LEHRBUCH

Thomas A. Runkler

Data Analytics

Models and Algorithms for Intelligent Data Analysis

Springer Vieweg

Data Analytics

Thomas A. Runkler

Data Analytics

Models and Algorithms
for Intelligent Data Analysis

Springer Vieweg

Prof. Dr. Thomas A. Runkler
Siemens AG, München und Technische Universität München
Germany

ISBN 978-3-8348-2588-9 ISBN 978-3-8348-2589-6 (eBook)
DOI 10.1007/978-3-8348-2589-6

Library of Congress Control Number: 2012912256

Vieweg+Teubner

Printed on acid-free paper

Springer Vieweg is part of Springer Science+Business Media (www.springer-vieweg.de)

Preface

The information in the world doubles every 20 months. Important data sources are business and industrial processes, text and structured data bases, image and biomedical data. Many applications show that data analytics can provide huge benefits. We need models and algorithms to collect, preprocess, analyze, and evaluate data, from various fields such as statistics, system theory, machine learning, pattern recognition, or computational intelligence. With this book you will learn about the most important methods and algorithms for data analytics. You will be able to choose appropriate methods for specific tasks and apply these in your own data analytics projects. You will understand the basic concepts of the growing field of data analytics, which will allow you to keep pace and to actively contribute to the advancement of the field.

This text is designed for undergraduate and graduate courses on data analytics for engineering, computer science, and math students. It is also suitable for practitioners working on data analytics projects. The book is structured according to typical practical data analytics projects. Only basic mathematics is required. This is the third edition of a book that has been used for more than ten years in numerous courses at the Technical University of Munich, Germany, in short courses at several other universities, and in tutorials at international scientific conferences. Much of the content is based on the results of industrial research and development projects at Siemens.

The author wishes to express his sincere appreciation to everybody who has supported this work, in particular Wilfried Brauer, Jim Bezdek, Javier Esparza, Ralph Grothmann, Hans Hellendoorn, Jürgen Hollatz, Eyke Hüllermeyer, Uzay Kaymak, Jim Keller, Frank Klawonn, Rudolf Kruse, Rainer Palm, Bernd Schürmann, João Sousa, and Hans-Georg Zimmermann, the reviewers and students for pointing out errors and suggesting improvements, and the editorial and publisher team for their professional collaboration.

Munich, August 2012 *Thomas Runkler*

Contents

Chapter 1
Introduction

Abstract This book deals with models and algorithms for the analysis of data sets, for example industrial process data, business data, text and structured data, image data, and biomedical data. We define the terms data analytics, data mining, knowledge discovery, and the KDD and CRISP-DM processes. Typical data analysis projects can be divided into several phases: preparation, preprocessing, analysis, and postprocessing. The chapters of this book are structured according to the main methods of data preprocessing and data analysis: data and relations, data preprocessing, visualization, correlation, regression, forecasting, classification, and clustering.

1.1 It's All About Data

The focus of this book is the analysis of large data sets, for example:

- Industrial process data: An increasing amount of data is acquired, stored and processed in order to automate and control industrial production, manufacturing, distribution, logistics and supply chain processes. Data are used on all levels of the automation pyramid: sensors and actuators at the field level, control signals at the control level, operation and monitoring data at the execution level, schedules and indicators at the planning level. The main purpose of data analysis in industry is to optimize processes and to improve the competitive position of the company.
- Business data: Data of business performance are analyzed to better understand and drive business processes. Important business domains to be analyzed include customers, portfolio, sales, marketing, pricing, financials, risk, and fraud. An example is shopping basket analysis that finds out which products customers purchase at the same time. This analysis aims to improve cross selling and thus increases sales. Another example for business data analysis is customer segmentation for tailored advertising and sales promotions.
- Text and structured data: The analysis of numerical data has been the focus of mathematical statistics for centuries. Today, text and structured data also serve

T. A. Runkler, *Data Analytics*, DOI 10.1007/978-3-8348-2589-6_1,

as important information sources: text documents, electronic messages (like e-mail), web documents, or web based data bases (the so-called deep web). The analysis of text and structured data helps to filter, search, extract, and structure information. Structured data (as opposed to unstructured text) uses particular organizational criteria such as record fields or object structures which are often enhanced by semantic and ontological models.

- Image data: An increasing number of image sensors ranging from smartphone cameras to satellite cameras provides large amounts of 2D and also 3D image data. Image data analysis finds and recognizes objects, analyzes and classifies scenes, and relates image data with other information sources.
- Biomedical data: Data from laboratory experiments are used to analyze, understand and exploit biological and medical processes, for example to analyze DNA sequences, to understand and annotate genome functions, to analyze gene and protein expressions or to model regulation networks.

1.2 Data Analytics, Data Mining, and Knowledge Discovery

The term *data mining* dates back to the 1980s [3]. The goal of data mining is to extract knowledge from data [1]. In this context, *knowledge* is defined as *interesting* patterns that are generally valid, novel, useful, and understandable to humans. Whether or not the extracted patterns are interesting depends on the particular application and needs to be verified by application experts. Based on expert feedback the knowledge extraction process is often interactively refined. The term *data analytics* became popular in the early 2000s [2, 6]. Data analytics is defined as the application of computer systems to the analysis of large data sets for the support of decisions. Data analytics is a very interdisciplinary field that has adopted aspects from many other scientific disciplines such as statistics, signal theory, pattern recognition, computational intelligence, machine learning, and operations research.

Typical data analysis projects can be divided into several phases. Data are assessed and selected, cleaned and filtered, visualized and analyzed, and the analysis results are finally interpreted and evaluated. The *knowledge discovery in databases* (KDD) process [1] comprises the six phases selection, preprocessing, transformation, data mining, interpretation, and evaluation. The *cross industry standard process for data mining* (CRISP-DM) [5] comprises the six phases business understanding, data understanding, data preparation, modeling, evaluation, and deployment. For simplicity we distinguish only four phases here: preparation, preprocessing, analysis, and postprocessing (Fig. 1.1). The main focus of this book is data preprocessing and data analysis. The chapters are structured according to the main methods of preprocessing and analysis: data and relations, data preprocessing, visualization, correlation, regression, forecasting, classification, and clustering.

This book gives a clear and concise overview of the most important methods and algorithms of data analysis. It enables the reader to gain a complete and compre-

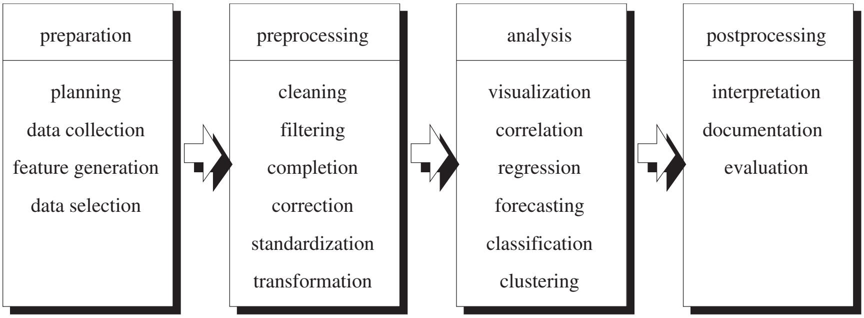

Fig. 1.1 Phases of data analysis projects.

hensive understanding of data analysis, to apply data analysis methods to his or her own projects, and to contribute to the progress of the field.

A large number a software tools for data mining are available today. Popular commercial software tools include MATLAB, SPSS, SAS, and STATISTICA. Popular free and open-source software tools include R, Rapid Miner, and WEKA. This book does *not* present, compare, or recommend any data mining software tools. For a comprehensive overview of current data mining software tools please refer to [4].

References

1. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, 1996.
2. R. Kohavi, N. J. Rothleder, and E. Simoudis. Emerging trends in business analytics. *Communications of the ACM*, 45(8):345–48, 2002.
3. M. C. Lovell. Data mining. *Review of Economics and Statistics*, 65(1):1–11, 1983.
4. R. Mikut and M. Reischl. Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):431–443, 2011.
5. C. Shearer. The CRISP-DM model: The new blueprint for data mining. *J Data Warehousing*, 5(4):13–22, 2000.
6. S. Tyagi. Using data analytics for greater profits. *Journal of Business Strategy*, 24(3):12–14, 2003.

Chapter 2
Data and Relations

Abstract The popular Iris benchmark set is used to introduce the basic concepts of data analysis. Data scales (nominal, ordinal, interval, ratio) must be accounted for because certain mathematical operations are only appropriate for specific scales. Numerical data can be represented by sets, vectors, or matrices. Data analysis is often based on dissimilarity measures (like inner product norms, Lebesgue/Minkowski norms) or on similarity measures (like cosine, overlap, Dice, Jaccard, Tanimoto). Sequences can be analyzed using sequence relations (like Hamming, Levenshtein, edit distance). Data can be extracted from continuous signals by sampling and quantization. The Nyquist condition allows sampling without loss of information.

2.1 The Iris Data Set

To introduce the basic concepts of data analysis we consider one of the most popular historic benchmark data sets: the *Iris* data set [1]. The Iris data set was originally created in 1935 by the American botanist Edgar Anderson who examined the geographic distribution of Iris flowers on the Gaspé peninsula in Quebec (Canada). In 1936, Sir Ronald Aylmer Fisher used Anderson's Iris data set as an example for multivariate discriminant analysis [4] (see chapter 8). Subsequently, the Iris data set became one of the most frequently used reference data set in statistics and data analysis.

The Iris data set comprises measurements from 150 Iris flower samples: 50 from each of the three species Iris Setosa, Iris Virginica, and Iris Versicolor. For each of the 150 flowers, values of four numerical features chosen by Anderson were measured: the length and the width of sepal and petal leaves in centimeters. For illustration, Table 2.1 shows the complete Iris data set. Notice that several *distinct* replicates of the original Iris data set have been used and published, so in experiments with this data set the version should be carefully checked [2]. The Iris data set as well as many other popular data sets are available, for example, through the machine learning data base at the University of California at Irvine (UCI).

T. A. Runkler, *Data Analytics*, DOI 10.1007/978-3-8348-2589-6_2,

Table 2.1 The Iris data set (from [1])

Setosa				Versicolor				Virginica			
sepal		petal		sepal		petal		sepal		petal	
length	width	length	width	length	width	length	width	length	width	length	width
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Table 2.2 Scales for numerical measurements.

scale	operations		example	statistics
nominal	$=$	$\neq$	Alice, Bob, Carol	mode
ordinal	$>$	$<$	A, B, C, D, F	median
interval	$+$	$-$	2015 A.D., $20^{\circ}C$	mean
ratio	$\cdot$	$/$	21 years, $273^{\circ}K$	generalized mean

In data analysis we call each of the 150 Iris flowers an *object*, each of the three species a *class*, and each of the four dimensions a *feature*. Here is a list of typical questions that we try to answer by data analysis:

- Which of the data might contain errors or false class assignments?
- What is the error caused by rounding the data off to one decimal place?
- What is the correlation between petal length and petal width?
- Which pair of dimensions is correlated most?
- None of the flowers in the data set has a sepal width of 1.8 centimeters. Which sepal length would we expect for a flower that did have 1.8 cm as its sepal width?
- Which species would an Iris with a sepal width of 1.8 centimeters belong to?
- Do the three species contain sub-species that can be identified from the data?

In this book you will find numerous methods and algorithms to answer these and other data analysis questions. For a better understanding of these data analysis methods and algorithms we first define and examine the fundamental properties of data and their relations.

2.2 Data Scales

Numerical measurements may have different semantic meanings, even if they are represented by the same numerical data. Depending on the semantic meaning different types of mathematical operations are appropriate. For the semantic meaning of numerical measurement Stevens [7] suggested the four different *scales* that are shown in Table 2.2. For nominal scaled data (first row) only tests for equality or inequality are valid. Examples for nominal features are names of persons or codes of objects. Data of a nominal feature can be represented by the *mode* which is defined as the value that occurs most frequently. For ordinal scaled data (second row) the operations ”greater than” and ”less than” are valid. For each scale level the operations and statistics of the lower scale levels are also valid, so for the ordinal scale we have equality, inequality, and the combinations ”greater than or equal” ($\geq$) and ”less than or equal” ($\leq$). The relation ”less than or equal” ($\leq$) defines a *total order*, such that for any x, y, z we have $(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$ (antisymmetry), $(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$ (transitivity), and $(x \leq y) \vee (y \leq x)$ (totality). Examples for ordinal features are school grades. Data of an ordinal feature can be represented by the *median* which is defined as the value for which (almost) as many smaller as