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 l-diversity Method 26
   3.6 The t-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

5.3 Query Auditing and Inference Control 34
6. Limitations of Privacy: The Curse of Dimensionality 37
7. Applications of Privacy-Preserving Data Mining 38
7.1 Medical Databases: The Scrub and Datafly Systems 38
7.2 Bioterrorism Applications 40
7.3 Homeland Security Applications 40
7.4 Genomic Privacy 42
8. Summary 43
References 43

3
A Survey of Inference Control Methods for Privacy-Preserving Data Mining 53
Josep Domingo-Ferrer
1. A classification of microdata protection methods 55
2. Perturbative masking methods 58
2.1 Additive noise 58
2.2 Microaggregation 59
2.3 Data swapping and rank swapping 61
2.4 Rounding 62
2.5 Resampling 62
2.6 PRAM 62
2.7 MASSC 63
3. Non-perturbative masking methods 63
3.1 Sampling 64
3.2 Global recoding 64
3.3 Top and bottom coding 65
3.4 Local suppression 65
4. Synthetic microdata generation 65
4.1 Synthetic data by multiple imputation 65
4.2 Synthetic data by bootstrap 66
4.3 Synthetic data by Latin Hypercube Sampling 66
4.4 Partially synthetic data by Cholesky decomposition 67
4.5 Other partially synthetic and hybrid microdata approaches 67
4.6 Pros and cons of synthetic microdata 68
5. Trading off information loss and disclosure risk 69
5.1 Score construction 69
5.2 R-U maps 71
5.3 k-anonymity 71
6. Conclusions and research directions 72
References 73

4
Measures of Anonymity 81
Suresh Venkatasubramanian
1. Introduction 81
1.1 What is privacy? 81
1.2 Data Anonymization Methods 83
1.3 A Classification Of Methods 84
2. Statistical Measures of Anonymity 85

2.1 Query Restriction 85
2.2 Anonymity via Variance 85
2.3 Anonymity via Multiplicity 86
3. Probabilistic Measures of Anonymity 86
3.1 Measures Based on Random Perturbation 87
3.2 Measures Based on Generalization 90
3.3 Utility vs Privacy 93
4. Computational Measures Of Anonymity 94
4.1 Anonymity via Isolation 96
5. Conclusions And New Directions 97
5.1 New Directions 98
References 98

5
k-Anonymous Data Mining: A Survey 103
V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati
1. Introduction 103
2. k-Anonymity 105
3. Algorithms for Enforcing k-Anonymity 108
4. k-Anonymity Threats from Data Mining 115
4.1 Association Rules 116
4.2 Classification Mining 116
5. k-Anonymity in Data Mining 118
6. Anonymize-and-Mine 121
7. Mine-and-Anonymize 124
7.1 Enforcing k-Anonymity on Association Rules 124
7.2 Enforcing k-Anonymity on Decision Trees 127
8. Conclusions 130
Acknowledgments 131
References 131

6
A Survey of Randomization Methods for Privacy-Preserving Data Mining 135
Charu C. Aggarwal, Philip S. Yu
1. Introduction 135
2. Reconstruction Methods for Randomization 137
2.1 The Bayes Reconstruction Method 137
2.2 The EM Reconstruction Method 139
2.3 Utility and Optimality of Randomization Models 141
3. Applications of Randomization 142
3.1 Privacy-Preserving Classification with Randomization 142
3.2 Privacy-Preserving OLAP 143
3.3 Collaborative Filtering 143
4. The Privacy-Information Loss Tradeoff 144
5. Vulnerabilities of the Randomization Method 147
6. Randomization of Time Series Data Streams 149
7. Multiplicative Noise for Randomization 150
7.1 Vulnerabilities of Multiplicative Randomization 151

7.2 Sketch Based Randomization 151
8. Conclusions and Summary 152
References 152

7
A Survey of Multiplicative Perturbation for Privacy-Preserving Data Mining 155
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

8
A Survey of Quantification of Privacy Preserving Data Mining Algorithms 181
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

9
A Survey of Utility-based Privacy-Preserving Data Transformation Methods 205
Ming Hua and Jian Pei
1. Introduction 206
   1.1 What is Utility-based Privacy Preservation? 207
2. Types of Utility-based Privacy Preservation Methods 208
   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 212
   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 217
   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 226
   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 237
Jayant R. Haritsa
1. Problem Framework 238
2. Evolution of the Literature 244
3. The FRAPP Framework 249
4. Sample Results 257
5. Closing Remarks 261
References 261

11
A Survey of Association Rule Hiding Methods for Privacy 265
Vassilios S. Verykios and Aris Gkoulalas-Divanis
1. Introduction 265
2. Terminology and Preliminaries 267
3. Taxonomy of Association Rule Hiding Algorithms 268
4. Classes of Association Rule Algorithms 269
   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 289
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 313
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 337

Jaideep Vaidya
1. Classification 339
1.1 Naïve Bayes Classification 342
1.2 Bayesian Network Structure Learning 343
1.3 Decision Tree Classification 344
2. Clustering 346
3. Association Rule Mining 347
4. Outlier detection 349
4.1 Algorithm 351
4.2 Security Analysis 352
4.3 Computation and Communication Analysis 354
5. Challenges and Research Directions 355
References 356

15
A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods 361
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

16
Private Data Analysis via Output Perturbation 385
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

4.2 SVD and PCA 405
4.3 Learning in the Statistical Queries Model 406
5. Beyond the Basics 407
5.1 Instance Based Noise and Smooth Sensitivity 407
5.2 The Sample-Aggregate Framework 409
5.3 A General Sanitization Mechanism 410
6. Related Work and Bibliographic Notes 411
References 413

17
A Survey of Query Auditing Techniques for Data Privacy 417
Shubha U. Nabar, Krishnaram Kenthapadi, Nina Mishra and Rajeev Motwani
1. Auditing Aggregate Queries 418
1.1 Offline Auditing 419
1.2 Online Auditing 420
2. Auditing Select-Project-Join Queries 428
3. Challenges in Auditing 429
4. Reading 431
References 432

18
Privacy and the Dimensionality Curse 435
Charu C. Aggarwal
1. Introduction 435
2. The Dimensionality Curse and the k-anonymity Method 438
3. The Dimensionality Curse and Condensation 443
4. The Dimensionality Curse and the Randomization Method 448
4.1 Effects of Public Information 449
4.2 Effects of High Dimensionality 452
4.3 Gaussian Perturbing Distribution 452
4.4 Uniform Perturbing Distribution 458
5. The Dimensionality Curse and l-diversity 461
6. Conclusions and Research Directions 462
References 462

19
Personalized Privacy Preservation 465
Yufei Tao and Xiaokui Xiao
1. Introduction 465
2. Formalization of Personalized Anonymity 468
2.1 Personal Privacy Requirements 468
2.2 Generalization 469
3. Combinatorial Process of Privacy Attack 471
3.1 Primary Case 472
3.2 Non-primary Case 473
4. Theoretical Foundation 475
4.1 Notations and Basic Properties 475
4.2 Derivation of the Breach Probability 476
5. Generalization Algorithm 477

5.1 The Greedy Framework 478
5.2 Optimal SA-generalization 480
6. Alternative Forms of Personalized Privacy Preservation 482
6.1 Extension of k-anonymity 483
6.2 Personalization in Location Privacy Protection 484
7. Summary and Future Work 486
References 489

20
Privacy-Preserving Data Stream Classification 491
Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Rong She, and Jian Pei
1. Introduction 491
1.1 Motivating Example 492
1.2 Contributions and Paper Outline 494
2. Related Works 496
3. Problem Statement 497
3.1 Secure Join Stream Classification 497
3.2 Naive Bayesian Classifiers 498
4. Our Approach 499
4.1 Initialization 500
4.2 Bottom-Up Propagation 500
4.3 Top-Down Propagation 502
4.4 Using NBC 503
4.5 Algorithm Analysis 504
5. Empirical Studies 505
5.1 Real-life Datasets 506
5.2 Synthetic Datasets 508
5.3 Discussion 510
6. Conclusions 511
References 512

Index 515