1. Record Nr. UNINA9910784996003321

Autore	Rokach Lior
Titolo	Data mining with decision trees [[electronic resource] /] : theory and applications / / Lior Rokach, Oded Maimon
Pubbl/distr/stampa	Singapore, : World Scientific, c2008
ISBN	1-281-91179-8 9786611911799 981-277-172-7
Descrizione fisica	1 online resource (263 p.)
Collana	Series in machine perception and artificial intelligence ; ; v. 69
Altri autori (Persone)	MaimonOded Z
Disciplina	006.312
Soggetti	Data mining Decision trees
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. 215-242) and index.
Nota di contenuto	Preface; Contents; 1. Introduction to Decision Trees; 1.1 Data Mining and Knowledge Discovery; 1.2 Taxonomy of Data Mining Methods; 1.3 Supervised Methods; 1.3.1 Overview; 1.4 Classification Trees; 1.5 Characteristics of Classification Trees; 1.5.1 Tree Size; 1.5.2 The hierarchical nature of decision trees; 1.6 Relation to Rule Induction; 2. Growing Decision Trees; 2.0.1 Training Set; 2.0.2 Definition of the Classification Problem; 2.0.3 Induction Algorithms; 2.0.4 Probability Estimation in Decision Trees; 2.0.4.1 Laplace Correction; 2.0.4.2 No Match 2.1 Algorithmic Framework for Decision Trees2.2 Stopping Criteria; 3. Evaluation of Classification Trees; 3.1 Overview; 3.2 Generalization Error; 3.2.1 Theoretical Estimation of Generalization Error; 3.2.2 Empirical Estimation of Generalization Error; 3.2.3 Alternatives to the Accuracy Measure; 3.2.4 The F-Measure; 3.2.5 Confusion Matrix; 3.2.6 Classifier Evaluation under Limited Resources; 3.2.6.1 ROC Curves; 3.2.6.2 Hit Rate Curve; 3.2.6.3 Qrecall (Quota Recall); 3.2.6.4 Lift Curve; 3.2.6.5 Pearson Correlation Coegfficient; 3.2.6.6 Area Under Curve (AUC); 3.2.6.7 Average Hit Rate 3.2.6.8 Average Qrecall3.2.6.9 Potential Extract Measure (PEM); 3.2.7 Which Decision Tree Classifier is Better?; 3.2.7.1 McNemar's Test; 3.2.7.2 A Test for the Difference of Two Proportions; 3.2.7.3 The

Resampled Paired t Test; 3.2.7.4 The k-fold Cross-validated Paired t Test; 3.3 Computational Complexity; 3.4 Comprehensibility; 3.5 Scalability to Large Datasets; 3.6 Robustness; 3.7 Stability; 3.8 Interestingness Measures; 3.9 Overfitting and Underfitting; 3.10 "No Free Lunch" Theorem; 4. Splitting Criteria; 4.1 Univariate Splitting Criteria; 4.1.1 Overview; 4.1.2 Impurity based Criteria
4.1.3 Information Gain4.1.4 Gini Index; 4.1.5 Likelihood Ratio Chi-squared Statistics; 4.1.6 DKM Criterion; 4.1.7 Normalized Impurity-based Criteria; 4.1.8 Gain Ratio; 4.1.9 Distance Measure; 4.1.10 Binary Criteria; 4.1.11 Twoing Criterion; 4.1.12 Orthogonal Criterion; 4.1.13 Kolmogorov-Smirnov Criterion; 4.1.14 AUC Splitting Criteria; 4.1.15 Other Univariate Splitting Criteria; 4.1.16 Comparison of Univariate Splitting Criteria; 4.2 Handling Missing Values; 5. Pruning Trees; 5.1 Stopping Criteria; 5.2 Heuristic Pruning; 5.2.1 Overview; 5.2.2 Cost Complexity Pruning
5.2.3 Reduced Error Pruning5.2.4 Minimum Error Pruning (MEP); 5.2.5 Pessimistic Pruning; 5.2.6 Error-Based Pruning (EBP); 5.2.7 Minimum Description Length (MDL) Pruning; 5.2.8 Other Pruning Methods; 5.2.9 Comparison of Pruning Methods; 5.3 Optimal Pruning; 6. Advanced Decision Trees; 6.1 Survey of Common Algorithms for Decision Tree Induction; 6.1.1 ID3; 6.1.2 C4.5; 6.1.3 CART; 6.1.4 CHAID; 6.1.5 QUEST.; 6.1.6 Reference to Other Algorithms; 6.1.7 Advantages and Disadvantages of Decision Trees; 6.1.8 Oblivious Decision Trees; 6.1.9 Decision Trees Inducers for Large Datasets
6.1.10 Online Adaptive Decision Trees

Sommario/riassunto

This is the first comprehensive book dedicated entirely to the field of decision trees in data mining and covers all aspects of this important technique. Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining, the science and technology of exploring large and complex bodies of data in order to discover useful patterns. The area is of great importance because it enables modeling and knowledge extraction from the abundance of data available. Both theoreticians and practitioners are continually seeking techniques to make the process more