Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

Bibliographic Notes for Chapter 1: Introduction

The book *Knowledge Discovery in Databases*, edited by Piatetsky-Shapiro and Frawley [PSF91], is an early collection of research papers on knowledge discovery from data. The book *Advances in Knowledge Discovery and Data Mining*, edited by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy [FPSSe96], is a collection of later research results on knowledge discovery and data mining. There have been many data mining books published in recent years, including *Predictive Data Mining* by Weiss and Indurkhya [WI98], *Data Mining Solutions: Methods and Tools for Solving Real-World Problems* by Westphal and Blaxton [WB98], *Mastering Data Mining: The Art and Science of Customer Relationship Management* by Berry and Linoff [BL99], *Building Data Mining Applications for CRM* by Berson, Smith, and Thearling [BST99], *Data Mining: Practical Machine Learning Tools and Techniques* by Witten and Frank [WF05], *Principles of Data Mining (Adaptive Computation and Machine Learning)* by Hand, Mannila, and Smyth [HMS01], *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman [HTF01], *Data Mining: Introductory and Advanced Topics* by Dunham, and *Data Mining: Multimedia, Soft Computing, and Bioinformatics* by Mitra and Acharya [MA03]. There are also books containing collections of papers on particular aspects of knowledge discovery, such as *Machine Learning and Data Mining: Methods and Applications* edited by Michalski, Brakto, and Kubat [MBK98], and *Relational Data Mining* edited by Dzeroski and Lavrac [De01], as well as many tutorial notes on data mining in major database, data mining and machine learning conferences.

KDnuggets News, moderated by Piatetsky-Shapiro since 1991, is a regular, free electronic newsletter containing information relevant to data mining and knowledge discovery. The *KDnuggets* web site, located at *www.kdnuggets.com*, contains a good collection of information relating to data mining.

The data mining community started its first international conference on knowledge discovery and data mining in 1995 [Fe95]. The conference evolved from the four international workshops on knowledge discovery in databases, held from 1989 to 1994 [PS89, PS91, FUe93, Fe94]. ACM-SIGKDD, a Special Interest Group on Knowledge Discovery in Databases was set up under ACM in 1998. In 1999, ACM-SIGKDD organized the fifth international conference on knowledge discovery and data mining (KDD'99). IEEE Computer Science Society has organized its annual data mining conference, International Conference on Data Mining (ICDM), since 2001. SIAM (Society on Industrial and Applied Mathematics) has organized its annual data mining conference, SIAM Data Mining conference (SDM), since 2002. A dedicated journal, *Data Mining and Knowledge Discovery*, published by Kluwers Publishers, has been available since 1997. ACM-SIGKDD also publishes a biannual newsletter, *SIGKDD Explorations*. There are a few other international or regional conferences on data mining, such as the Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), and the International Conference on Data Warehousing and Knowledge Discovery (DaWaK).

Research in data mining has also been published in books, conferences, and journals on databases, statistics, machine learning, and data visualization. References to such sources are listed below.

Popular textbooks on database systems include *Database Systems: The Complete Book* by Garcia-Molina, Ullman, and Widom [GMUW02], *Database Management Systems* by Ramakrishnan and Gehrke [RG03], *Database System Concepts* by Silberschatz, Korth, and Sudarshan [SKS02], and *Fundamentals of Database Systems* by Elmasri and Navathe [EN03]. For an edited collection of seminal articles on database systems, see *Readings in Database Systems* by Hellerstein and Stonebraker [HS05]. Many books on data warehouse technology, systems, and applications have been published in the last several years, such as *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* by Kimball and Ross [KR02], *The Data Warehouse Lifecycle Toolkit:*

Expert Methods for Designing, Developing, and Deploying Data Warehouses by Kimball, Reeves, Ross, et al. [KRRT98], *Mastering Data Warehouse Design: Relational and Dimensional Techniques* by Imhoff, Galemmo, and Geiger [IGG03], *Building the Data Warehouse* by Inmon [Inm96], and *OLAP Solutions: Building Multidimensional Information Systems* by Thomsen [Tho97]. A set of research papers on materialized views and data warehouse implementations were collected in *Materialized Views: Techniques, Implementations, and Applications* by Gupta and Mumick [GM99]. Chaudhuri and Dayal [CD97] present a comprehensive overview of data warehouse technology.

Research results relating to data mining and data warehousing have been published in the proceedings of many international database conferences, including the *ACM-SIGMOD International Conference on Management of Data (SIGMOD)*, the *International Conference on Very Large Data Bases (VLDB),* the *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, the *International Conference on Data Engineering (ICDE)*, the *International Conference on Extending Database Technology (EDBT)*, the *International Conference on Database Theory (ICDT)*, the *International Conference on Information and Knowledge Management (CIKM)*, the *International Conference on Database and Expert Systems Applications (DEXA)*, and the *International Symposium on Database Systems for Advanced Applications (DASFAA)*. Research in data mining is also published in major database journals, such as *IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Database Systems (TODS), Journal of ACM (JACM), Information Systems, The VLDB Journal, Data and Knowledge Engineering*, *International Journal of Intelligent Information Systems (JIIS)*, and *Knowledge and Information Systems (KAIS)*.

Many effective data mining methods have been developed by statisticians and pattern recognition researchers, and introduced in a rich set of textbooks. An overview of classification from a statistical pattern recognition perspective can be found in in *Pattern Classification* by Duda, Hart, and Stork [DHS01]. There are also many textbooks covering different topics in statistical analysis, such as *Mathematical Statistics: Basic Ideas and Selected Topics* by Bickel and Doksum [BD01], *The Statistical Sleuth: A Course in Methods of Data Analysis* by Ramsey and Schafer [RS01], *Applied Linear Statistical Models* by Neter, Kutner, Nachtsheim, and Wasserman [NKNW96], *An Introduction to Generalized Linear Models* by Dobson [Dob01], *Applied Statistical Time Series Analysis* by Shumway [Shu88], and *Applied Multivariate Statistical Analysis* by Johnson and Wichern [JW02].

Research in statistics is published in the proceedings of several major statistical conferences, including *Joint Statistical Meetings*, *International Conference of the Royal Statistical Society*, and *Symposium on the Interface: Computing Science and Statistics*. Other sources of publication include the *Journal of the Royal Statistical Society*, *The Annals of Statistics*, *Journal of American Statistical Association*, *Technometrics*, and *Biometrika*.

Textbooks and reference books on machine learning include *Machine Learning, An Artificial Intelligence Approach*, Vols. 1–4, edited by Michalski et al. [MCM83, MCM86, KM90, MT94], *C4.5: Programs for Machine Learning* by Quinlan [Qui93], *Elements of Machine Learning* by Langley [Lan96], and *Machine Learning* by Mitchell [Mit97]. The book *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems* by Weiss and Kulikowski [WK91] compares classification and prediction methods from several different fields. For an edited collection of seminal articles on machine learning, see *Readings in Machine Learning* by Shavlik and Dietterich [SD90].

Machine learning research is published in the proceedings of several large machine learning and artificial intelligence conferences, including the *International Conference on Machine Learning (ML)*, the *ACM Conference on Computational Learning Theory (COLT)*, the *International Joint Conference on Artificial Intelligence (IJCAI)*, and the *American Association of Artificial Intelligence Conference (AAAI)*. Other sources of publication include major machine learning, artificial intelligence, pattern recognition, and knowledge system journals, some of which have been mentioned above. Others include *Machine Learning (ML), Artificial Intelligence Journal (AI)*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, and *Cognitive Science*.

Pioneering work on data visualization techniques is described in *The Visual Display of Quantitative Information* [Tuf83], *Envisioning Information* [Tuf90], and *Visual Explanations : Images and Quantities, Evidence and Narrative* [Tuf97], all by Tufte, in addition to *Graphics and Graphic Information Processing* by Bertin [Ber81], *Visualizing Data* by Cleveland [Cle93], and *Information Visualization in Data Mining and Knowledge Discovery* edited by Fayyad, Grinstein, and Wierse [FGW01]. Major conferences and symposiums on visualization include *ACM Human Factors in Computing Systems (CHI)*, *Visualization*, and the *International Symposium on Infor-*

mation Visualization. Research on visualization is also published in *Transactions on Visualization and Computer Graphics*, *Journal of Computational and Graphical Statistics*, and *IEEE Computer Graphics and Applications*.

The DMQL data mining query language was proposed by Han, Fu, Wang, et al. [HFW+96] for the *DBMiner* data mining system. Other examples include *Discovery Board* (formerly *Data Mine*) by Imielinski, Virmani, and Abdulghani [IVA96], and MSQL by Imielinski and Virmani [IV99]. MINE RULE, an SQL-like operator for mining single-dimensional association rules, was proposed by Meo, Psaila, and Ceri [MPC96] and extended by Baralis and Psaila [BP97]. Microsoft Corporation has made a major data mining standardization effort by proposing OLE DB for Data Mining (DM) [Cor00] and the DMX language [TM05, TMK05]. An introduction to the data mining language primitives of DMX can be found in the appendix of this book. Other standardization efforts include PMML (Programming data Model Markup Language) [Ras04], described at *www.dmg.org*, and CRISP-DM (CRoss-Industry Standard Process for Data Mining), described at *www.crisp-dm.org*.

Architectures of data mining systems have been discussed by many researchers in conference panels and meetings. The recent design of data mining languages, such as [BP97, IV99, Cor00, Ras04], the proposal of on-line analytical mining, such as [Han98], and the study of optimization of data mining queries, such as [NLHP98, STA98, LNHP99], can be viewed as steps toward the tight integration of data mining systems with database systems and data warehouse systems. For relational or object-relational systems, data mining primitives as proposed by Sarawagi, Thomas, and Agrawal [STA98] may be used as building blocks for the efficient implementation of data mining in such database systems.

Bibliography

[BD01] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1.* Prentice Hall, 2001.

[Ber81] J. Bertin. *Graphics and Graphic Information Processing.* Berlin, 1981.

[BL99] M. J. A. Berry and G. Linoff. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons, 1999.

[BP97] E. Baralis and G. Psaila. Designing templates for mining association rules. *J. Intelligent Information Systems*, 9:7–32, 1997.

[BST99] A. Berson, S. J. Smith, and K. Thearling. *Building Data Mining Applications for CRM.* McGraw-Hill, 1999.

[CD97] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26:65–74, 1997.

[Cle93] W. Cleveland. *Visualizing Data.* Hobart Press, 1993.

[Cor00] Microsoft Corporation. OLEDB for Data Mining draft specification, version 0.9. In *www.microsoft.com/data/oledb/dm*, Feb. 2000.

[De01] S. Dzeroski and N. Lavrac (eds.). *Relational Data Mining.* Springer, 2001.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification* (2nd ed.). John Wiley & Sons, 2001.

[Dob01] A. J. Dobson. *An Introduction to Generalized Linear Models* (2nd ed.). Chapman and Hall, 2001.

[EN03] R. Elmasri and S. B. Navathe. *Fundamental of Database Systems* (4th ed.). Addison Wesley, 2003.

[Fe94] U. M. Fayyad and R. Uthurusamy (eds.). *Notes of AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94).* Seattle, WA, July 1994.

[Fe95] U. M. Fayyad and R. Uthurusamy (eds.). *Proc. 1995 Int. Conf. Knowledge Discovery and Data Mining (KDD'95).* AAAI Press, Aug. 1995.

[FGW01] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann, 2001.

[FPSSe96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[FUe93] U. M. Fayyad, R. Uthurusamy, and G. Piatetsky-Shapiro (eds.). *Notes of AAAI'93 Workshop Knowledge Discovery in Databases (KDD'93).* Washington, DC, July 1993.

[GM99] A. Gupta and I. S. Mumick. *Materialized Views: Techniques, Implementations, and Applications.* MIT Press, 1999.

[GMUW02] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, 2002.

[Han98] J. Han. Towards on-line analytical mining in large databases. *SIGMOD Record*, 27:97–107, 1998.

[HFW+96] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaïane. DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, OR, Aug. 1996.

[HMS01] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[HS05] J. M. Hellerstein and M. Stonebraker. *Readings in Database Systems* (4th ed.). MIT Press, 2005.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[IGG03] C. Imhoff, N. Galemmo, and J. G. Geiger. *Mastering Data Warehouse Design : Relational and Dimensional Techniques*. John Wiley & Sons, 2003.

[Inm96] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.

[IV99] T. Imielinski and A. Virmani. MSQL: A query language for database mining. *Data Mining and Knowledge Discovery*, 3:373–408, 1999.

[IVA96] T. Imielinski, A. Virmani, and A. Abdulghani. DataMine—application programming interface and query language for KDD applications. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 256–261, Portland, OR, Aug. 1996.

[JW02] R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis* (5th ed.). Prentice Hall, 2002.

[KM90] Y. Kodratoff and R. S. Michalski. *Machine Learning, An Artificial Intelligence Approach, Vol. 3*. Morgan Kaufmann, 1990.

[KR02] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). John Wiley & Sons, 2002.

[KRRT98] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite. *The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses*. John Wiley & Sons, 1998.

[Lan96] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann, 1996.

[LNHP99] L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang. Optimization of constrained frequent set queries with 2-variable constraints. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 157–168, Philadelphia, PA, June 1999.

[MA03] S. Mitra and T. Acharya. *Data Mining: Multimedia, Soft Computing, and Bioinformatics*. John Wiley & Sons, 2003.

[MBK98] R. S. Michalski, I. Brakto, and M. Kubat. *Machine Learning and Data Mining: Methods and Applications*. John Wiley & Sons, 1998.

[MCM83] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 1*. Morgan Kaufmann, 1983.

[MCM86] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 2*. Morgan Kaufmann, 1986.

[Mit97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[MPC96] R. Meo, G. Psaila, and S. Ceri. A new SQL-like operator for mining association rules. In *Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96)*, pages 122–133, Bombay, India, Sept. 1996.

[MT94] R. S. Michalski and G. Tecuci. *Machine Learning, A Multistrategy Approach, Vol. 4*. Morgan Kaufmann, 1994.

[NKNW96] J. Neter, M. H. Kutner, C. J. Nachtsheim, and L. Wasserman. *Applied Linear Statistical Models* (4th ed.). Irwin, 1996.

[NLHP98] R. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 13–24, Seattle, WA, June 1998.

[PS89] G. Piatetsky-Shapiro. *Notes of IJCAI'89 Workshop Knowledge Discovery in Databases (KDD'89)*. Detroit, MI, July 1989.

[PS91] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–238. AAAI/MIT Press, 1991.

[PSF91] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Ras04] S. Raspl. PMML version 3.0—overview and status. In *Proc. 2004 KDD Worshop on Data Mining Standards, Services and Platforms (DM-SSP04)*, Seattle, WA, Aug. 2004.

[RG03] R. Ramakrishnan and J. Gehrke. *Database Management Systems* (3rd ed.). McGraw-Hill, 2003.

[RS01] F. Ramsey and D. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury Press, 2001.

[SD90] J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.

[Shu88] R. H. Shumway. *Applied Statistical Time Series Analysis*. Prentice Hall, 1988.

[SKS02] A. Silberschatz, H. F. Korth, and S. Sudarshan. *Database System Concepts* (4th ed.). McGraw-Hill, 2002.

[STA98] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 343–354, Seattle, WA, June 1998.

[Tho97] E. Thomsen. *OLAP Solutions: Building Multidimensional Information Systems*. John Wiley & Sons, 1997.

[TM05] Z. Tang and J. MacLennan. *Data Mining with SQL Server 2005*. John Wiley & Sons, 2005.

[TMK05] Z. Tang, J. Maclennan, and P. P. Kim. Building data mining solutions with OLE DB for DM and XML analysis. *SIGMOD Record*, 34:80–85, June 2005.

[Tuf83] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[Tuf90] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.

[Tuf97] E. R. Tufte. *Visual Explanations : Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.

[WB98] C. Westphal and T. Blaxton. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley & Sons, 1998.

[WF05] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, 2005.

[WI98] S. M. Weiss and N. Indurkhya. *Predictive Data Mining.* Morgan Kaufmann, 1998.

[WK91] S. M. Weiss and C. A. Kulikowski. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* Morgan Kaufman, 1991.