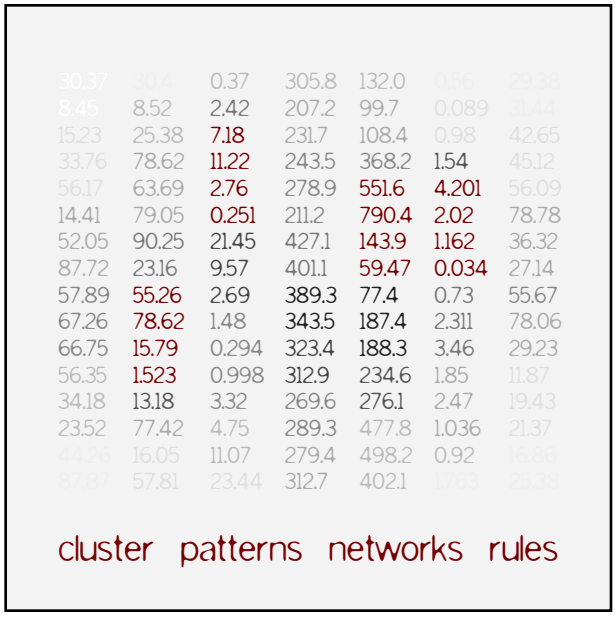
Johann-Adolf Mueller
Frank Lemke

Self-organising Data Mining

An Intelligent Approach To Extract Knowledge From Data

Theory
Real World Applications
Software
Future Technologies



Mueller, Johann-Adolf; Lemke, Frank:
Self-Organising Data Mining.
An Intelligent Approach To Extract Knowledge From Data.
Berlin , Dresden 1999
1. Edition

Contents

Preface

1 Knowledge Discovery from Data *9*

1.1 Models and their application in decision making
1.2 Relevance and value of forecasts
1.3 Theory driven approach
1.4 Data driven approach
1.5 Data mining
References

2 Self-organising Data Mining *31*

2.1 Involvement of users in the data mining process
2.2 Automatic model generation
• regression based models • rule based modelling
• symbolic modelling • nonparametric models
2.3 Self-organising data mining
References

3 Self-organising Modelling Technologies *57*

3.1 Statistical Learning Networks
3.2 Inductive approach - The GMDH algorithm
• induction • principles • model of optimal complexity
References

4 Parametric GMDH Algorithms *77*

4.1 Elementary models (neurons)
4.2 Generation of alternate model variants
4.3 Nets of active neurons
4.4 Criteria of model selection
4.5 Validation
References

5 Nonparametric Algorithms *103*

5.1 Objective Cluster Analysis
5.2 Analog Complexing
5.3 Self-organising Fuzzy Rule Induction
5.4 Logic based rules
References

6 Application of Self-organising Data Mining ***125***

6.1 Spectrum of self-organising data mining methods

6.2 Choice of appropriate modelling methods

6.3 Application fields

6.4 Synthesis

6.5 Software tools

References

7 KnowledgeMiner ***147***

7.1 General features

7.2 GMDH implementation
• elementary models and active neurons • generation of alternate model variants • criteria of model selection • systems of equations

7.3 Analog Complexing implementation
• features • example

7.4 Fuzzy Rule Induction implementation
• fuzzification • rule induction • defuzzification • example

7.5 Using models
• the model base • finance module

8 Sample Applications ***177***

8.1 ... From Economics
• national economy • stock prediction • balance sheet
• sales prediction • solvency checking • energy consumption

8.2 ... From Ecology
• water pollution • water quality

8.3 ... From other Fields
• heart disease • U.S. congressional voting behavior

References

This book is dedicated to
Prof. A.G. Ivakhnenko,
the father of GMDH,
to his eighty fifth' birthday

Preface

The rapid development of information technology, continuing computerization in almost every field of human activity and distributed computing has led to a flood of data stored in data bases and data warehouses. In the 1960s, Management Information Systems (MIS) and then, in the 1970s, Decision Support Systems (DSS) were praised for their potential to supply executives with mountains of data needed to carry out their jobs. While these systems have supplied some useful information for executives, they have not lived up to their proponents' expectations. They simply supplied too much data and not enough information to be generaly useful.

Today, there is an increased need for information - contextual data - non obvious and valuable for decision making from a large collection of data. This is an interactive and iterative process of various subtasks and decisions and is called Knowledge Discovery from Data. The engine of Knowledge Discovery - where data is transformed into knowledge for decision making - is Data Mining.

There are very different data mining tools available and many papers are published describing data mining techniques. We think that it is most important for a more sophisticated data mining technique to limit the user involvement in the entire data mining process to the inclusion of well-known a priori knowledge. This makes the process more automated and more objective. Most users' primary interest is in generating useful and valid model results without having to have extensive knowledge of mathematical, cybernetic and statistical techniques or sufficient time for complex dialog driven modelling tools. Soft computing, i.e., Fuzzy Modelling, Neural Networks, Genetic Algorithms and other methods of automatic model generation, is a way to mine data by generating mathematical models from empirical data more or less automatically.

In the past years there has been much publicity about the ability of Artificial Neural Networks to learn and to generalize despite important problems with design, development and application of Neural Networks:

- Neural Networks have no explanatory power by default to describe why results are as they are. This means that the knowledge (models) extracted by Neural Networks is still hidden and distributed over the network.
- There is no systematical approach for designing and developing Neural Networks. It is a trial-and-error process.
- Training of Neural Networks is a kind of statistical estimation often using algorithms that are slower and less effective than algorithms used in statistical software.
- If noise is considerable in a data sample, the generated models systematically tend to being overfitted.

In contrast to Neural Networks that use

- Genetic Algorithms as an external procedure to optimize the network architecture and
- several pruning techniques to counteract overtraining,

this book introduces principles of evolution - inheritance, mutation and selection - for generating a network structure systematically enabling automatic model structure synthesis and model validation. Models are generated from the data in the form of networks of active neurons in an evolutionary fashion of repetitive generation of populations of competing models of growing complexity and their validation and selection until an optimal complex model - not too simple and not too complex - has been created. That is, growing a tree-like network out of seed information (input and output variables' data) in an evolutionary fashion of pairwise combination and survival-of-the-fittest selection from a simple single individual (neuron) to a desired final, not overspecialized behavior (model). Neither, the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron is predefined. All this is adjusted during the process of self-organisation, and therefore, is called self-organising data mining.

A self-organising data mining creates optimal complex models systematically and autonomously by employing both parameter *and* structure identification. An optimal complex model is a model that optimally balances model quality on a given learning data set ("closeness of fit") and its generalisation power on new, not previously seen data with respect to the data's noise level and the task of modelling (prediction, classification, modelling, etc.). It thus solves the basic problem of experimental systems analysis of systematically avoiding "overfitted" models based on the data's information only. This makes self-organising data mining a most automated, fast and very efficient supplement and alternative to other data mining methods.

The differences between Neural Networks and this new approach focus on Statistical Learning Networks and induction. The first Statistical Learning Network algorithm of this new type, the Group Method of Data Handling (GMDH), was developed by A.G. Ivakhnenko in 1967. Considerable improvements were introduced in the 1970s and 1980s by versions of the Polynomial Network Training algorithm (PNETTR) by Barron and the Algorithm for Synthesis of Polynomial Networks (ASPN) by Elder when Adaptive Learning Networks and GMDH were flowing together. Further enhancements of the GMDH algorithm have been realized in the “KnowledgeMiner” software described and enclosed in this book.

KnowledgeMiner is a powerful and easy-to-use modelling and prediction tool designed to support the knowledge extraction process on a highly automated level and has implemented three advanced self-organising modelling technologies: GMDH, Analog Complexing and self-organising Fuzzy Rule Induction. There are three different GMDH modelling algorithms implemented - active neurons, enhanced network synthesis and creation of systems of equations - to make knowledge extraction systematically, fast and easy-to-use even for large and complex systems. The Analog Complexing algorithm is suitable for prediction of the most fuzzy processes like financial or other markets. It is a multidimensional search engine to select most similar past system states compared with a chosen (actual) reference state from a given data set. All selected patterns will be synthesized to a most likely, most optimistic and most pessimistic prediction. KnowledgeMiner does this in an objective way using GMDH finding out the optimal number of synthesized patterns and their composition. Fuzzy modelling is an approach to form a system model using a description language based on fuzzy logic with fuzzy predicates. Such a language can describe a dynamic multi-input/multi-output system qualitatively by means of a system of fuzzy rules.

Therefore, the generated models can be

- linear/nonlinear time series models,
- static/dynamic linear/nonlinear multi-input/single-output models,
- systems of linear/nonlinear difference equations (multi-input/multi-output models),
- systems of static/dynamic multi-input/multi-output fuzzy rules described analytically in all four cases, as well as

- nonparametric models obtained by Analog Complexing.

This book provides a thorough introduction to self-organising data mining technologies for business executives, decision makers and specialists involved in developing Executive Information Systems (EIS) or in modelling, data mining or knowledge discovery projects. It is a book for working professionals in many fields of decision making: Economics (banking, financing, marketing), business oriented computer science, ecology, medicine and biology, sociology, engineering sciences and all other fields of modelling of ill-defined systems.

Each chapter includes some practical examples and a reference list for further reading. The accompanying diskette/internet download contains the KnowledgeMiner Demo version and several executable examples. This book offers a comprehensive view to all major issues related to self-organising data mining and its practical application for solving real-world problems. It gives not only an introduction to self-organising data mining, but provides answers to questions like:

- what is self-organising data mining compared with other known data mining techniques,
- what are the pros, cons and difficulties of the main data mining approaches,
- what problems can be solved by self-organising data mining, specifically by using the KnowledgeMiner modelling and prediction tool,
- what is the basic methodology for self-organising data mining and application development using a set of real-world business problems exemplarily,
- how to use KnowledgeMiner and how to prepare a problem for solution.

The book spans eight chapters. Chapter 1 discusses several aspects of knowledge discovery from data as an introductory overview and understanding, such as why it is worth building models for decision support and how we think forecasting can be applied today to get valuable predictive control solutions. Also considered are the pros, cons and difficulties of the two main approaches of modelling: Theory-driven and data-driven modelling.

Chapter 2 explains the idea of a self-organising data mining and put it in context to several automated data-driven modelling approaches. The algorithm of a self-organising data mining is introduced and we describe how self-organisation works generally, what conditions it requires, and how existing theoretical knowledge can be embedded into the process.

Chapter 3 introduces and describes some important terms in self-organising modelling: Statistical Learning Networks, inductive approach, GMDH, nonphysical models, and model of optimal complexity.

Chapter 4 focuses on parametric regression based GMDH algorithms. Several algorithms on the principles of self-organisation are considered, and also the important problem of selection criteria choice and some model validation aspects are discussed.

In chapter 5, three nonparametric algorithms are discussed. First, there is the Objective Cluster Analysis algorithm that operates on pairs of closely spaced sample points. For the most fuzzy objects, the Analog Complexing algorithm is recommended selecting the most similar patterns from a given data set. Thirdly, a self-organising fuzzy-rule induction can help to describe and predict complex objects qualitatively.

In chapter 6 we want to point to some application opportunities of self-organising data mining from our own experience. Selected application fields and ideas on how a self-organising modelling approach can contribute to improve results of other modelling methods - simulation, Neural Networks and econometric modelling (statistics) - are suggested. Also included in this chapter is a discussion on a synthesis of model results, its goals and its options while the last part gives a short overview of existing self-organising data mining software.

In chapter 7 the KnowledgeMiner software is described in more detail to give the reader an understanding of its self-organising modelling implementations and to help examining the examples included in the accompanied diskette or Internet download.

Chapter 8 explains based on several sample applications from economics, ecology, medicine and sociology how it is possible to solve complex modelling, prediction, classification or diagnosis tasks systematically and fast using the knowledge extraction capabilities of a self-organising data mining approach.

Since self-organising data mining will evolve quickly, especially the KnowledgeMiner software, the following Internet addresses can be referenced to for news, updates and new versions, but also for new research results and other discussions and comments reflecting this book:

`http://www.knowledgeminer.net`

`http://www.informatik.htw-dresden.de/~muellerj .`

We would like to extend our thanks to our wives and our entire families for their encouragement and understanding during the writing of this book. A special thanks to Julian Miller from Script Software International for his supporting work and for his active promotion of the KnowledgeMiner software from the beginning. Thank you also to Russell Gum for proof reading the manuscript and to all persons who helped with their comments, suggestions and critics.

This electronic edition is a preprint of the book, and it serves all KnowledgeMiner users as a documentation and guide about theory and application of self-organising data mining. It may also form the basis for discussing these items in the KnowledgeMiner discussion forum:

`http://network54.com/Hide/Forum/goto?forumid=13476 .`

Comments and remarks are appreciated.

September 17, 1999

Johann-Adolf Mueller *Frank Lemke*