

GLOBAL
EDITION



Introduction to Data Mining

SECOND EDITION

Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar



INTRODUCTION TO DATA MINING

Introduction to Data Mining eBook: Global Edition

Table of Contents

Front Cover

Title Page

Copyright Page

Dedication

Preface to the Second Edition

Contents

1 Introduction

1.1 What Is Data Mining?

1.2 Motivating Challenges

1.3 the Origins of Data Mining

1.4 Data Mining Tasks

1.5 Scope and Organization of the Book

1.6 Bibliographic Notes

1.7 Exercises

2 Data

2.1 Types of Data

2.1.1 Attributes and Measurement

2.1.2 Types of Data Sets

2.2 Data Quality

2.2.1 Measurement and Data Collection Issues

2.2.2 Issues Related to Applications

Table of Contents

2.3 Data Preprocessing

- 2.3.1 Aggregation
- 2.3.2 Sampling
- 2.3.3 Dimensionality Reduction
- 2.3.4 Feature Subset Selection
- 2.3.5 Feature Creation
- 2.3.6 Discretization and Binarization
- 2.3.7 Variable Transformation

2.4 Measures of Similarity and Dissimilarity

- 2.4.1 Basics
- 2.4.2 Similarity and Dissimilarity Between Simple Attributes
- 2.4.3 Dissimilarities Between Data Objects
- 2.4.4 Similarities Between Data Objects
- 2.4.5 Examples of Proximity Measures
- 2.4.6 Mutual Information
- 2.4.7 Kernel Functions*
- 2.4.8 Bregman Divergence*
- 2.4.9 Issues in Proximity Calculation
- 2.4.10 Selecting the Right Proximity Measure

2.5 Bibliographic Notes

2.6 Exercises

3 Classification: Basic Concepts and Techniques

3.1 Basic Concepts

3.2 General Framework for Classification

3.3 Decision Tree Classifier

- 3.3.1 A Basic Algorithm to Build a Decision Tree
- 3.3.2 Methods for Expressing Attribute Test Conditions
- 3.3.3 Measures for Selecting an Attribute Test Condition
- 3.3.4 Algorithm for Decision Tree Induction

Table of Contents

3.3.5 Example Application: Web Robot Detection

3.3.6 Characteristics of Decision Tree Classifiers

3.4 Model Overfitting

3.4.1 Reasons for Model Overfitting

3.5 Model Selection

3.5.1 Using a Validation Set

3.5.2 Incorporating Model Complexity

3.5.3 Estimating Statistical Bounds

3.5.4 Model Selection for Decision Trees

3.6 Model Evaluation

3.6.1 Holdout Method

3.6.2 Cross-validation

3.7 Presence of Hyper-parameters

3.7.1 Hyper-parameter Selection

3.7.2 Nested Cross-validation

3.8 Pitfalls of Model Selection and Evaluation

3.8.1 Overlap Between Training and Test Sets

3.8.2 Use of Validation Error as Generalization Error

3.9 Model Comparison*

3.9.1 Estimating the Confidence Interval for Accuracy

3.9.2 Comparing the Performance of Two Models

3.10 Bibliographic Notes

3.11 Exercises

4 Association Analysis: Basic Concepts and Algorithms

4.1 Preliminaries

4.2 Frequent Itemset Generation

4.2.1 The Apriori Principle

4.2.2 Frequent Itemset Generation in the Algorithm

Table of Contents

4.2.3 Candidate Generation and Pruning

4.2.4 Support Counting

4.2.5 Computational Complexity

4.3 Rule Generation

4.3.1 Confidence-based Pruning

4.3.2 Rule Generation in Algorithm

4.3.3 an Example: Congressional Voting Records

4.4 Compact Representation of Frequent Itemsets

4.4.1 Maximal Frequent Itemsets

4.4.2 Closed Itemsets

4.5 Alternative Methods for Generating Frequent Itemsets*

4.6 FP-Growth Algorithm*

4.6.1 FP-Tree Representation

4.6.2 Frequent Itemset Generation in FP-Growth Algorithm

4.7 Evaluation of Association Patterns

4.7.1 Objective Measures of Interestingness

4.7.2 Measures Beyond Pairs of Binary Variables

4.7.3 Simpsons Paradox

4.8 Effect of Skewed Support Distribution

4.9 Bibliographic Notes

4.10 Exercises

5 Cluster Analysis: Basic Concepts and Algorithms

5.1 Overview

5.1.1 What Is Cluster Analysis?

5.1.2 Different Types of Clusterings

5.1.3 Different Types of Clusters

Road Map

5.2 K-means

Table of Contents

5.2.1 The Basic K-means Algorithm

5.2.2 K-means: Additional Issues

5.2.3 Bisecting K-means

5.2.4 K-means and Different Types of Clusters

5.2.5 Strengths and Weaknesses

5.2.6 K-means as an Optimization Problem

5.3 Agglomerative Hierarchical Clustering

5.3.1 Basic Agglomerative Hierarchical Clustering Algorithm

5.3.2 Specific Techniques

5.3.3 The Lance-williams Formula for Cluster Proximity

5.3.4 Key Issues in Hierarchical Clustering

5.3.5 Outliers

5.3.6 Strengths and Weaknesses

5.4 DBSCAN

5.4.1 Traditional Density: Center-based Approach

5.4.2 The Dbscan Algorithm

5.4.3 Strengths and Weaknesses

5.5 Cluster Evaluation

5.5.1 Overview

5.5.2 Unsupervised Cluster Evaluation Using Cohesion and Separation

5.5.3 Unsupervised Cluster Evaluation Using the Proximity Matrix

5.5.4 Unsupervised Evaluation of Hierarchical Clustering

5.5.5 Determining the Correct Number of Clusters

5.5.6 Clustering Tendency

5.5.7 Supervised Measures of Cluster Validity

5.5.8 Assessing the Significance of Cluster Validity Measures

5.5.9 Choosing a Cluster Validity Measure

5.6 Bibliographic Notes

5.7 Exercises

Table of Contents

6 Classification: Alternative Techniques

6.1 Types of Classifiers

6.2 Rule-Based Classifier

6.2.1 How a Rule-Based Classifier Works

6.2.2 Properties of a Rule Set

6.2.3 Direct Methods for Rule Extraction

6.2.4 Indirect Methods for Rule Extraction

6.2.5 Characteristics of Rule-Based Classifiers

6.3 Nearest Neighbor Classifiers

6.3.1 Algorithm

6.3.2 Characteristics of Nearest Neighbor Classifiers

6.4 Naïve Bayes Classifier

6.4.1 Basics of Probability Theory

6.4.2 Naïve Bayes Assumption

6.5 Bayesian Networks

6.5.1 Graphical Representation

6.5.2 Inference and Learning

6.5.3 Characteristics of Bayesian Networks

6.6 Logistic Regression

6.6.1 Logistic Regression as a Generalized Linear Model

6.6.2 Learning Model Parameters

6.6.3 Characteristics of Logistic Regression

6.7 Artificial Neural Network (ann)

6.7.1 Perceptron

6.7.2 Multi-layer Neural Network

6.7.3 Characteristics of Ann

6.8 Deep Learning

6.8.1 Using Synergistic Loss Functions

6.8.2 Using Responsive Activation Functions

Table of Contents

6.8.3 Regularization

6.8.4 Initialization of Model Parameters

6.8.5 Characteristics of Deep Learning

6.9 Support Vector Machine (svm)

6.9.1 Margin of a Separating Hyperplane

6.9.2 Linear SVM

6.9.3 Soft-margin SVM

6.9.4 Nonlinear SVM

6.9.5 Characteristics of SVM

6.10 Ensemble Methods

6.10.1 Rationale for Ensemble Method

6.10.2 Methods for Constructing an Ensemble Classifier

6.10.3 Bias-Variance Decomposition

6.10.4 Bagging

6.10.5 Boosting

6.10.6 Random Forests

6.10.7 Empirical Comparison Among Ensemble Methods

6.11 Class Imbalance Problem

6.11.1 Building Classifiers with Class Imbalance

6.11.2 Evaluating Performance with Class Imbalance

6.11.3 Finding an Optimal Score Threshold

6.11.4 Aggregate Evaluation of Performance

6.12 Multiclass Problem

6.13 Bibliographic Notes

6.14 Exercises

7 Association Analysis: Advanced Concepts

7.1 Handling Categorical Attributes

7.2 Handling Continuous Attributes

Table of Contents

7.2.1 Discretization-Based Methods

7.2.2 Statistics-Based Methods

7.2.3 Non-Discretization Methods

7.3 Handling a Concept Hierarchy

7.4 Sequential Patterns

7.4.1 Preliminaries

7.4.2 Sequential Pattern Discovery

7.4.3 Timing Constraints*

7.4.4 Alternative Counting Schemes*

7.5 Subgraph Patterns

7.5.1 Preliminaries

7.5.2 Frequent Subgraph Mining

7.5.3 Candidate Generation

7.5.4 Candidate Pruning

7.5.5 Support Counting

7.6 Infrequent Patterns*

7.6.1 Negative Patterns

7.6.2 Negatively Correlated Patterns

7.6.3 Comparisons Among Infrequent Patterns, Negative Patterns, and
Negatively Correlated Patterns

7.6.4 Techniques for Mining Interesting Infrequent Patterns

7.6.5 Techniques Based on Mining Negative Patterns

7.6.6 Techniques Based on Support Expectation

7.7 Bibliographic Notes

7.8 Exercises

8 Cluster Analysis: Additional Issues and Algorithms

8.1 Characteristics of Data, Clusters, and Clustering Algorithms

8.1.1 Example: Comparing K-means and Dbscan

8.1.2 Data Characteristics

Table of Contents

8.1.3 Cluster Characteristics

8.1.4 General Characteristics of Clustering Algorithms

Road Map

8.2 Prototype-based Clustering

8.2.1 Fuzzy Clustering

8.2.2 Clustering Using Mixture Models

8.2.3 Self-organizing Maps (SOM)

8.3 Density-Based Clustering

8.3.1 Grid-Based Clustering

8.3.2 Subspace Clustering

8.3.3 Denclue: A Kernel-Based Scheme for Density-based Clustering

8.4 Graph-Based Clustering

8.4.1 Sparsification

8.4.2 Minimum Spanning Tree (MST) Clustering

8.4.3 Opossum: Optimal Partitioning of Sparse Similarities Using Metis

8.4.4 Chameleon: Hierarchical Clustering with Dynamic Modeling

8.4.5 Spectral Clustering

8.4.6 Shared Nearest Neighbor Similarity

8.4.7 the Jarvis-patrick Clustering Algorithm

8.4.8 SNN Density

8.4.9 SNN Density-Based Clustering

8.5 Scalable Clustering Algorithms

8.5.1 Scalability: General Issues and Approaches

8.5.2 Birch

8.5.3 Cure

8.6 Which Clustering Algorithm?

8.7 Bibliographic Notes

8.8 Exercises

9 Anomaly Detection

Table of Contents

9.1 Characteristics of Anomaly Detection Problems

9.1.1 A Definition of an Anomaly

9.1.2 Nature of Data

9.1.3 How Anomaly Detection is Used

9.2 Characteristics of Anomaly Detection Methods

9.3 Statistical Approaches

9.3.1 Using Parametric Models

9.3.2 Using Non-Parametric Models

9.3.3 Modeling Normal and Anomalous Classes

9.3.4 Assessing Statistical Significance

9.3.5 Strengths and Weaknesses

9.4 Proximity-Based Approaches

9.4.1 Distance-Based Anomaly Score

9.4.2 Density-Based Anomaly Score

9.4.3 Relative Density-Based Anomaly Score

9.4.4 Strengths and Weaknesses

9.5 Clustering-Based Approaches

9.5.1 Finding Anomalous Clusters

9.5.2 Finding Anomalous Instances

9.5.3 Strengths and Weaknesses

9.6 Reconstruction-Based Approaches

9.6.1 Strengths and Weaknesses

9.7 One-Class Classification

9.7.1 Use of Kernels

9.7.2 The Origin Trick

9.7.3 Strengths and Weaknesses

9.8 Information Theoretic Approaches

9.8.1 Strengths and Weaknesses

9.9 Evaluation of Anomaly Detection

Table of Contents

9.10 Bibliographic Notes

9.11 Exercises

10 Avoiding False Discoveries

10.1 Preliminaries: Statistical Testing

10.1.1 Significance Testing

10.1.2 Hypothesis Testing

10.1.3 Multiple Hypothesis Testing

10.1.4 Pitfalls in Statistical Testing

10.2 Modeling Null and Alternative Distributions

10.2.1 Generating Synthetic Data Sets

10.2.2 Randomizing Class Labels

10.2.3 Resampling Instances

10.2.4 Modeling the Distribution of the Test Statistic

10.3 Statistical Testing for Classification

10.3.1 Evaluating Classification Performance

10.3.2 Binary Classification as Multiple Hypothesis Testing

10.3.3 Multiple Hypothesis Testing in Model Selection

10.4 Statistical Testing for Association Analysis

10.4.1 Using Statistical Models

10.4.2 Using Randomization Methods

10.5 Statistical Testing for Cluster Analysis

10.5.1 Generating a Null Distribution for Internal Indices

10.5.2 Generating a Null Distribution for External Indices

10.5.3 Enrichment

10.6 Statistical Testing for Anomaly Detection

10.7 Bibliographic Notes

10.8 Exercises

Author Index

Table of Contents

Subject Index

Copyright Permissions

Back Cover