

MASSIVE COMPUTING

KNOWLEDGE DISCOVERY AND DATA MINING

The Info-Fuzzy Network (IFN) Methodology

Oded Maimon and Mark Last

KLUWER ACADEMIC PUBLISHERS

Knowledge Discovery and Data Mining

The Info-Fuzzy Network (IFN) Methodology

by

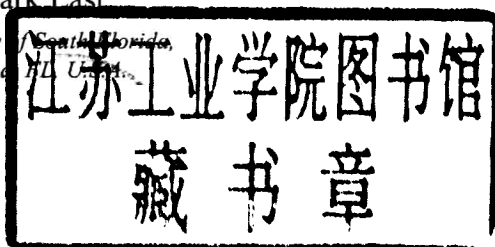
Oded Maimon

*Tel-Aviv University,
Tel-Aviv, Israel*

and

Mark Last

*University of South Florida,
Tampa, FL, U.S.A.*



KLUWER ACADEMIC PUBLISHERS

DORDRECHT / BOSTON / LONDON

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

ISBN 0-7923-6647-6

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved
© 20001 Kluwer Academic Publishers
No part of the material protected by this copyright notice may be reproduced or
utilized in any form or by any means, electronic or mechanical,
including photocopying, recording or by any information storage and
retrieval system, without written permission from the copyright owner.

Printed in the Netherlands.

KNOWLEDGE DISCOVERY AND DATA MINING

To our families

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 varied types of input data. DM is a part of KDD, which is the overall process for Knowledge Discovery in Data. 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 Internet site (<http://wwwie.eng.tau.ac.il/ifn-kdd/index.htm>). 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 theory and algorithms of the IFN. 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 needed theoretical background and present descriptions of the 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 led to the field of KDD, 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 validated by many empirical tests that also show high classification accuracy along with remarkable stability of results.

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

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

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 the leading features of the IFN is the attribute reduction. The experiments with IFN show that in most cases less than 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 and 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}

Contents

List of Figures	xi
List of Tables	xii
Acknowledgements	xv
Preface	xvii
Part I: Information-Theoretic Approach to Knowledge Discovery	1
1. Introduction	3
1.1 Data explosion in the Internet Age	3
1.2 Knowledge discovery in databases (KDD)	4
1.3 Verification-Based Methods of Data Mining	6
1.4 Discovery-Oriented Data Mining	8
1.5 Feature Selection Methods	11
1.6 Learning issues	14
1.7 Information theory – the data mining perspective	16
1.8 Data Modelling	19
1.9 Book Organization	20
2. Automated data pre-processing	23
2.1 Discretization of Ordinal Features	23
2.2 Static Discretization Algorithm	25
2.3 The Partitioning Procedure	26
2.4 Computational Complexity of the Static Algorithm	28
2.5 Static Discretization and Dimensionality Reduction	29
3. Information-Theoretic Connectionist Networks	31
3.1 A Unified Approach to Data Modelling	31
3.2 Constant Structure Information-Theoretic Networks	32
3.3 Multi-Layer Information-Theoretic Network	39
3.4 Dynamic Discretization of Ordinal Attributes	47
4. Post-Processing of Data Mining Results	53
4.1 Rule Extraction and Reduction	53
4.2 Prediction	55
4.3 From Local to Global Modelling	57

Part II: Application Methodology and Case Studies	61
5. Methodology of Application	63
5.1 Overview of the Discovery Process	63
5.2 Understanding the Problem Domain	64
5.3 Obtaining and Understanding the Data	64
5.4 Preparation of the Data	66
5.5 Construction of the Knowledge Model from Data	68
5.6 Evaluation of the Model	69
5.7 Using the Model (Interpretation and Post-Processing)	69
6. Case Studies	71
6.1 Design and Manufacturing	71
6.2 Education (Student Admission)	86
6.3 Health Care (Medical Diagnosis Database)	93
Part III: Comparative Study and Advanced Issues	105
7. Comparative Study	107
7.1 Overview	107
7.2 Dimensionality Reduction	108
7.3 Prediction Power	110
7.4 Stability of Results	113
8. Advanced data mining methods	125
8.1 Anytime Algorithm for Knowledge Discovery	125
8.2 Data reliability	131
9. Summary and Some Open Problems	137
9.1 Method Benefits and Limitations	137
9.2 Future Research	144
References	145
Appendix A	149
A.1 Entropy $H(X)$	149
A.2 Joint Entropy and Conditional Entropy	149
A.3 Relative Entropy and Mutual Information	150
Appendix B	153
B.1 Breast Cancer Database	153
B.2 Chess Endgames	154
B.3 Credit Approval Database	156
B.4 Diabetes Database	157

B.5	Glass Identification Database	158
B.6	Heart Disease (Cleveland) Database	159
B.7	Iris Plants Database	160
B.8	Liver Database	161
B.9	Lung Cancer Database	162
B.10	Wine Database	164
Index		165

List of Figures

<i>Figure 1. Taxonomy of Data Mining Methods</i>	6
<i>Figure 2. Information-Theoretic Network: Credit Dataset</i>	51
<i>Figure 3. Process Engineering Dataset - Distribution of Attribute H</i>	85
<i>Figure 4. Information-Theoretic Network: Iris Dataset</i>	112
<i>Figure 5. Average Training Errors</i>	117
<i>Figure 6. Average Testing Errors</i>	117
<i>Figure 7. Maximum Gaps - Training Errors</i>	118
<i>Figure 8. Maximum Gaps - Testing Errors</i>	119
<i>Figure 9. Average Gaps between Training and Testing Errors</i>	120
<i>Figure 10. Maximum Gaps - Number of Terminal Nodes</i>	120
<i>Figure 11. Maximum Gaps - Number of Predicting Attributes</i>	121
<i>Figure 12. Sets of Predicting Attributes</i>	122
<i>Figure 13. Fuzzy Information Gain as a function of MI, for three different values of b</i>	128
<i>Figure 14. Performance profile of the information-theoretic algorithm</i>	130
<i>Figure 15. Reliability Degree for Different Values of Beta</i>	135

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 Database – Dimensionality Reduction Procedure	74
Table 4 WIP Database – Highest Positive Connection Weights.....	76
Table 5 WIP Database – 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 Database – Dimensionality Reduction.....	83
Table 9 Process Engineering Database – Rule Extraction (Attribute H).....	85
Table 10 Dimensionality Reduction – Registration Database	90
Table 11 Prediction Power – Using Fano Inequality.....	93
Table 12 Target Attribute – Grouping of Diagnoses.....	96
Table 13 Medical Diagnosis Database – Dimensionality Reduction Procedure	98
Table 14 Discretization Scheme - <i>Age</i>	99
Table 15 Discretization Scheme - <i>Month</i>	100
Table 16 Ischaemic Heart Disease: Highest Positive Rule Weights	102
Table 17 Ischaemic Heart Disease: Lowest Negative Rule Weights	102
Table 18 Motor Vehicle Traffic Accidents: Highest Positive Rule Weights	102
Table 19 Motor Vehicle Traffic Accidents: Lowest Negative Rule Weights	103
Table 20 Low Reliability Records (Reliability < 0.1%).....	104
Table 21 Dimensionality Reduction - Summary Table	110
Table 22 Prediction Power – Comparison to Other Methods	111
Table 23 Prediction Power – Using Fano Inequality.....	113
Table 24. Stability - List of Datasets.....	116

Table 25. Feature selection: summary of results.....	131
Table 26. The Relational schema - Wisconsin Breast Cancer Database	153
Table 27. Dimensionality Reduction Procedure - Wisconsin Breast Cancer Database.....	154
Table 28. The Relational schema – Chess Endgame.....	154
Table 29. Dimensionality Reduction Procedure – Chess Endgame	155
Table 30. The Relational schema - Australian Credit Approval	156
Table 31. Dimensionality Reduction Procedure – Credit Approval.....	157
Table 32. The Relational schema - Diabetes Database.....	157
Table 33. Dimensionality Reduction Procedure – Diabetes Database	157
Table 34. The Relational schema – Glass Identification Database	158
Table 35. Dimensionality Reduction Procedure – Glass Database	159
Table 36. The Relational schema – Heart Disease (Cleveland) Database..	159
Table 37. Dimensionality Reduction Procedure – Heart Database	160
Table 38. The Relational schema – Iris Plants Database.....	160
Table 39. Dimensionality Reduction Procedure – Iris Database	161
Table 40. The Relational schema – Liver Database.....	161
Table 41. Dimensionality Reduction Procedure – Liver Database	161
Table 42. The Relational schema – Lung Cancer Database	162
Table 43. Dimensionality Reduction Procedure – Lung Cancer Database.	163
Table 44. The Relational schema – Wine Database	164
Table 45. Dimensionality Reduction Procedure – Wine Database	164

PART I

INFORMATION-THEORETIC APPROACH TO
KNOWLEDGE DISCOVERY

