Intelligent Systems Reference Library

Volume 144

Series editors

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Lakhmi C. Jain, University of Canberra, Canberra, Australia;
Bournemouth University, UK;
KES International, UK
e-mail: jainlc2002@yahoo.co.uk; jainlakhmi@gmail.com
URL: http://www.kesinternational.org/organisation.php

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

More information about this series at http://www.springer.com/series/8578

Boris Kovalerchuk

Visual Knowledge Discovery and Machine Learning

Springer

Boris Kovalerchuk
Central Washington University
Ellensburg, WA
USA

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-319-73039-4 ISBN 978-3-319-73040-0 (eBook)
https://doi.org/10.1007/978-3-319-73040-0

Library of Congress Control Number: 2017962977

© Springer International Publishing AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my family

Preface

Emergence of Data Science placed knowledge discovery, machine learning, and data mining in multidimensional data, into the forefront of a wide range of current research, and application activities in computer science, and many domains far beyond it.

Discovering patterns, in multidimensional data, using a combination of visual and analytical machine learning means are an attractive visual analytics opportunity. It allows the injection of the unique human perceptual and cognitive abilities, directly into the process of discovering multidimensional patterns. While this opportunity exists, the long-standing problem is that we cannot see the n-D data with a naked eye. Our cognitive and perceptual abilities are perfected only in the 3-D physical world. We need enhanced visualization tools ("n-D glasses") to represent the n-D data in 2-D completely, without loss of information, which is important for knowledge discovery. While multiple visualization methods for the n-D data have been developed and successfully used for many tasks, many of them are non-reversible and lossy. Such methods *do not represent the n-D data fully* and do not allow the restoration of the n-D data completely from their 2-D representation. Respectively, our abilities to discover the n-D data patterns, from such incomplete 2-D representations, are limited and potentially erroneous. The number of available approaches, to overcome these limitations, is quite limited itself. The Parallel Coordinates and the Radial/Star Coordinates, today, are the most powerful reversible and lossless n-D data visualization methods, while suffer from occlusion.

There is a need to *extend* the class of *reversible and lossless n-D data visual representations*, for the knowledge discovery in the n-D data. A new class of such representations, called the **General Line Coordinate (GLC)** and several of their specifications, are the focus of this book. This book describes the GLCs, and their advantages, which include analyzing the data of the Challenger disaster, World hunger, semantic shift in humorous texts, image processing, medical computer-aided diagnostics, stock market, and the currency exchange rate predictions. Reversible methods for visualizing the n-D data have the advantages as *cognitive enhancers,* of the human cognitive abilities, to discover the n-D data patterns. This book reviews the state of the

art in this area, outlines the challenges, and describes the solutions in the framework of the General Line Coordinates.

This book expands the methods of the visual analytics for the knowledge discovery, by presenting the *visual and hybrid methods*, which combine the analytical *machine learning and the visual means*. New approaches are explored, from both the theoretical and the experimental viewpoints, using the modeled and real data. The inspiration, for a new large class of coordinates, is twofold. The first one is the marvelous success of the Parallel Coordinates, pioneered by Alfred Inselberg. The second inspiration is the absence of a "silver bullet" visualization, which is perfect for the pattern discovery, in the all possible n-D datasets. Multiple GLCs can serve as a collective "silver bullet." This multiplicity of GLCs increases the chances that the humans will reveal the hidden n-D patterns in these visualizations.

The topic of this book is related to the prospects of both the *super-intelligent machines* and the *super-intelligent humans,* which can far surpass the current human intelligence, significantly lifting the human cognitive limitations. This book is about a technical way for reaching some of the aspects of super-intelligence, which are beyond the current human cognitive abilities. It is to overcome the inabilities to analyze a large amount of abstract, numeric, and *high-dimensional* data; and to find the complex *pattern*s, in these data, with a *naked eye, supported by the analytical means of machine learning*. The new algorithms are presented for the reversible GLC visual representations of high-dimensional data and knowledge discovery. The advantages of GLCs are shown, both mathematically and using the different datasets. These advantages form a basis, for the future studies, in this super-intelligence area.

This book is organized as follows. Chapter 1 presents the goal, motivation, and the approach. Chapter 2 introduces the concept of the General Line Coordinates, which is illustrated with multiple examples. Chapter 3 provides the rigorous mathematical definitions of the GLC concepts along with the mathematical statements of their properties. A reader, interested only in the applied aspects of GLC, can skip this chapter. A reader, interested in implementing GLC algorithms, may find Chap. 3 useful for this. Chapter 4 describes the methods of the simplification of visual patterns in GLCs for the better human perception.

Chapter 5 presents several GLC case studies, on the real data, which show the GLC capabilities. Chapter 6 presents the results of the experiments on discovering the visual features in the GLCs by multiple participants, with the analysis of the human shape perception capabilities with over hundred dimensions, in these experiments. Chapter 7 presents the linear GLCs combined with machine learning, including hybrid, automatic, interactive, and collaborative versions of linear GLC, with the data classification applications from medicine to finance and image processing. Chapter 8 demonstrates the hybrid, visual, and analytical knowledge discovery and the machine learning approach for the investment strategy with GLCs. Chapter 9 presents a hybrid, visual, and analytical machine learning approach in text mining, for discovering the incongruity in humor modeling. Chapter 10 describes the capabilities of the GLC visual means to enhance evaluation of accuracy and errors of machine learning algorithms. Chapter 11 shows an approach,

to how the GLC visualization benefits the exploration of the multidimensional Pareto front, in multi-objective optimization tasks. Chapter 12 outlines the vision of a virtual data scientist and the super-intelligence with visual means. Chapter 13 concludes this book with a comparison and the fusion of methods and the discussion of the future research. The final note is on the topics, which are outside of this book. These topics are "goal-free" visualizations that are not related to the specific knowledge discovery tasks of supervised and unsupervised learning, and the Pareto optimization in the n-D data. The author's Web site of this book is located at http://www.cwu.edu/∼borisk/visualKD, where additional information and updates can be found.

Ellensburg, USA Boris Kovalerchuk

Acknowledgements

First of all thanks to my family for supporting this endeavor for years. My great appreciation goes to my collaborators: Vladimir Grishin, Antoni Wilinski, Michael Kovalerchuk, Dmytro Dovhalets, Andrew Smigaj, and Evgenii Vityaev. This book is based on a series of conference and journal papers, written jointly with them. These papers are listed in the reference section in Chap. 1 under respective names. This book would not be possible without their effort; and the effort by the graduate and undergraduate students: James Smigaj, Abdul Anwar, Jacob Brown, Sadiya Syeda, Abdulrahman Gharawi, Mitchell Hanson, Matthew Stalder, Frank Senseney, Keyla Cerna, Julian Ramirez, Kyle Discher, Chris Cottle, Antonio Castaneda, Scott Thomas, and Tommy Mathan, who have been involved in writing the code and the computational explorations. Over 70 Computer Science students from the Central Washington University (CWU) in the USA and the West Pomeranian Technical University (WPTU) in Poland participated in visual pattern discovery and experiments described in Chap. 6. The visual pattern discovery demonstrated its universal nature, when students at CWU in the USA, WPTU in Poland, and Nanjing University of Aeronautics and Astronautics in China were able to discover the visual pattern in the n-D data GLC visualizations during my lectures and challenged me with interesting questions. Discussion of the work of students involved in GLC development with the colleagues: Razvan Andonie, Szilard Vajda, and Donald Davendra helped, in writing this book, too.

I would like to thank Andrzej Piegat and the anonymous reviewers of our journal and conference papers, for their critical readings of those papers. I owe much to William Sumner and Dale Comstock for the critical readings of multiple parts of the manuscript. The remaining errors are mine, of course.

My special appreciation is to Alfred Inselberg, for his role in developing the Parallel Coordinates and the personal kindness in our communications, which inspired me to work on this topic and book. The importance of his work is in developing the Parallel Coordinates as a powerful tool for the reversible n-D data visualization and establishing their mathematical properties. It is a real marvel in its

elegance and power. As we know now, Parallel Coordinates were originated in nineteenth century. However, for almost 100 years, they have been forgotten. Mathematics, in Cartesian Coordinates, continues to dominate in science for the last 400 years, providing tremendous benefits, while other known coordinate systems play a much more limited role. The emergence of Data Science requires going beyond the Cartesian Coordinates. Alfred Inselberg likely was the first person to recognize this need, long before even the term Data Science was coined. This book is a further step in Data Science beyond the Cartesian Coordinates, in this long-term journey.

Contents

1 Motivation, Problems and Approach 1
1.1 Motivation 1
1.2 Visualization: From n-D Points to 2-D Points 2
1.3 Visualization: From n-D Points to 2-D Structures 4
1.4 Analysis of Alternatives 7
1.5 Approach 10
References 12

2 General Line Coordinates (GLC) 15
2.1 Reversible General Line Coordinates 15
2.1.1 Generalization of Parallel and Radial Coordinates 15
2.1.2 n-Gon and Circular Coordinates 18
2.1.3 Types of GLC in 2-D and 3-D 21
2.1.4 In-Line Coordinates 23
2.1.5 Dynamic Coordinates 26
2.1.6 Bush and Parallel Coordinates with Shifts 28
2.2 Reversible Paired Coordinates 29
2.2.1 Paired Orthogonal Coordinates 29
2.2.2 Paired Coordinates with Non-linear Scaling 33
2.2.3 Partially Collocated and Non-orthogonal Collocated Coordinates 34
2.2.4 Paired Radial (Star) Coordinates 35
2.2.5 Paired Elliptical Coordinates 38
2.2.6 Open and Closed Paired Crown Coordinates 40
2.2.7 Clutter Suppressing in Paired Coordinates 44
2.3 Discussion on Reversible and Non-reversible Visualization Methods 45
References 47

3 Theoretical and Mathematical Basis of GLC 49
3.1 Graphs in General Line Coordinates 49
3.2 Steps and Properties of Graph Construction Algorithms 55
3.3 Fixed Single Point Approach 58
3.3.1 Single Point Algorithm 58
3.3.2 Statements Based on Single Point Algorithm 59
3.3.3 Generalization of a Fixed Point to GLC 62
3.4 Theoretical Limits to Preserve n-D Distances in 2-D: Johnson-Lindenstrauss Lemma 64
3.5 Visual Representation of n-D Relations in GLC 65
3.5.1 Hyper-cubes and Clustering in CPC 67
3.5.2 Comparison of Linear Dependencies in PC, CPC and SPC 68
3.5.3 Visualization of n-D Linear Functions and Operators in CPC, SPC and PC 71
References 75

4 Adjustable GLCs for Decreasing Occlusion and Pattern Simplification 77
4.1 Decreasing Occlusion by Shifting and Disconnecting Radial Coordinates 77
4.2 Simplifying Patterns by Relocating and Scaling Parallel Coordinates 78
4.2.1 Shifting and Tilting Parallel Coordinates 78
4.2.2 Shifting and Reordering of Parallel Coordinates 80
4.3 Simplifying Patterns and Decreasing Occlusion by Relocating, Reordering, and Negating Shifted Paired Coordinates 82
4.3.1 Negating Shifted Paired Coordinates for Removing Crossings 82
4.3.2 Relocating Shifted Paired Coordinates for Making the Straight Horizontal Lines 85
4.3.3 Relocating Shifted Paired Coordinates for Making a Single 2-D Point 85
4.4 Simplifying Patterns by Relocating and Scaling Circular and n-Gon Coordinates 86
4.5 Decreasing Occlusion with the Expanding and Shrinking Datasets 90
4.5.1 Expansion Alternatives 90
4.5.2 Rules and Classification Accuracy for Vicinity in E1 91

4.6 Case Studies for the Expansion E1 92
4.7 Discussion 99
References 99

5 GLC Case Studies 101
5.1 Case Study 1: Glass Processing with CPC, APC and SPC 101
5.2 Case Study 2: Simulated Data with PC and CPC 103
5.3 Case Study 3: World Hunger Data 105
5.4 Case Study 4: Challenger USA Space Shuttle Disaster with PC and CPC 107
5.5 Case Study 5: Visual n-D Feature Extraction from Blood Transfusion Data with PSPC 109
5.6 Case Study 6: Health Monitoring with PC and CPC 111
5.7 Case Study 7: Iris Data Classification in Two-Layer Visual Representation 114
5.7.1 Extended Convex Hulls for Iris Data in CPC 115
5.7.2 First Layer Representation 116
5.7.3 Second Layer Representation for Classes 2 and 3 118
5.7.4 Comparison with Parallel Coordinates, Radvis and SVM 119
5.8 Case Study 8: Iris Data with PWC 122
5.9 Case Study 9: Car Evaluation Data with PWC 127
5.10 Case Study 10: Car Data with CPC, APC, SPC, and PC 130
5.11 Case Study 11: Glass Identification Data with Bush Coordinates and Parallel Coordinates 133
5.12 Case Study 12: Seeds Dataset with In-Line Coordinates and Shifted Parallel Coordinates 135
5.13 Case Study 13: Letter Recognition Dataset with SPC 137
5.14 Conclusion 140
References 140

6 Discovering Visual Features and Shape Perception Capabilities in GLC 141
6.1 Discovering Visual Features for Prediction 141
6.2 Experiment 1: CPC Stars Versus Traditional Stars for 192-D Data 145
6.3 Experiment 2: Stars Versus PC for 48-D, 72-D and 96-D Data 147
6.3.1 Hyper-Tubes Recognition 147
6.3.2 Feature Selection 149
6.3.3 Unsupervised Learning Features for Classification 151
6.3.4 Collaborative N-D Visualization and Feature Selection in Data Exploration 152

6.4 Experiment 3: Stars and CPC Stars Versus PC for 160-D Data . . . 153
6.4.1 Experiment Goal and Setting . . . 153
6.4.2 Task and Solving Hints . . . 155
6.4.3 Results . . . 156
6.5 Experiment 4: CPC Stars, Stars and PC for Feature Extraction on Real Data in 14-D and 170-D . . . 158
6.5.1 Closed Contour Lossless Visual Representation . . . 158
6.5.2 Feature Extraction Algorithm . . . 161
6.5.3 Comparison with Parallel Coordinates . . . 163
6.6 Discussion . . . 164
6.6.1 Comparison of Experiments 1 and 3 . . . 164
6.6.2 Application Scope of CPC Stars . . . 165
6.6.3 Prospects for Higher Data Dimensions . . . 166
6.6.4 Shape Perception Capabilities: Gestalt Law . . . 167
6.7 Collaborative Visualization . . . 168
6.8 Conclusion . . . 171
References . . . 171

7 Interactive Visual Classification, Clustering and Dimension Reduction with GLC-L . . . 173
7.1 Introduction . . . 173
7.2 Methods: Linear Dependencies for Classification with Visual Interactive Means . . . 174
7.2.1 Base GLC-L Algorithm . . . 174
7.2.2 Interactive GLC-L Algorithm . . . 177
7.2.3 Algorithm GLC-AL for Automatic Discovery of Relation Combined with Interactions . . . 179
7.2.4 Visual Structure Analysis of Classes . . . 181
7.2.5 Algorithm GLC-DRL for Dimension Reduction . . . 181
7.2.6 Generalization of the Algorithms for Discovering Non-linear Functions and Multiple Classes . . . 182
7.3 Case Studies . . . 183
7.3.1 Case Study 1 . . . 183
7.3.2 Case Study 2 . . . 187
7.3.3 Case Study 3 . . . 193
7.3.4 Case Study 4 . . . 195
7.3.5 Case Study 5 . . . 197
7.4 Discussion and Analysis . . . 203
7.4.1 Software Implementation, Time and Accuracy . . . 203
7.4.2 Comparison with Other Studies . . . 206
7.5 Conclusion . . . 212
References . . . 215

8 Knowledge Discovery and Machine Learning for Investment Strategy with CPC 217
8.1 Introduction 217
8.2 Process of Preparing of the Strategy 220
8.2.1 Stages of the Process 220
8.2.2 Variables 221
8.2.3 Analysis 223
8.2.4 Collocated Paired Coordinates Approach 225
8.3 Visual Method for Building Investment Strategy in 2D Space 228
8.4 Results of Investigation in 2D Space 230
8.5 Results of Investigation in 3D Space 235
8.5.1 Strategy Based on Number of Events in Cubes 235
8.5.2 Strategy Based on Quality of Events in Cubes 237
8.5.3 Discussion 242
8.6 Conclusion 246
References 247

9 Visual Text Mining: Discovery of Incongruity in Humor Modeling 249
9.1 Introduction 249
9.2 Incongruity Resolution Theory of Humor and Garden Path Jokes 250
9.3 Establishing Meanings and Meaning Correlations 252
9.3.1 Vectors of Word Association Frequencies Using Web Mining 252
9.3.2 Correlation Coefficients and Differences 253
9.4 Dataset Used in Visualizations 255
9.5 Visualization 1: Collocated Paired Coordinates 255
9.6 Visualization 2: Heat Maps 258
9.7 Visualization 3: Model Space Using Monotone Boolean Chains 259
9.8 Conclusion 262
References 263

10 Enhancing Evaluation of Machine Learning Algorithms with Visual Means 265
10.1 Introduction 265
10.1.1 Preliminaries 265
10.1.2 Challenges of k-Fold Cross Validation 266
10.2 Method 267
10.2.1 Shannon Function 267
10.2.2 Interactive Hybrid Algorithm 269
10.3 Case Studies 269

10.3.1 Case Study 1: Linear SVM and LDA in 2-D on Modeled Data . . . 270
10.3.2 Case Study 2: GLC-AL and LDA on 9-D on Wisconsin Breast Cancer Data . . . 271
10.4 Discussion and Conclusion . . . 274
References . . . 276

11 Pareto Front and General Line Coordinates . . . 277
11.1 Introduction . . . 277
11.2 Pareto Front with GLC-L . . . 279
11.3 Pareto Front and Its Approximations with CPC . . . 282
References . . . 286

12 Toward Virtual Data Scientist and Super-Intelligence with Visual Means . . . 289
12.1 Introduction . . . 289
12.2 Deficiencies . . . 290
12.3 Visual n-D ML Models: Inspiration from Success in 2-D . . . 292
12.4 Visual n-D ML Models at Different Generalization Levels . . . 294
12.5 Visual Defining and Curating ML Models . . . 298
12.6 Summary on the Virtual Data Scientist from the Visual Perspective . . . 301
12.7 Super Intelligence for High-Dimensional Data . . . 301
References . . . 305

13 Comparison and Fusion of Methods and Future Research . . . 307
13.1 Comparison of GLC with Chernoff Faces and Time Wheels . . . 307
13.2 Comparison of GLC with Stick Figures . . . 309
13.3 Comparison of Relational Information in GLCs and PC . . . 312
13.4 Fusion GLC with Other Methods . . . 313
13.5 Capabilities . . . 313
13.6 Future Research . . . 315
References . . . 316

List of Abbreviations

APC	Anchored Paired Coordinates
ATC	Anchored Tripled Coordinates
CF	Chernoff Face
CPC	Collocated Paired Coordinates
CTC	Collocated Tripled Coordinates
CV	Cross Validation
DM	Data Mining
GLC	General Line Coordinates
GLC-AL	GLC-L algorithm for automatic discovery
GLC-B	Basic GLC graph-constructing algorithm
GLC-CC1	Graph-constructing algorithm that generalizes CPC
GLC-CC2	Graph-constructing algorithm that generalizes CPC and SC
GLC-DRL	GLC-L algorithm for dimension reduction
GLC-IL	Interactive GLC-L algorithm
GLC-L	GLC for linear functions
GLC-PC	Graph-constructing algorithm that generalizes PC
GLC-SC1	Forward graph-constructing algorithm that generalizes SC
GLC-SC2	Backward graph-constructing algorithm that generalizes SC
ILC	In-Line Coordinates
IPC	In-Plane Coordinates
LDA	Linear Discriminant Analysis
MDF	Multiple Disk Form
MDS	Multidimensional Scaling
ML	Machine Learning
MOO	Multiobjective Optimization
PC	Parallel Coordinates
PCA	Principal Component Analysis
PCC	Partially Collocated Coordinates

PF	Pareto Front
P-to-G representation	Mapping an n-D point to a graph
P-to-P representation	Mapping an n-D point to a 2-D point
PWC	Paired Crown Coordinate
SC	Star Coordinate
SF	Stick Figure
SME	Subject Matter Expert
SOM	Self-Organized Map
SPC	Shifted Paired Coordinate
STP	Shifted Tripled Coordinate
SVM	Support Vector Machine
URC	Unconnected Radial Coordinates

Abstract

This book combines the advantages of the high-dimensional data visualization and machine learning for discovering complex n-D data patterns. It vastly expands the class of reversible lossless 2-D and 3-D visualization methods which preserve the n-D information for the knowledge discovery. This class of visual representations, called the General Lines Coordinates (GLCs), is accompanied by a set of algorithms for n-D data classification, clustering, dimension reduction, and Pareto optimization. The mathematical and theoretical analyses and methodology of GLC are included. The usefulness of this new approach is demonstrated in multiple case studies. These case studies include the Challenger disaster, the World hunger data, health monitoring, image processing, the text classification, market prediction for a currency exchange rate, and computer-aided medical diagnostics. Students, researchers, and practitioners in the emerging Data Science are the intended readership of this book.