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

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 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 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

List of Tables

PART I

INFORMATION-THEORETIC APPROACH TO KNOWLEDGE DISCOVERY