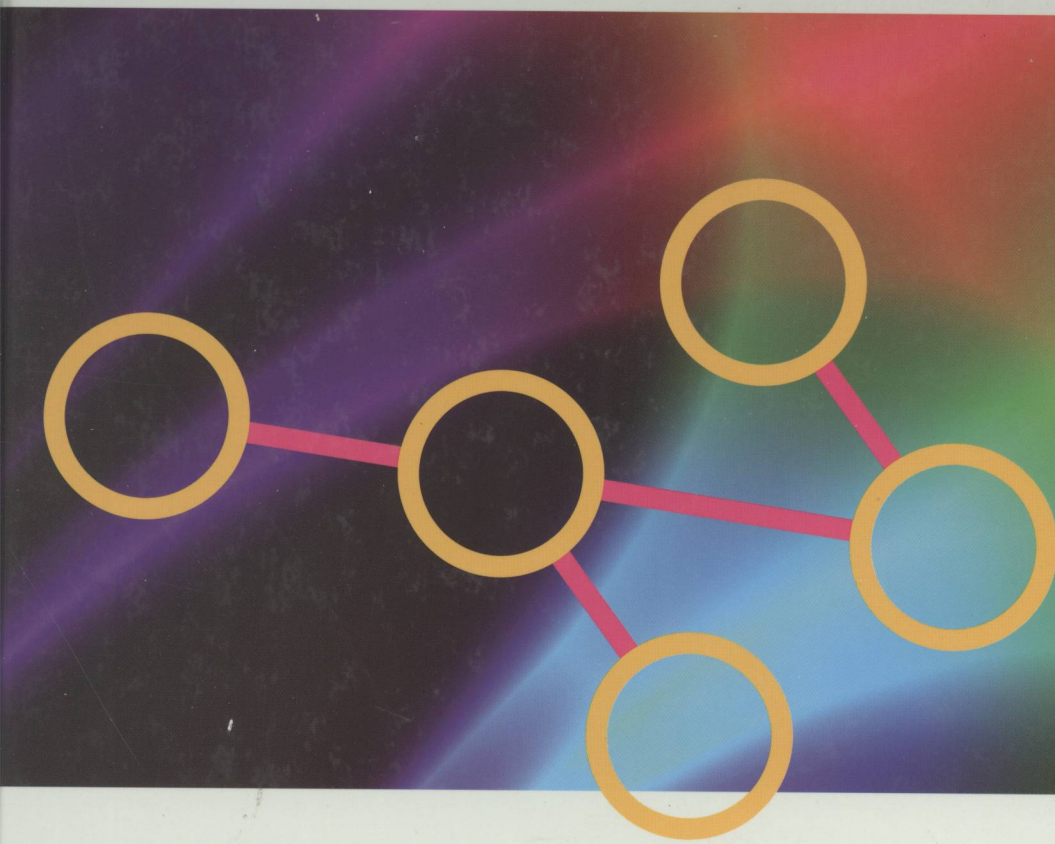
WILEY SERIES IN COMPUTATIONAL STATISTICS



Christian Borgelt,
Matthias Steinbrecher and Rudolf Kruse

GRAPHICAL MODELS

REPRESENTATIONS FOR LEARNING, REASONING AND DATA MINING, 2ND EDITION

WILEY

Graphical Models

Representations for Learning, Reasoning and Data Mining

Second Edition

Christian Borgelt
European Centre for Soft Computing, Spain

Matthias Steinbrecher & Rudolf Kruse
Otto-von-Guericke University Magdeburg, Germany

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data
Record on file

A catalogue record for this book is available from the British Library.

ISBN 978-0-470-72210-7

Typeset in 10/12 cmr10 by Laserwords Private Limited, Chennai, India
Printed in Great Britain by TJ International Ltd, Padstow, Cornwall

Graphical Models

Wiley Series in Computational Statistics

Wiley Series in Computational Statistics is comprised of practical guides and cutting edge research books on new developments in computational statistics. It features quality authors with a strong applications focus. The texts in the series provide detailed coverage of statistical concepts, methods and case studies in areas at the interface of statistics, computing, and numerics.

With sound motivation and a wealth of practical examples, the books show in concrete terms how to select and to use appropriate ranges of statistical computing techniques in particular fields of study. Readers are assumed to have a basic understanding of introductory terminology.

The series concentrates on applications of computational methods in statistics to fields of bioinformatics, genomics, epidemiology, business, engineering, finance and applied statistics.

Preface

Although the origins of graphical models can be traced back to the beginning of the 20th century, they have become truly popular only since the mid-eighties, when several researchers started to use Bayesian networks in expert systems. But as soon as this start was made, the interest in graphical models grew rapidly and is still growing to this day. The reason is that graphical models, due to their explicit and sound treatment of (conditional) dependences and independences, proved to be clearly superior to naive approaches like certainty factors attached to if-then-rules, which had been tried earlier.

Data Mining, also called Knowledge Discovery in Databases, is a another relatively young area of research, which has emerged in response to the flood of data we are faced with nowadays. It has taken up the challenge to develop techniques that can help humans discover useful patterns in their data. In industrial applications patterns found with these methods can often be exploited to improve products and processes and to increase turnover.

This book is positioned at the boundary between these two highly important research areas, because it focuses on learning graphical models from data, thus exploiting the recognized advantages of graphical models for learning and data analysis. Its special feature is that it is not restricted to probabilistic models like Bayesian and Markov networks. It also explores relational graphical models, which provide excellent didactical means to explain the ideas underlying graphical models. In addition, possibilistic graphical models are studied, which are worth considering if the data to analyze contains imprecise information in the form of sets of alternatives instead of unique values.

Looking back, this book has become longer than originally intended. However, although it is true that, as C.F. von Weizsäcker remarked in a lecture, anything ultimately understood can be said briefly, it is also evident that anything said too briefly is likely to be incomprehensible to anyone who has not yet understood completely. Since our main aim was comprehensibility, we hope that a reader is remunerated for the length of this book by an exposition that is clear and self-contained and thus easy to read.

Christian Borgelt, Matthias Steinbrecher, Rudolf Kruse
Oviedo and Magdeburg, March 2009

Contents

Chapter 1

Introduction

Due to modern information technology, which produces ever more powerful computers and faster networks every year, it is possible today to collect, transfer, combine, and store huge amounts of data at very low costs. Thus an ever-increasing number of companies and scientific and governmental institutions can afford to compile huge archives of tables, documents, images, and sounds in electronic form. The thought is compelling that if you only have enough data, you can solve any problem—at least in principle.

A closer examination reveals though, that data alone, however voluminous, are not sufficient. We may say that in large databases we cannot see the wood for the trees. Although any single bit of information can be retrieved and simple aggregations can be computed (for example, the average monthly sales in the Frankfurt area), general patterns, structures, and regularities usually go undetected. However, often these patterns are especially valuable, for example, because they can easily be exploited to increase turnover. For instance, if a supermarket discovers that certain products are frequently bought together, the number of items sold can sometimes be increased by appropriately arranging these products on the shelves of the market (they may, for example, be placed adjacent to each other in order to invite even more customers to buy them together, or they may be offered as a bundle).

However, to find these patterns and thus to exploit more of the information contained in the available data turns out to be fairly difficult. In contrast to the abundance of data there is a lack of tools to transform these data into useful knowledge. As John Naisbett remarked [Fayyad *et al.* 1996]:

> We are drowning in information, but starving for knowledge.

As a consequence a new area of research has emerged, which has been named *Knowledge Discovery in Databases (KDD)* or *Data Mining (DM)* and which has taken up the challenge to develop techniques that can help humans to discover useful patterns and regularities in their data.

Graphical Models: Representations for Learning, Reasoning and Data Mining, 2nd Edition
C. Borgelt, M. Steinbrecher and R. Kruse

In this introductory chapter we provide a brief overview on knowledge discovery in databases and data mining, which is intended to show the context of this book. In a first step, we try to capture the difference between "data" and "knowledge" in order to attain precise notions by which it can be made clear why it does not suffice just to gather data and why we must strive to turn them into knowledge. As an illustration we will discuss and interpret a well-known example from the history of science. Secondly, we explain the process of discovering knowledge in databases (the *KDD process*), of which data mining is just one, though very important, step. We characterize the standard data mining tasks and position the work of this book by pointing out for which tasks the discussed methods are well suited.

1.1 Data and Knowledge

In this book we distinguish between *data* and *knowledge*. Statements like "Columbus discovered America in 1492" or "Mrs Jones owns a VW Golf" are *data*. For these statements to qualify as data, we consider it to be irrelevant whether we already know them, whether we need these specific pieces of information at this moment, etc. For our discussion, the essential property of these statements is that they refer to single events, cases, objects, persons, etc., in general, to single instances. Therefore, even if they are true, their range of validity is very restricted and thus is their usefulness.

In contrast to the above, *knowledge* consists of statements like "All masses attract each other." or "Every day at 17:00 hours there runs an InterCity (a specific type of train of German Rail) from Magdeburg to Braunschweig." Again we neglect the relevance of the statement for our current situation and whether we already know it. The essential property is that these statements do not refer to single instances, but are general laws or rules. Therefore, provided they are true, they have a wide range of validity, and, above all else, they allow us to make predictions and thus they are very useful.

It has to be admitted, though, that in daily life statements like "Columbus discovered America in 1492." are also called knowledge. However, we disregard this way of using the term "knowledge", regretting that full consistency of terminology with daily life language cannot be achieved. Collections of statements about single instances do not qualify as knowledge.

Summarizing, data and knowledge can be characterized as follows:

Data

- refer to single instances
  (single objects, persons, events, points in time, etc.)
- describe individual properties
- are often available in huge amounts
  (databases, archives)

- are usually easy to collect or to obtain
  (for example cash registers with scanners in supermarkets, Internet)
- do not allow us to make predictions

Knowledge

- refers to *classes* of instances
  (*sets* of objects, persons, events, points in time, etc.)
- describes general patterns, structures, laws, principles, etc.
- consists of as few statements as possible
  (this is an objective, see below)
- is usually hard to find or to obtain
  (for example natural laws, education)
- allows us to make predictions

From these characterizations we can clearly see that usually knowledge is much more valuable than (raw) data. It is mainly the generality of the statements and the possibility to make predictions about the behavior and the properties of new cases that constitute its superiority.

However, not just any kind of knowledge is as valuable as any other. Not all general statements are equally important, equally substantial, equally useful. Therefore knowledge must be evaluated and assessed. The following list, which we do not claim to be complete, names some important criteria:

Criteria to Assess Knowledge

- correctness (probability, success in tests)
- generality (range of validity, conditions for validity)
- usefulness (relevance, predictive power)
- comprehensibility (simplicity, clarity, parsimony)
- novelty (previously unknown, unexpected)

In science correctness, generality, and simplicity (parsimony) are at the focus of attention: One way to characterize science is to say that it is the search for a minimal correct description of the world. In business and industry greater emphasis is placed on usefulness, comprehensibility, and novelty: the main goal is to get a competitive edge and thus to achieve higher profit. Nevertheless, none of the two areas can afford to neglect the other criteria.

Tycho Brahe and Johannes Kepler

Tycho Brahe (1546–1601) was a Danish nobleman and astronomer, who in 1576 and in 1584, with the financial support of Frederic II, King of Denmark and Norway, built two observatories on the island of Sen, about 32 km to

the north-east of Copenhagen. Using the best equipment of his time (telescopes were unavailable then—they were used only later by Galileo Galilei (1564–1642) and Johannes Kepler (see below) for celestial observations) he determined the positions of the sun, the moon, and the planets with a precision of less than one minute of arc, thus surpassing by far the exactitude of all measurements carried out earlier. He achieved in practice the theoretical limit for observations with the unaided eye. Carefully he recorded the motions of the celestial bodies over several years [Greiner 1989, Zey 1997].

Tycho Brahe gathered data about our planetary system. Huge amounts of data—at least from a 16th century point of view. However, he could not discern the underlying structure. He could not combine his data into a consistent scheme—to some extent, because be adhered to the geocentric system. He could tell exactly in what position Mars had been on a specific day in 1585, but he could not relate the positions on different days in such a way as to fit his highly accurate observational data. All his hypotheses were fruitless. He developed the so-called Tychonic planetary model, according to which the sun and the moon revolve around the earth, but all other planets revolve around the sun, but this model, though popular in the 17th century, did not stand the test of time. Today we may say that Tycho Brahe had a "data mining" or "knowledge discovery" problem. He had the necessary data, but he could not extract the knowledge contained in it.

Johannes Kepler (1571–1630) was a German astronomer and mathematician and assistant to Tycho Brahe. He advocated the Copernican planetary model, and during his whole life he endeavored to find the laws that govern the motions of the celestial bodies. He strove to find a mathematical description, which, in his time, was a virtually radical approach. His starting point were the catalogs of data Tycho Brahe had compiled and which he continued in later years. After several unsuccessful trials and long and tedious calculations, Johannes Kepler finally managed to condense Tycho Brahe's data into three simple laws, which have been named after him. Having discovered in 1604 that the course of Mars is an ellipse, he published the first two laws in "Astronomia Nova" in 1609, the third ten years later in his principal work "Harmonica Mundi" [Feynman *et al.* 1963, Greiner 1989, Zey 1997].

1. Each planet moves around the sun on an elliptical course, with the sun at one focus of the ellipse.
2. The radius vector from the sun to the planet sweeps out equal areas in equal intervals of time.
3. The squares of the periods of any two planets are proportional to the cubes of the semi-major axes of their respective orbits: $T \sim a^{\frac{3}{2}}$.

Tycho Brahe had collected a large amount of celestial data, Johannes Kepler found the laws by which they can be explained. He discovered the hidden knowledge and thus became one of the most famous "data miners" in history.

Today the works of Tycho Brahe are almost forgotten. His catalogs are merely of historical value. No textbook on astronomy contains extracts from his measurements. His observations and minute recordings are raw data and thus suffer from a decisive disadvantage: They do not provide us with any insight into the underlying mechanisms and therefore they do not allow us to make predictions. Kepler's laws, however, are treated in all textbooks on astronomy and physics, because they state the principles that govern the motions of planets as well as comets. They combine all of Brahe's measurements into three fairly simple statements. In addition, they allow us to make predictions: If we know the position and the velocity of a planet at a given moment, we can compute, using Kepler's laws, its future course.

1.2 Knowledge Discovery and Data Mining

How did Johannes Kepler discover his laws? How did he manage to extract from Tycho Brahe's long tables and voluminous catalogs those simple laws that revolutionized astronomy? We know only fairly little about this. He must have tested a large number of hypotheses, most of them failing. He must have carried out long and complicated computations. Presumably, outstanding mathematical talent, tenacious work, and a considerable amount of good luck finally led to success. We may safely guess that he did not know any universal method to discover physical or astronomical laws.

Today we still do not know such a method. It is still much simpler to gather data, by which we are virtually swamped in today's "information society" (whatever that means), than to obtain knowledge. We even need not work diligently and perseveringly any more, as Tycho Brahe did, in order to collect data. Automatic measurement devices, scanners, digital cameras, and computers have taken this load from us. Modern database technology enables us to store an ever-increasing amount of data. It is indeed as John Naisbett remarked: We are drowning in information, but starving for knowledge.

If it took such a distinguished mind like Johannes Kepler several years to evaluate the data gathered by Tycho Brahe, which today seem to be negligibly few and from which he even selected only the data on the course of Mars, how can we hope to cope with the huge amounts of data available today? "Manual" analysis has long ceased to be feasible. Simple aids like, for example, representations of data in charts and diagrams soon reach their limits. If we refuse to simply surrender to the flood of data, we are forced to look for intelligent computerized methods by which data analysis can be automated at least partially. These are the methods that are sought for in the research areas called *Knowledge Discovery in Databases (KDD)* and *Data Mining (DM)*. It is true, these methods are still very far from replacing people like Johannes Kepler, but it is not entirely implausible that he, if supported by these methods, would have reached his goal a little sooner.

Often the terms *Knowledge Discovery* and *Data Mining* are used interchangeably. However, we distinguish them here. By *Knowledge Discovery in Databases (KDD)* we mean a process consisting of several steps, which is usually characterized as follows [Fayyad *et al.* 1996]:

> Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

One step of this process, though definitely one of the most important, is *Data Mining.* In this step modeling and discovery techniques are applied.

1.2.1 The KDD Process

In this section we structure the KDD process into two preliminary and five main steps or phases. However, the structure we discuss here is by no means binding: it has proven difficult to find a single scheme that everyone in the scientific community can agree on. However, an influential suggestion and detailed exposition of the KDD process, which is close to the scheme presented here and which has had considerable impact, because it is backed by several large companies like NCR and DaimlerChrysler, is the CRISP-DM model (CRoss Industry Standard Process for Data Mining) [Chapman *et al.* 1999].

Preliminary Steps

- estimation of potential benefit
- definition of goals, feasibility study

Main Steps

- check data availability, data selection, if necessary, data collection
- preprocessing (usually 60–90% of total overhead)
  - unification and transformation of data formats
  - data cleaning
    (error correction, outlier detection, imputation of missing values)
  - reduction / focusing
    (sample drawing, feature selection, prototype generation)
- **Data Mining** (using a variety of methods)
- visualization
  (also in parallel to preprocessing, data mining, and interpretation)
- interpretation, evaluation, and test of results
- deployment and documentation

The preliminary steps mainly serve the purpose to decide whether the main steps should be carried out. Only if the potential benefit is high enough and the demands can be met by data mining methods, can it be expected that some profit results from the usually expensive main steps.

In the main steps the data to be analyzed for hidden knowledge are first collected (if necessary), appropriate subsets are selected, and they are transformed into a unique format that is suitable for applying data mining techniques. Then they are cleaned and reduced to improve the performance of the algorithms to be applied later. These preprocessing steps usually consume the greater part of the total costs. Depending on the data mining task that was identified in the goal definition step (see below for a list), data mining methods are applied (see farther below for a list), the results of which, in order to interpret and evaluate them, can be visualized. Since the desired goal is rarely achieved in the first go, usually several steps of the preprocessing phase (for example feature selection) and the application of data mining methods have to be reiterated in order to improve the result. If it has not been obvious before, it is clear now that KDD is an interactive process, rather than completely automated. A user has to evaluate the results, check them for plausibility, and test them against hold-out data. If necessary, he/she modifies the course of the process to make it meet his/her requirements.

1.2.2 Data Mining Tasks

In the course of time typical tasks have been identified, which data mining methods should be able to solve (although, of course, not every single method is required to be able to solve all of them—it is the combination of methods that makes them powerful). Among these are especially those named in the—surely incomplete—list below. We tried to characterize them not only by their name, but also by a typical question [Nakhaeizadeh 1998b].

- classification
  Is this customer credit-worthy?
- segmentation, clustering
  What groups of customers do I have?
- concept description
  Which properties characterize fault-prone vehicles?
- prediction, trend analysis
  What will the exchange rate of the dollar be tomorrow?
- dependence/association analysis
  Which products are frequently bought together?
- deviation analysis
  Are there seasonal or regional variations in turnover?