1. Record Nr. UNISA996211655103316

| | |
|---|---|
| Autore | Dasu Tamraparni |
| Titolo | Exploratory data mining and data cleaning [[electronic resource] /] / Tamraparni Dasu, Theorodre Johnson |
| Pubbl/distr/stampa | New York, : Wiley-Interscience, 2003 |
| ISBN | 1-280-36625-7<br>9786610366255<br>0-470-30781-1<br>0-471-45864-3<br>0-471-44835-4 |
| Descrizione fisica | 1 online resource (226 p.) |
| Collana | Wiley series in probability and statistics |
| Altri autori (Persone) | JohnsonTheodore |
| Disciplina | 005.741<br>006.3<br>006.312 |
| Soggetti | Data mining<br>Electronic data processing - Data preparation<br>Electronic data processing - Quality control |
| Lingua di pubblicazione | Inglese |
| Formato | Materiale a stampa |
| Livello bibliografico | Monografia |
| Note generali | Description based upon print version of record. |
| Nota di bibliografia | Includes bibliographical references (p. 189-195) and index. |
| Nota di contenuto | Exploratory Data Mining and Data Cleaning; Contents; Preface; 1. Exploratory Data Mining and Data Cleaning: An Overview; 1.1 Introduction; 1.2 Cautionary Tales; 1.3 Taming the Data; 1.4 Challenges; 1.5 Methods; 1.6 EDM; 1.6.1 EDM Summaries-Parametric; 1.6.2 EDM Summaries-Nonparametric; 1.7 End-to-End Data Quality (DQ); 1.7.1 DQ in Data Preparation; 1.7.2 EDM and Data Glitches; 1.7.3 Tools for DQ; 1.7.4 End-to-End DQ: The Data Quality Continuum; 1.7.5 Measuring Data Quality; 1.8 Conclusion; 2. Exploratory Data Mining; 2.1 Introduction; 2.2 Uncertainty; 2.2.1 Annotated Bibliography 2.3 EDM: Exploratory Data Mining2.4 EDM Summaries; 2.4.1 Typical Values; 2.4.2 Attribute Variation; 2.4.3 Example; 2.4.4 Attribute Relationships; 2.4.5 Annotated Bibliography; 2.5 What Makes a Summary Useful?; 2.5.1 Statistical Properties; 2.5.2 Computational Criteria; 2.5.3 Annotated Bibliography; 2.6 Data-Driven Approach-Nonparametric Analysis; 2.6.1 The Joy of Counting; 2.6.2 Empirical |

Cumulative Distribution Function (ECDF); 2.6.3 Univariate Histograms; 2.6.4 Annotated Bibliography; 2.7 EDM in Higher Dimensions; 2.8 Rectilinear Histograms; 2.9 Depth and Multivariate Binning
2.9.1 Data Depth2.9.2 Aside: Depth-Related Topics; 2.9.3 Annotated Bibliography; 2.10 Conclusion; 3. Partitions and Piecewise Models; 3.1 Divide and Conquer; 3.1.1 Why Do We Need Partitions?; 3.1.2 Dividing Data; 3.1.3 Applications of Partition-Based EDM Summaries; 3.2 Axis-Aligned Partitions and Data Cubes; 3.2.1 Annotated Bibliography; 3.3 Nonlinear Partitions; 3.3.1 Annotated Bibliography; 3.4 DataSpheres (DS); 3.4.1 Layers; 3.4.2 Data Pyramids; 3.4.3 EDM Summaries; 3.4.4 Annotated Bibliography; 3.5 Set Comparison Using EDM Summaries; 3.5.1 Motivation; 3.5.2 Comparison Strategy
3.5.3 Statistical Tests for Change3.5.4 Application-Two Case Studies; 3.5.5 Annotated Bibliography; 3.6 Discovering Complex Structure in Data with EDM Summaries; 3.6.1 Exploratory Model Fitting in Interactive Response Time; 3.6.2 Annotated Bibliography; 3.7 Piecewise Linear Regression; 3.7.1 An Application; 3.7.2 Regression Coefficients; 3.7.3 Improvement in Fit; 3.7.4 Annotated Bibliography; 3.8 One-Pass Classification; 3.8.1 Quantile-Based Prediction with Piecewise Models; 3.8.2 Simulation Study; 3.8.3 Annotated Bibliography; 3.9 Conclusion; 4. Data Quality; 4.1 Introduction
4.2 The Meaning of Data Quality4.2.1 An Example; 4.2.2 Data Glitches; 4.2.3 Conventional Definition of DQ; 4.2.4 Times Have Changed; 4.2.5 Annotated Bibliography; 4.3 Updating DQ Metrics: Data Quality Continuum; 4.3.1 Data Gathering; 4.3.2 Data Delivery; 4.3.3 Data Monitoring; 4.3.4 Data Storage; 4.3.5 Data Integration; 4.3.6 Data Retrieval; 4.3.7 Data Mining/Analysis; 4.3.8 Annotated Bibliography; 4.4 The Meaning of Data Quality Revisited; 4.4.1 Data Interpretation; 4.4.2 Data Suitability; 4.4.3 Dataset Type; 4.4.4 Attribute Type; 4.4.5 Application Type
4.4.6 Data Quality-A Many Splendored Thing

Sommario/riassunto

Written for practitioners of data mining, data cleaning and database management. Presents a technical treatment of data quality including process, metrics, tools and algorithms.Focuses on developing an evolving modeling strategy through an iterative data exploration loop and incorporation of domain knowledge.Addresses methods of detecting, quantifying and correcting data quality issues that can have a significant impact on findings and decisions, using commercially available tools as well as new algorithmic approaches.Uses case studies to illustrate applications in real