KNOWLEDGE DISCOVERY AND DATA MINING

Knowledge Discovery and Data Mining

The Info-Fuzzy Network (IFN) Methodology

by

Oded Maimon
Tel-Aviv University,
Tel-Aviv, Israel

and

Mark Last
University of South Florida,
Tampa, FL, U.S.A.

Springer-Science+Business Media, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4419-4842-7 ISBN 978-1-4757-3296-2 (eBook)
DOI 10.1007/978-1-4757-3296-2

Printed on acid-free paper

Originally published by Kluwer Academic Publishers 2001.
Softcover reprint of the hardcover 1st edition 2001

To our families

Contents

List of Figures x

List of Tables xi

Acknowledgements xiii

Preface xv

Part I 1

INFORMATION-THEORETIC APPROACH TO KNOWLEDGE DISCOVERY 1

1 3

INTRODUCTION 3

1. Data explosion in the Internet Age 3
2. Knowledge discovery in databases (KDD) 4
3. Verification-Based Methods of Data Mining 6
4. Discovery-Oriented Data Mining 8
5. Feature Selection Methods 11
6. Learning issues 14
7. Information theory – the data mining perspective 16
8. Data Modelling 19
9. Book Organization 20

2 23

AUTOMATED DATA PRE-PROCESSING 23

1. Discretization of Ordinal Features 23
2. Static Discretization Algorithm 25
3. The Partitioning Procedure 26
4. Computational Complexity of the Static Algorithm 28
5. Static Discretization and Dimensionality Reduction 29

3 31

INFORMATION-THEORETIC CONNECTIONIST NETWORKS 31

1. A Unified Approach to Data Modelling 31

2. Constant Structure Information-Theoretic Networks....32
3. Multi-Layer Information-Theoretic Network....39
4. Dynamic Discretization of Ordinal Attributes....47
4 53
POST-PROCESSING OF DATA MINING RESULTS 53
1. Rule Extraction and Reduction....53
2. Prediction....55
3. From Local to Global Modelling....57

Part II 61

APPLICATION METHODOLOGY AND CASE STUDIES 61
5 63
METHODOLOGY OF APPLICATION 63
1. Overview of the Discovery Process....63
2. Understanding the Problem Domain....64
3. Obtaining and Understanding the Data....65
4. Preparation of the Data....66
5. Construction of the Knowledge Model from Data....68
6. Evaluation of the Model....69
7. Using the Model (Interpretation and Post-Processing)....70
6 71
CASE STUDIES 71
1. Design and Manufacturing....71
2. Education (Student Admission)....86
3. Health Care (Medical Diagnosis Database)....93

Part III 105

COMPARATIVE STUDY AND ADVANCED ISSUES 105
7 107
COMPARATIVE STUDY 107
1. Overview....107
2. Dimensionality Reduction....108
3. Prediction Power....110
4. Stability of Results....113
8 123
ADVANCED DATA MINING METHODS 123
1. Anytime Algorithm for Knowledge Discovery....123
2. Data reliability....129

9 135
SUMMARY AND SOME OPEN PROBLEMS 135
1. Method Benefits and Limitations 135
2. Future Research 140
APPENDIX A 145
1. Entropy H(X) 145
2. Joint Entropy and Conditional Entropy 145
3. Relative Entropy and Mutual Information 146
APPENDIX B 149
1. Breast Cancer Database 149
2. Chess Endgames 150
3. Credit Approval Database 152
4. Diabetes Database 153
5. Glass Identification Database 154
6. Heart Disease (Cleveland) Database 155
7. Iris Plants Database 156
8. Liver Database 157
9. Lung Cancer Database 158
10. Wine Database 160

Index 163

List of Figures

Figure 1. Taxonomy of Data Mining Methods 6
Figure 2. Information-Theoretic Network: Credit Dataset 50
Figure 3. Process Engineering Dataset - Distribution of Attribute H 85
Figure 4. Information-Theoretic Network: Iris Dataset 111
Figure 5. Average Training Errors 116
Figure 6. Average Testing Errors 116
Figure 7. Maximum Gaps - Training Errors 117
Figure 8. Maximum Gaps - Testing Errors 117
Figure 9. Average Gaps between Training and Testing Errors 118
Figure 10. Maximum Gaps - Number of Terminal Nodes 119
Figure 11. Maximum Gaps - Number of Predicting Attributes 120
Figure 12. Sets of Predicting Attributes 120
Figure 13. Fuzzy Information Gain as a function of MI, for three different values of b 125
Figure 14. Performance profile of the information-theoretic algorithm 128
Figure 15. Reliability Degree for Different Values of Beta 132

List of Tables

Table 1 The Relation Scheme - Work-in-Process 73
Table 2 Attribute Encoding - Work-in-Process 74
Table 3 Work-in-Process Dataset – Dimensionality Reduction Procedure 74
Table 4 WIP Dataset – Highest Positive Connection Weights 76
Table 5 WIP Dataset – Lowest Negative Connection Weights 77
Table 6 . Low Reliability Records - Work-in-Process 79
Table 7 WIP Reliability by Operation 79
Table 8 Process Engineering Dataset – Dimensionality Reduction 83
Table 9 Process Engineering Dataset – Rule Extraction (Attribute H) 85
Table 10 Dimensionality Reduction – Registration Dataset 90
Table 11 Prediction Power – Using Fano Inequality 93
Table 12 Target Attribute – Grouping of Diagnoses 96
Table 13 Medical Diagnosis Dataset – Dimensionality Reduction Procedure 97
Table 14 Discretization Scheme - *Age* 99
Table 15 Discretization Scheme - *Month* 99
Table 16 Ischaemic Heart Disease: Highest Positive Rule Weights 101
Table 17 Ischaemic Heart Disease: Lowest Negative Rule Weights 101
Table 18 Motor Vehicle Traffic Accidents: Highest Positive Rule Weights 101
Table 19 Motor Vehicle Traffic Accidents: Lowest Negative Rule Weights 102
Table 20 Low Reliability Records (Reliability < 0.1%) 103
Table 21 Dimensionality Reduction - Summary Table 109
Table 22 Prediction Power – Comparison to Other Methods 111
Table 23 Prediction Power – Using Fano Inequality 112

Table 24. Stability - List of Datasets 115
Table 25. Feature selection: summary of results 129
Table 26. The Relational schema - Wisconsin Breast Cancer Database 149
Table 27. Dimensionality Reduction Procedure - Wisconsin Breast Cancer Database 150
Table 28. The Relational schema – Chess Endgame 150
Table 29. Dimensionality Reduction Procedure – Chess Endgame 151
Table 30. The Relational schema - Australian Credit Approval 152
Table 31. Dimensionality Reduction Procedure – Credit Approval 153
Table 32. The Relational schema - Diabetes Database 153
Table 33. Dimensionality Reduction Procedure – Diabetes Database 154
Table 34. The Relational schema – Glass Identification Database 154
Table 35. Dimensionality Reduction Procedure – Glass Database 155
Table 36. The Relational schema – Heart Disease (Cleveland) Database 155
Table 37. Dimensionality Reduction Procedure – Heart Database 156
Table 38. The Relational schema – Iris Plants Database 156
Table 39. Dimensionality Reduction Procedure – Iris Database 157
Table 40. The Relational schema – Liver Database 157
Table 41. Dimensionality Reduction Procedure – Liver Database 158
Table 42. The Relational schema – Lung Cancer Database 158
Table 43. Dimensionality Reduction Procedure – Lung Cancer Database 160
Table 44. The Relational schema – Wine Database 160
Table 45. Dimensionality Reduction Procedure – Wine Database 161

Acknowledgements

We are thankful to the editorial staff of Kluwer Academic Publishers, especially John Martindale and Angela Quilici for their interest, helpfulness, and efficiency in bringing this project to a successful completion.

We would like to acknowledge our colleagues whose comments and suggestions have helped us in developing the IFN methodology. These include:
Professors Yishay Mansour, Tova Milo, and Irad Ben-Gal from Tel Aviv University;
Professor Abe Kandel from the University of South Florida;
Professor Martin Golumbic from Bar-Ilan University;
Dr. Abraham Meidan from Wizsoft;
Lior Rokach and Einat Minkov, graduate students from the IE Department at Tel-Aviv University.
We would also like to thank the Computing Division of the Ministry of Health in Israel and AVX Corporation for providing some of the databases for testing the IFN algorithm.

Preface

This book presents a specific and unified approach to Knowledge Discovery and Data Mining, termed IFN for Information Fuzzy Network methodology. Data Mining (DM) is the science of modelling and generalizing common patterns from large sets of multi-type data. DM is a part of KDD, which is the overall process for Knowledge Discovery in Databases. The accessibility and abundance of information today makes this a topic of particular importance and need.

The book has three main parts complemented by appendices as well as software and project data that are accessible from the book's web site (http://www.eng.tau.ac.il/~maimon/ifn-kdd/). Part I (Chapters 1-4) starts with the topic of KDD and DM in general and makes reference to other works in the field, especially those related to the information theoretic approach. The remainder of the book presents our work, starting with the IFN theory and algorithms. Part II (Chapters 5-6) discusses the methodology of application and includes case studies. Then in Part III (Chapters 7-9) a comparative study is presented, concluding with some advanced methods and open problems.

The IFN, being a generic methodology, applies to a variety of fields, such as manufacturing, finance, health care, medicine, insurance, and human resources. The appendices expand on the relevant theoretical background and present descriptions of sample projects (including detailed results). Finally, we refer the readers to the book's web site, where a copy of IFN program and data can be downloaded and experimented with. This is a "live" web site, meaning that we will update the program periodically and add more examples and case studies.

Data Mining has always been (under different names) of great interest to scientists. The existing methodologies of Data Mining can be historically categorized to five main approaches:

- Logic Based (for example Inductive Models)
- Classical Statistical (such as Regression Models and ANOVA)
- Non-Linear Classifiers (including Neural Networks and Pattern Recognition)
- Probabilistic (such as Bayesian Models)
- Information Theoretic (where IFN belongs)

All approaches are still being developed, and there are other taxonomies.

The challenge of Data Mining vis-à-vis the availability of large and dynamic data sets has led to the study of the KDD process, which includes the following main steps:

1. Data Pre-Processing (treating missing data and data cleansing)
2. Attribute Extraction (transformations and adding new features to the original data)
3. Feature Selection (trimming and identifying the most important features)
4. Data Mining (discovering patterns and rules)
5. Post Processing (assessing the importance of the rules and evaluating data reliability)

In developing the IFN, we have achieved two major goals, one in DM and one in KDD. In DM, the information-theoretic nature of the IFN is providing a quantitative trade-off to the major issue of generalization from data (finding the common patterns) versus specialization (recognizing that cases present different phenomena).

In the KDD process (see above), the IFN provides a unified approach to steps 3-5 that were traditionally treated by different methodologies. Steps 1-2 are problem specific and cannot be handled by a general approach. The IFN solves the feature selection problem, data mining, and post processing issues in a single run of the algorithm and with the same methodology (thus saving computational and modelling efforts).

In addition, the IFN provides models that are understandable, robust, and scalable. Understandability is provided by the weighted causality-type structure of the network. Robustness to noisy and incomplete data is achieved by a special built-in statistical significance testing. Scalability is apparent from analyzing the computational complexity of the algorithm, and it is confirmed by many experiments that show high classification accuracy along with remarkable stability of results.

The IFN method is designed as an *anytime* algorithm in the sense that it starts by revealing the most important features of the model and refines itself over time. Thus, the solution is of value given any type of time limitation,

which is important in time-constrained situations (subject to accuracy threshold).

IFN can handle datasets of mixed nature, including numerical, binary and categorical (non ordinal) data. Discretization of continuous attributes is performed automatically to maximize the information gain.

One of IFN's leading features is the attribute reduction. The experiments with IFN show that in most cases less then 10 ranked attributes affect a target. The importance of this result is that with so few attributes, phenomena can be understood and analyzed as a physical law. The stability of the IFN method allows the rules to stay the same with minor changes in the training set unless the dynamics of the data represents underlying phenomena changes.

This book is only a starting point for further development of the theory and the applications based on the IFN methodology. Applications can include efficient data warehouse design, queries in large distributed databases, personalization, information security, and knowledge extraction to personal communication devices.

This book can be used by researchers in the fields of information systems, engineering (especially industrial and electrical), computer science, statistics and management, who are searching for a unified theoretical approach to the KDD process. In addition, social sciences, psychology, medicine, genetics, and other fields that are interested in understanding the underlying phenomena from data can much benefit from the IFN approach. The book can also serve as a 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 KDD projects performed with IFN.

We hope you will enjoy the book, learn from it, and then share your ideas with us as you explore the fascinating topic of knowledge discovery. We invite you to continue the interaction by staying tuned to the book's web site.

Oded Maimon and Mark Last

Tel Aviv, May 2000
{maimon@eng.tau.ac.il}{ mlast@csee.usf.edu}