Honghua Dai · James N.K. Liu
Evgueni Smirnov *Editors*

Reliable Knowledge Discovery

Springer

Reliable Knowledge Discovery

Honghua Dai • James N.K. Liu • Evgueni Smirnov
Editors

Reliable Knowledge Discovery

Springer

Editors
Honghua Dai
Deakin University
Burwood, Victoria, Australia

James N.K. Liu
The Hong Kong Polytechnic University
Hong Kong

Evgueni Smirnov
Maastricht University
Maastricht, The Netherlands

ISBN 978-1-4614-1902-0 e-ISBN 978-1-4614-1903-7
DOI 10.1007/978-1-4614-1903-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2012932410

Preface

1 Description

With the rapid development of the data mining and knowledge discovery, a key issue which could significantly affect the real world applications of data mining is the reliability issues of knowledge discovery. It is natural that people will ask if the discovered knowledge is reliable. Why do we trust the discovered knowledge? How much can we trust the discovered knowledge? When it could go wrong. All these questions are very essential to data mining. It is especial crucial to the real world applications.

One of the essential requirements of data mining is validity. This means both the discovery process itself and the discovered knowledge should be valid. Reliability is a necessary but not sufficient condition for validity. Reliability could be viewed as stability, equivalence and consistency in some ways.

This special volume of the book on the reliability issues of Data Mining and Knowledge Discovery will focus on the theory and techniques that can ensure the discovered knowledge is reliable and to identify under which conditions the discovered knowledge is reliable or in which cases the discovery process is robust. In the last 20 years, many data mining algorithms have been developed for the discovery of knowledge from given data bases. However in some cases, the discovery process is not robust or the discovered knowledge is not reliable or even incorrect in certain cases. We could also find that in some cases, the discovered knowledge may not necessary be the real reflection of the data. Why does this happen? What are the major factors that affect the discovery process? How can we make sure that the discovered knowledge is reliable? What are the conditions under which a reliable discovery can be assured? These are some interesting questions to be investigated in this book.

2 Scope and Topics of this Book

The topics of this book covers the following:

- The theories on reliable knowledge discovery
- Reliable knowledge discovery methods
- Reliability measurement criteria of knowledge discovery
- Reliability estimation methods
- General reliability issues on knowledge discovery
- Domain specific reliability issues on knowledge discovery
- The criteria that can be used to assess the reliability of discovered knowledge.
- The conditions under which we can confidently say that the discovered knowledge is reliable.
- The techniques which can improve reliability of knowledge discovery
- Practical approaches that can be used to solve reliability problems of data mining systems.
- The theoretical work on data mining reliability
- The practical approaches which can be used to assess if the discovered knowledge is reliable.
- The analysis of the factors that affect data mining reliability
- How reliability can be assessed
- In which condition, the reliability of the discovered knowledge is assured.

3 The Theme and Related Resources

The main purpose of this book is to encourage the use of Reliable Knowledge Discovery from Databases (RKDD) in critical-domain applications related to society, science, and technology. The book is intended for practitioners, researchers, and advanced-level students. It can be employed primarily as a reference work and it is a good compliment to the excellent book on reliable prediction Algorithmic learning in a random world by Vladimir Vovk, Alex Gammerman, and Glenn Shafer (New York: Springer, 2005). Extra information sources are the proceedings of the workshops Reliability Issues in Knowledge Discovery held in conjunction with the IEEE International Conferences on Data Mining. Other relevant conferences are the Annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), the International Conference on Machine Learning (ICML), The pacific-Asia Conference on Knowledge Discovery (PAKDD), and the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). Many AI-related journals regularly publish work in RKDD. Among others it is worth mentioning the Journal of Data Mining and Knowledge Discovery, the Journal of Machine Learning Research, and the Journal of Intelligent Data Analysis.

4 An Overview of the Book

This book presents the recent advances in the emerging field of **Reliable Knowledge Discovery from Data (RKDD)**. In this filed the knowledge is considered as reliable in the sense that its generalization performance can be set in advance. Hence, RKDD has a potential for a broad spectrum of applications, especially in critical domains like medicine, finance, military etc. The main material presented in the book is based on three consequent workshops Reliability Issues in Knowledge Discovery held in conjunction with the IEEE International Conferences on Data Mining (ICDM) in 2006, 2008, and 2010, respectively. In addition we provided an opportunity to authors to publish the results of their newest research related to RKDD.

This book is organized in seventeen chapters divided into four parts.

Part I includes three chapters on Reliability Estimation.

Chapter 1 provides an overview of typicalness and transductive reliability estimation frameworks. The overview is employed for introducing an approach for accessing reliability of individual classifications called joint confidence machine. Chapter 1 describes an approach that compensates the weaknesses of typicalness-based confidence estimation and transductive reliability estimation by integrating them into a joint confidence machine. It provides better interpretation of the performance of any classifiers. Experimental results performed with different machine learning algorithms in several problem domains show that there is no reduction of discrimination performance and is more suitable for applications with risk-sensitive problems with strict confidence limits.

Chapter 2 introduces new approaches to estimating and correcting individual predictions in the context of stream mining. It investigates the online reliability estimation of individual predictions. It proposes different strategies and explores techniques based on local variance and local bias, of local sensitivity analysis and online bagging of predictors. Comparison results on benchmark data are given to demonstrate the improvement of prediction accuracy.

Chapter 3 deals with the problem of quantifying the reliability in the context of neural networks. It elaborates on new approaches to estimation of confidence and prediction intervals for polynomial neural networks.

Part II includes seven chapters on Reliable Knowledge Discovery Methods.

Chapter 4 investigates outliers in regression targeting robust diagnostic regression. The chapter discusses both robust regression and regression diagnostics, presents several contemporary methods through numerical examples in linear regression.

Chapter 5 presents a conventional view on the definition of reliability; points out the three major categories of factors that affect the reliability of knowledge discovery, examined the impact of model complexity, weak links, varying sample sizes and the ability of different learners to the reliability of graphical model discovery, proposed reliable graph discovery approaches.

Chapter 6 provides a generalization of version spaces for reliable classification implemented using support vector machines.

Chapter 7 presents a unified generative model ONM which characterizes the life cycle of a ticket. The model uses maximum likelihood estimation to capture reliable ticket transfer profiles which can reflect how the information contained in a ticket is used by human experts to make reliable ticket routing decisions.

Chapter 8 applies the methods of aggregation functions for the reliable web based knowledge discovery from network trafic data.

Chapter 9 gives two new versions of SVM for the regression study of features in the problem domain. It provides means for feature selection and weighting based on the correlation analysis to give better and reliable result.

Chapter 10 describes in detail an application of transductive confidence machines for reliable handwriting recognition. It introduces a TCM framework which can enhance classifiers to reduce the computational costs and memory consumption required for updating the non-conformity scores in the offline learning setup of TCMs. Results are found to have outperformed previous methods on both relatively easy data and on difficult test samples.

PART III includes four Chapters on Reliability Analysis.

Chapter 11 addresses the problem of reliable feature selection. It introduces a generic-feature-selection measure together with a new search approach for globally optimal feature-subset selection. It discusses the reliability in the feature-selection process of a real pattern-recognition system, provides formal measurements and allows consistent search for relevant features in order to attain global optimal solution.

Chapter 12 provides three detailed case studies to show how the reliability of an induced classifier can be influenced. The case study results reveal the impact of data-oriented factors to the relaibility of the discovered knowledge.

Chapter 13 analyzes recently-introduced instance-based penalization techniques capable of providing more accurate predictions.

Chapter 14 investigates subsequence frequency measurement and its impact on the reliability of knowledge discovery in single sequences.

PART IV includes three chapters on Reliability Improvement Methods.

Chapter 15 proposed to use the inexact field learning method and parameter optimized one-class classifiers to improving reliability of unbalanced text mining by reducing performance bias.

Chapter 16 proposes a formal description technique for ontology representation and verification using a high level Petri net approach. It provides the capability of detection and identification of potential anomalies in ontology for the improvement of the discovered knolwedge.

Chapter 17 presents an UGDSS framework to provide reliable support for multicriteria decision making in uncertainty problem domain. It gives the system design and architecture.

5 Acknowledgement

We would like to thank many people that made this book possible. We start with the organizers of the workshops held in conjunction with the IEEE International Conferences on Data Mining (ICDM): Shusaku Tsumoto, Francesco Bonchi, Bettina Berendt, Wei Fan and Wynne Hsu. We express our gratitude to the authors whose contributions can be found in the book. Finally, we thank our colleagues from Springer that made the publication process possible in a short period.

Burwood Victoria (Australia), *Honghua Dai*
Hong Kong (China), *James Liu*
Maastricht (The Netherlands), *Evgueni Smirnov*
August 2011

Contents