

Contents

1	Introduction	3
1.1	What motivated data mining? Why is it important?	3
1.2	So, what is data mining?	6
1.3	Data mining — on what kind of data?	8
1.3.1	Relational databases	9
1.3.2	Data warehouses	11
1.3.3	Transactional databases	12
1.3.4	Advanced database systems and advanced database applications	13
1.4	Data mining functionalities — what kinds of patterns can be mined?	13
1.4.1	Concept/class description: characterization and discrimination	13
1.4.2	Association analysis	14
1.4.3	Classification and prediction	15
1.4.4	Clustering analysis	16
1.4.5	Evolution and deviation analysis	16
1.5	Are all of the patterns interesting?	17
1.6	A classification of data mining systems	18
1.7	Major issues in data mining	19
1.8	Summary	21

Contents

2	Data Warehouse and OLAP Technology for Data Mining	3
2.1	What is a data warehouse?	3
2.2	A multidimensional data model	6
2.2.1	From tables to data cubes	6
2.2.2	Stars, snowflakes, and fact constellations: schemas for multidimensional databases	8
2.2.3	Examples for defining star, snowflake, and fact constellation schemas	11
2.2.4	Measures: their categorization and computation	13
2.2.5	Introducing concept hierarchies	14
2.2.6	OLAP operations in the multidimensional data model	15
2.2.7	A starlet query model for querying multidimensional databases	18
2.3	Data warehouse architecture	19
2.3.1	Steps for the design and construction of data warehouses	19
2.3.2	A three-tier data warehouse architecture	20
2.3.3	OLAP server architectures: ROLAP vs. MOLAP vs. HOLAP	22
2.3.4	SQL extensions to support OLAP operations	24
2.4	Data warehouse implementation	24
2.4.1	Efficient computation of data cubes	25
2.4.2	Indexing OLAP data	30
2.4.3	Efficient processing of OLAP queries	30
2.4.4	Metadata repository	31
2.4.5	Data warehouse back-end tools and utilities	32
2.5	Further development of data cube technology	32
2.5.1	Discovery-driven exploration of data cubes	33
2.5.2	Complex aggregation at multiple granularities: Multifeature cubes	36
2.6	From data warehousing to data mining	38
2.6.1	Data warehouse usage	38
2.6.2	From on-line analytical processing to on-line analytical mining	39
2.7	Summary	41

Contents

3	Data Preprocessing	3
3.1	Why preprocess the data?	3
3.2	Data cleaning	5
3.2.1	Missing values	5
3.2.2	Noisy data	6
3.2.3	Inconsistent data	7
3.3	Data integration and transformation	8
3.3.1	Data integration	8
3.3.2	Data transformation	8
3.4	Data reduction	10
3.4.1	Data cube aggregation	10
3.4.2	Dimensionality reduction	11
3.4.3	Data compression	13
3.4.4	Numerosity reduction	14
3.5	Discretization and concept hierarchy generation	19
3.5.1	Discretization and concept hierarchy generation for numeric data	19
3.5.2	Concept hierarchy generation for categorical data	23
3.6	Summary	25

Contents

4 Primitives for Data Mining	3
4.1 Data mining primitives: what defines a data mining task?	3
4.1.1 Task-relevant data	4
4.1.2 The kind of knowledge to be mined	6
4.1.3 Background knowledge: concept hierarchies	7
4.1.4 Interestingness measures	10
4.1.5 Presentation and visualization of discovered patterns	12
4.2 A data mining query language	12
4.2.1 Syntax for task-relevant data specification	15
4.2.2 Syntax for specifying the kind of knowledge to be mined	15
4.2.3 Syntax for concept hierarchy specification	18
4.2.4 Syntax for interestingness measure specification	20
4.2.5 Syntax for pattern presentation and visualization specification	20
4.2.6 Putting it all together — an example of a DMQL query	21
4.3 Designing graphical user interfaces based on a data mining query language	22
4.4 Summary	22

Contents

5	Concept Description: Characterization and Comparison	1
5.1	What is concept description?	1
5.2	Data generalization and summarization-based characterization	2
5.2.1	Data cube approach for data generalization	3
5.2.2	Attribute-oriented induction	3
5.2.3	Presentation of the derived generalization	7
5.3	Efficient implementation of attribute-oriented induction	10
5.3.1	Basic attribute-oriented induction algorithm	10
5.3.2	Data cube implementation of attribute-oriented induction	11
5.4	Analytical characterization: Analysis of attribute relevance	12
5.4.1	Why perform attribute relevance analysis?	12
5.4.2	Methods of attribute relevance analysis	13
5.4.3	Analytical characterization: An example	15
5.5	Mining class comparisons: Discriminating between different classes	17
5.5.1	Class comparison methods and implementations	17
5.5.2	Presentation of class comparison descriptions	19
5.5.3	Class description: Presentation of both characterization and comparison	20
5.6	Mining descriptive statistical measures in large databases	22
5.6.1	Measuring the central tendency	22
5.6.2	Measuring the dispersion of data	23
5.6.3	Graph displays of basic statistical class descriptions	25
5.7	Discussion	28
5.7.1	Concept description: A comparison with typical machine learning methods	28
5.7.2	Incremental and parallel mining of concept description	30
5.7.3	Interestingness measures for concept description	30
5.8	Summary	31

Contents

6	Mining Association Rules in Large Databases	3
6.1	Association rule mining	3
6.1.1	Market basket analysis: A motivating example for association rule mining	3
6.1.2	Basic concepts	4
6.1.3	Association rule mining: A road map	5
6.2	Mining single-dimensional Boolean association rules from transactional databases	6
6.2.1	The Apriori algorithm: Finding frequent itemsets	6
6.2.2	Generating association rules from frequent itemsets	9
6.2.3	Variations of the Apriori algorithm	10
6.3	Mining multilevel association rules from transaction databases	12
6.3.1	Multilevel association rules	12
6.3.2	Approaches to mining multilevel association rules	14
6.3.3	Checking for redundant multilevel association rules	16
6.4	Mining multidimensional association rules from relational databases and data warehouses	17
6.4.1	Multidimensional association rules	17
6.4.2	Mining multidimensional association rules using static discretization of quantitative attributes	18
6.4.3	Mining quantitative association rules	19
6.4.4	Mining distance-based association rules	21
6.5	From association mining to correlation analysis	23
6.5.1	Strong rules are not necessarily interesting: An example	23
6.5.2	From association analysis to correlation analysis	23
6.6	Constraint-based association mining	24
6.6.1	Metarule-guided mining of association rules	25
6.6.2	Mining guided by additional rule constraints	26
6.7	Summary	29

Contents

7	Classification and Prediction	3
7.1	What is classification? What is prediction?	3
7.2	Issues regarding classification and prediction	5
7.3	Classification by decision tree induction	6
7.3.1	Decision tree induction	7
7.3.2	Tree pruning	9
7.3.3	Extracting classification rules from decision trees	10
7.3.4	Enhancements to basic decision tree induction	11
7.3.5	Scalability and decision tree induction	12
7.3.6	Integrating data warehousing techniques and decision tree induction	13
7.4	Bayesian classification	15
7.4.1	Bayes theorem	15
7.4.2	Naive Bayesian classification	16
7.4.3	Bayesian belief networks	17
7.4.4	Training Bayesian belief networks	19
7.5	Classification by backpropagation	19
7.5.1	A multilayer feed-forward neural network	20
7.5.2	Defining a network topology	21
7.5.3	Backpropagation	21
7.5.4	Backpropagation and interpretability	24
7.6	Association-based classification	25
7.7	Other classification methods	27
7.7.1	k -nearest neighbor classifiers	27
7.7.2	Case-based reasoning	28
7.7.3	Genetic algorithms	28
7.7.4	Rough set theory	28
7.7.5	Fuzzy set approaches	29
7.8	Prediction	30
7.8.1	Linear and multiple regression	30
7.8.2	Nonlinear regression	32
7.8.3	Other regression models	32
7.9	Classifier accuracy	33
7.9.1	Estimating classifier accuracy	33
7.9.2	Increasing classifier accuracy	34
7.9.3	Is accuracy enough to judge a classifier?	34
7.10	Summary	35

Contents

8 Cluster Analysis	3
8.1 What is cluster analysis?	3
8.2 Types of data in clustering analysis	4
8.2.1 Dissimilarities and similarities: Measuring the quality of clustering	5
8.2.2 Interval-scaled variables	6
8.2.3 Binary variables	7
8.2.4 Nominal, ordinal, and ratio-scaled variables	9
8.2.5 Variables of mixed types	10
8.3 A categorization of major clustering methods	11
8.4 Partitioning methods	12
8.4.1 Classical partitioning methods: k -means and k -medoids	12
8.4.2 Partitioning methods in large databases: from k -medoids to CLARANS	15
8.5 Hierarchical methods	16
8.5.1 Agglomerative and divisive hierarchical clustering	16
8.5.2 BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies	17
8.5.3 CURE: Clustering Using REpresentatives	18
8.5.4 CHAMELEON: A hierarchical clustering algorithm using dynamic modeling	20
8.6 Density-based clustering methods	21
8.6.1 DBSCAN: A density-based clustering method based on connected regions with sufficiently high density	21
8.6.2 OPTICS: Ordering Points To Identify the Clustering Structure	22
8.6.3 DENCLUE: Clustering based on density distribution functions	23
8.7 Grid-based clustering methods	24
8.7.1 STING: A Statistical Information Grid Approach	25
8.7.2 WaveCluster: Clustering using wavelet transformation	26
8.7.3 CLIQUE: Clustering high-dimensional space	28
8.8 Model-based clustering methods	29
8.9 Outlier analysis	29
8.9.1 Statistical approach for outlier detection	30
8.9.2 Distance-based outlier detection	31
8.9.3 Deviation-based outlier detection	32
8.10 Summary	33

Contents

9 Mining Complex Types of Data	3
9.1 Generalization and Multidimensional Analysis of Complex Data Objects	3
9.1.1 Generalization on structured data	3
9.1.2 Aggregation and approximation in spatial and multimedia data generalization	4
9.1.3 Generalization of object identifiers and class/subclass hierarchies	5
9.1.4 Generalization on inherited and derived properties	5
9.1.5 Generalization on class composition hierarchies	5
9.1.6 Class-based generalization and mining object data cubes	6
9.2 Mining Spatial Databases	6
9.2.1 Spatial data cube construction and spatial OLAP	7
9.2.2 Spatial characterization	7
9.2.3 Spatial association analysis	7
9.2.4 Spatial classification and prediction	7
9.2.5 Spatial clustering methods	7
9.3 Mining Time-Series Databases and Temporal Databases	7
9.3.1 Similarity search in time-series analysis	7
9.3.2 Trend analysis	7
9.3.3 Periodicity analysis	7
9.3.4 Sequential pattern mining	7
9.3.5 Plan mining by divide-and-conquer	8
9.4 Mining Text Databases	8
9.4.1 Text data analysis and information retrieval	8
9.4.2 Keyword-based association analysis	8
9.4.3 Document classification analysis	8
9.4.4 Automated extraction of structures in text documents	8
9.5 Mining Multimedia Databases	8
9.5.1 Similarity search in multimedia data	8
9.5.2 Multi-dimensional analysis of multimedia data	8
9.5.3 Mining associations in multimedia data	8
9.6 Mining the World-Wide-Web	8
9.6.1 Web mining and a classification of Web mining tasks	9
9.6.2 Web usage mining	9
9.6.3 Web structure mining	9
9.6.4 Web content mining	9
9.7 Summary	9

Contents

- 10 Data Mining Applications and Trends in Data Mining** **3**
- 10.1 Data Mining Applications 3
 - 10.1.1 Customized Data Mining Tools for Domain-Specific Applications 3
 - 10.1.2 Intelligent Query Answering with Data Mining Techniques 3
- 10.2 Other Themes on Data Mining 3
 - 10.2.1 Visual and audio data mining 3
 - 10.2.2 Scientific data mining 3
 - 10.2.3 Commercial Data Mining Systems and Prototypes 3
- 10.3 Social Impacts of Data Mining 3
- 10.4 Trends and Research Issues in Data Mining 4
- 10.5 Summary 4