reviewing well-known measures and criteria, the statistical validity of selecting the most interesting rules by performing a large number of tests is investigated. An original, bootstrap-based validation method is proposed that controls, for a given level, the number of false discoveries. The potential value of this method is illustrated by several examples.
- **Chapter 12: Comparing Classification Results between N -ary and Binary Problems**, by Felkin, deals with supervised learning and the quality of classifiers. This chapter presents a practical tool that will enable the data-analyst to apply quality measures to a classification task. More specifically, the tool can be used during the pre-processing step, when the analyst is considering different formulations of the task at hand. This tool is well suited for illustrating the choices for the number of possible class values to be used to define a classification problem and the relative difficulties of the problems that result from these choices.

Topics

The topics of the book include:

- Measures for data quality
- Objective vs subjective measures
- Interestingness measures for rules, patterns, and summaries
- Quality measures for classification, clustering, pattern discovery, etc.
- Theoretical properties of quality measures
- Human-centered quality measures for knowledge validation
- Aggregation of measures
- Quality measures for different stages of the data mining process,
- Evaluation of measure properties via simulation
- Application of quality measures and case studies

Review Committee

All published chapters have been reviewed by at least 2 referees.

- Henri Briand (LINA, University of Nantes, France)
- Rgis Gras (LINA, University of Nantes, France)
- Yves Kodratoff (LRI, University of Paris-Sud, France)
- Vipin Kumar (University of Minnesota, USA)
- Pascale Kuntz (LINA, University of Nantes, France)
- Robert Hilderman (University of Regina, Canada)
- Ludovic Lebart (ENST, Paris, France)
- Philippe Lenca (ENST-Bretagne, Brest, France)
- Bing Liu (University of Illinois at Chicago, USA)
- Amro Napoli (LORIA, University of Nancy, France)
- Gregory Piatetsky-Shapiro (KDNuggets, USA)
- Gilbert Ritschard (Geneve University, Switzerland)
- Sigal Sahar (Intel, USA)
- Gilbert Saporta (CNAM, Paris, France)
- Dan Simovici (University of Massachusetts Boston, USA)
- Jaideep Srivastava (University of Minnesota, USA)
- Einoshin Suzuki (Yokohama National University, Japan)
- Pang-Ning Tan (Michigan State University, USA)
- Alexander Tuzhilin (Stern School of Business, USA)
- Djamel Zighed (ERIC, University of Lyon 2, France)

Associated Reviewers

Jérôme Azé,
 Laure Berti-Equille,
 Libei Chen,
 Peter Christen,
 Béatrice Duval,
 Mary Felkin,
 Liqiang Geng,

Karl Goiser,
 Stéphane Lallich,
 Rajesh Natajaran,
 Ansaf Salleb,
 Benoît Vaillant

Acknowledgments

The editors would like to thank the chapter authors for their insights and contributions to this book.

The editors would also like to acknowledge the member of the review committee and the associated referees for their involvement in the review process of the book. Without their support the book would not have been satisfactorily completed.

A special thank goes to D. Zighed and H. Briand for their kind support and encouragement.

Finally, we thank Springer and the publishing team, and especially T. Ditzinger and J. Kacprzyk, for their confidence in our project.

Regina, Canada and Nantes, France,
May 2006

Fabrice Guillet
Howard Hamilton



<http://www.springer.com/978-3-540-44911-9>

Quality Measures in Data Mining

Guillet, F.; Hamilton, H.J. (Eds.)

2007, XIV, 314 p., Hardcover

ISBN: 978-3-540-44911-9