PRIVACY-PRESERVING DATA MINING: MODELS AND ALGORITHMS
PRIVACY-PRESERVING DATA MINING:
MODELS AND ALGORITHMS

Edited by
CHARU C. AGGARWAL
IBM T. J. Watson Research Center, Hawthorne, NY 10532

PHILIP S. YU
University of Illinois at Chicago, Chicago, IL 60607

Kluwer Academic Publishers
Boston/Dordrecht/London
Contents

List of Figures xv
List of Tables xx
Preface xxi

1 An Introduction to Privacy-Preserving Data Mining 1
Charu C. Aggarwal, Philip S. Yu
1. Introduction 1
2. Privacy-Preserving Data Mining Algorithms 3
3. Conclusions and Summary 7
References 8

2 A General Survey of Privacy-Preserving Data Mining Models and Algorithms 11
Charu C. Aggarwal, Philip S. Yu
1. Introduction 11
2. The Randomization Method 13
2.1 Privacy Quantification 15
2.2 Adversarial Attacks on Randomization 18
2.3 Randomization Methods for Data Streams 18
2.4 Multiplicative Perturbations 19
2.5 Data Swapping 19
3. Group Based Anonymization 20
3.1 The $k$-Anonymity Framework 20
3.2 Personalized Privacy-Preservation 24
3.3 Utility Based Privacy Preservation 24
3.4 Sequential Releases 25
3.5 The $t$-diversity Method 26
3.6 The $l$-closeness Model 27
3.7 Models for Text, Binary and String Data 27
4. Distributed Privacy-Preserving Data Mining 28
4.1 Distributed Algorithms over Horizontally Partitioned Data Sets 30
4.2 Distributed Algorithms over Vertically Partitioned Data 31
4.3 Distributed Algorithms for $k$-Anonymity 31
5. Privacy-Preservation of Application Results 32
5.1 Association Rule Hiding 33
5.2 Downgrading Classifier Effectiveness 34
7.2 Sketch Based Randomization 151
8. Conclusions and Summary 152
References 152

A Survey of Multiplicative Perturbation for Privacy-Preserving Data Mining
Keke Chen and Ling Liu

1. Introduction 156
   1.1 Data Privacy vs. Data Utility 157
   1.2 Outline 158
2. Definition of Multiplicative Perturbation 159
   2.1 Notations 159
   2.2 Rotation Perturbation 159
   2.3 Projection Perturbation 160
   2.4 Sketch-based Approach 162
   2.5 Geometric Perturbation 162
3. Transformation Invariant Data Mining Models 163
   3.1 Definition of Transformation Invariant Models 164
   3.2 Transformation-Invariant Classification Models 164
   3.3 Transformation-Invariant Clustering Models 165
4. Privacy Evaluation for Multiplicative Perturbation 166
   4.1 A Conceptual Multidimensional Privacy Evaluation Model 166
   4.2 Variance of Difference as Column Privacy Metric 167
   4.3 Incorporating Attack Evaluation 168
   4.4 Other Metrics 169
5. Attack Resilient Multiplicative Perturbations 169
   5.1 Naive Estimation to Rotation Perturbation 169
   5.2 ICA-Based Attacks 171
   5.3 Distance-Inference Attacks 172
   5.4 Attacks with More Prior Knowledge 174
   5.5 Finding Attack-Resilient Perturbations 175
6. Conclusion 175
References 176

A Survey of Quantification of Privacy Preserving Data Mining Algorithms
Elisa Bertino and Dan Lin and Wei Jiang

1. Metrics for Quantifying Privacy Level 184
   1.1 Data Privacy 184
   1.2 Result Privacy 189
2. Metrics for Quantifying Hiding Failure 190
3. Metrics for Quantifying Data Quality 191
   3.1 Quality of the Data Resulting from the PPDM Process 191
   3.2 Quality of the Data Mining Results 196
4. Complexity Metrics 198
5. How to Select a Proper Metric 199
6. Conclusion and Research Directions 200
References 200
Contents

9

A Survey of Utility-based Privacy-Preserving Data Transformation Methods
Ming Hua and Jian Pei
1. Introduction
   1.1 What is Utility-based Privacy Preservation? 207
2. Types of Utility-based Privacy Preservation Methods
   2.1 Privacy Models 208
   2.2 Utility Measures 210
   2.3 Summary of the Utility-Based Privacy Preserving Methods 212
3. Utility-Based Anonymization Using Local Recoding
   3.1 Global Recoding and Local Recoding 213
   3.2 Utility Measure 214
   3.3 Anonymization Methods 215
   3.4 Summary and Discussion 217
4. The Utility-based Privacy Preserving Methods in Classification Problems
   4.1 The Top-Down Specialization Method 218
   4.2 The Progressive Disclosure Algorithm 222
   4.3 Summary and Discussion 226
5. Anonymized Marginal: Injecting Utility into Anonymized Data Sets
   5.1 Anonymized Marginal 227
   5.2 Utility Measure 228
   5.3 Injecting Utility Using Anonymized Marginals 229
   5.4 Summary and Discussion 231
6. Summary 232
References 232

10

Mining Association Rules under Privacy Constraints
Jayant R. Haritsa
1. Problem Framework
2. Evolution of the Literature
3. The FRAPP Framework
4. Sample Results
5. Closing Remarks
References 261

11

A Survey of Association Rule Hiding Methods for Privacy
Vassilios S. Verykios and Aris Gkoulalas-Divanis
1. Introduction
2. Terminology and Preliminaries
3. Taxonomy of Association Rule Hiding Algorithms
4. Classes of Association Rule Algorithms
   4.1 Heuristic Approaches 270
   4.2 Border-based Approaches 275
   4.3 Exact Approaches 276
5. Other Hiding Approaches 277
6. Metrics and Performance Analysis 279
7. Discussion and Future Trends 282
8. Conclusions 283
References 284

12
A Survey of Statistical Approaches to Preserving Confidentiality of Contingency Table Entries
Stephen E. Fienberg and Aleksandra B. Slavkovic
1. Introduction 289
2. The Statistical Approach Privacy Protection 290
3. Datamining Algorithms, Association Rules, and Disclosure Limitation 292
4. Estimation and Disclosure Limitation for Multi-way Contingency Tables 293
5. Two Illustrative Examples 299
5.1 Example 1: Data from a Randomized Clinical Trial 299
5.2 Example 2: Data from the 1993 U.S. Current Population Survey 303
6. Conclusions 306
References 307

13
A Survey of Privacy-Preserving Methods Across Horizontally Partitioned Data
Murat Kantarcioglu
1. Introduction 313
2. Basic Cryptographic Techniques for Privacy-Preserving Distributed Data Mining 315
3. Common Secure Sub-protocols Used in Privacy-Preserving Distributed Data Mining 318
4. Privacy-preserving Distributed Data Mining on Horizontally Partitioned Data 323
5. Comparison to Vertically Partitioned Data Model 326
6. Extension to Malicious Parties 327
7. Limitations of the Cryptographic Techniques Used in Privacy-Preserving Distributed Data Mining 329
8. Privacy Issues Related to Data Mining Results 330
9. Conclusion 332
References 332

14
A Survey of Privacy-Preserving Methods across Vertically Partitioned Data
A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods

Kun Liu, Chris Giannella, and Hillol Kargupta

1. Introduction 362
2. Definitions and Notation 362
3. Attacking Additive Data Perturbation 363
   3.1 Eigen-Analysis and PCA Preliminaries 364
   3.2 Spectral Filtering 365
   3.3 SVD Filtering 366
   3.4 PCA Filtering 367
   3.5 MAP Estimation Attack 368
   3.6 Distribution Analysis Attack 369
   3.7 Summary 370
4. Attacking Matrix Multiplicative Data Perturbation 371
   4.1 Known I/O Attacks 372
   4.2 Known Sample Attack 375
   4.3 Other Attacks Based on ICA 376
   4.4 Summary 377
5. Attacking k-Anonymization 378
6. Conclusion 379

Acknowledgments 379
References 379

Private Data Analysis via Output Perturbation

Kobbi Nissim

1. The Abstract Model – Statistical Databases, Queries, and Sanitizers 387
2. Privacy 390
   2.1 Interpreting the Privacy Definition 392
3. The Basic Technique: Calibrating Noise to Sensitivity 396
   3.1 Applications: Functions with Low Global Sensitivity 398
4. Constructing Sanitizers for Complex Functionalities 402
   4.1 k-Means Clustering 403
Privacy-Preserving Data Stream Classification
Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Rong She, and Jian Pei

1. Introduction
   1.1 Motivating Example
   1.2 Contributions and Paper Outline

2. Related Works

3. Problem Statement
   3.1 Secure Join Stream Classification
   3.2 Naive Bayesian Classifiers

4. Our Approach
   4.1 Initialization
   4.2 Bottom-Up Propagation
   4.3 Top-Down Propagation
   4.4 Using NBC
   4.5 Algorithm Analysis

5. Empirical Studies
   5.1 Real-life Datasets
   5.2 Synthetic Datasets
   5.3 Discussion

6. Conclusions

References

Index