Chapter 1

INTRODUCTION TO KNOWLEDGE DISCOVERY IN DATABASES

Oded Maimon
Department of Industrial Engineering
Tel-Aviv University
maimon@eng.tau.ac.il

Lior Rokach
Department of Industrial Engineering
Tel-Aviv University
liorr@eng.tau.ac.il

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. *Data Mining* (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction.

The accessibility and abundance of data today makes knowledge discovery and Data Mining a matter of considerable importance and necessity. Given the recent growth of the field, it is not surprising that a wide variety of methods is now available to the researchers and practitioners. No one method is superior to others for all cases. The handbook of Data Mining and Knowledge Discovery from Data aims to organize all significant methods developed in the field into a coherent and unified catalog; presents performance evaluation approaches and techniques; and explains with cases and software tools the use of the different methods.

The goals of this introductory chapter are to explain the KDD process, and to position DM within the information technology tiers. Research and devel-

opment challenges for the next generation of the science of KDD and DM are also defined. The rationale, reasoning and organization of the handbook are presented in this chapter. In this chapter there are six sections followed by a brief reference primer list containing leading papers, books, conferences and journals in the field:

1. The KDD Process
2. Taxonomy of Data Mining Methods
3. Data Mining within the Complete Decision Support System
4. KDD & DM Research Opportunities and Challenges
5. KDD & DM Trends
6. The Organization of the Handbook

The special recent aspects of data availability that are promoting the rapid development of KDD and DM are the electronically readiness of data (though of different types and reliability). The internet and intranet fast development in particular promote data accessibility. Methods that were developed before the Internet revolution considered smaller amounts of data with less variability in data types and reliability.

Since the information age, the accumulation of data has become easier and storing it inexpensive. It has been estimated that the amount of stored information doubles every twenty months. Unfortunately, as the amount of electronically stored information increases, the ability to understand and make use of it does not keep pace with its growth. Data Mining is a term coined to describe the process of sifting through large databases for interesting patterns and relationships. The studies today aim at evidence-based modeling and analysis, as is the leading practice in medicine, finance and many other fields.

The data availability is increasing exponentially, while the human processing level is almost constant. Thus the gap increases exponentially. This gap is the opportunity for the KDD\DM field, which therefore becomes increasingly important and necessary.

1. The KDD Process

The knowledge discovery process (Figure 1.1) is iterative and interactive, consisting of nine steps.

Note that the process is iterative at each step, meaning that moving back to previous steps may be required. The process has many “artistic” aspects in the sense that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type. Thus it is required to understand the process and the different needs and possibilities in each step.

Taxonomy is appropriate for the Data Mining methods and is presented in the next section.

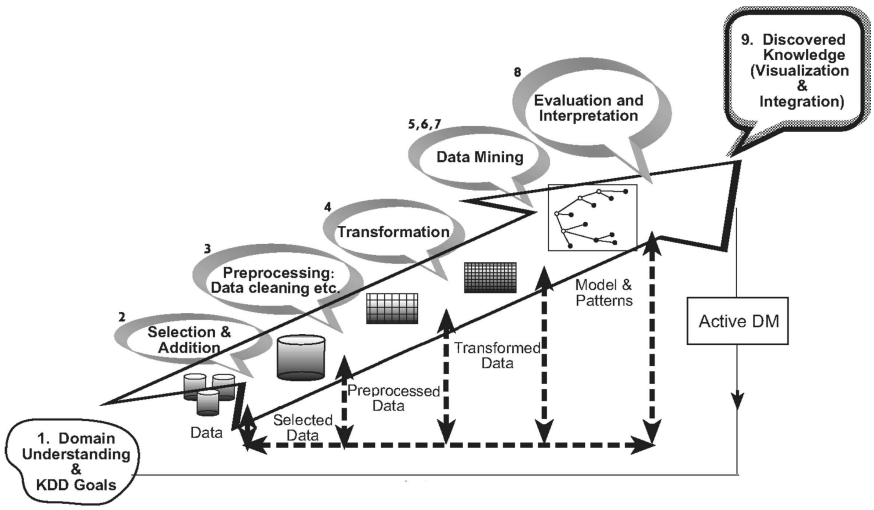

Figure 1.1. The Process of Knowledge Discovery in Databases.

The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. Then the loop is closed - the Active Data Mining part starts (which is beyond the scope of this book and the process defined here). As a result, changes would have to be made in the application domain (such as offering different features to mobile phone users in order to reduce churning). This closes the loop, and the effects are then measured on the new data repositories, and the KDD process is launched again.

Following is a brief description of the nine-step KDD process, starting with a managerial step:

1. Developing an understanding of the application domain This is the initial preparatory step. It prepares the scene for understanding what should be done with the many decisions (about transformation, algorithms, representation, etc.). The people who are in charge of a KDD project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior knowledge). As the KDD process proceeds, there may be even a revision of this step.

Having understood the KDD goals, the preprocessing of the data starts, defined in the next three steps (note that some of the methods here are

similar to Data Mining algorithms, but are used in the preprocessing context):

2. **Selecting and creating a data set on which discovery will be performed.** Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From this respect, the more attributes are considered, the better. On the other hand, to collect, organize and operate complex data repositories is expensive and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. This starts with the best available data set and later expands and observes the effect in terms of knowledge discovery and modeling.

3. **Preprocessing and cleansing.** In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. There are many methods explained in the handbook, from doing nothing to becoming the major part (in terms of time consumed) of a KDD project in certain projects. It may involve complex statistical methods or using a Data Mining algorithm in this context. For example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors. In any case, studying the aspects is important and often revealing by itself, regarding enterprise information systems.

4. **Data transformation.** In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step can be crucial for the success of the entire KDD project, and it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. In marketing, we may need to consider effects beyond our control as well as

efforts and temporal issues (such as studying the effect of advertising accumulation). However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed.

Having completed the above four steps, the following four steps are related to the Data Mining part, where the focus is on the algorithmic aspects employed for each project:

5. **Choosing the appropriate Data Mining task.** We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta-learning for the particular set of available data.

6. **Choosing the Data Mining algorithm.** Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers). For example, in considering precision versus understandability, the former is better with neural networks, while the latter is better with decision trees. For each strategy of meta-learning there are several possibilities of how it can be accomplished. Meta-learning focuses on explaining what causes a Data Mining algorithm to be successful or not in a particular problem. Thus, this approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate. Each algorithm has parameters and tactics of learning (such as ten-fold cross-validation or another division for training and testing).

7. **Employing the Data Mining algorithm.** Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.

8. **Evaluation.** In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step.

Here we consider the preprocessing steps with respect to their effect on the Data Mining algorithm results (for example, adding features in Step 4, and repeating from there). This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage.

The last step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining:

9. **Using the discovered knowledge.** We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we may make changes to the system and measure the effects. Actually the success of this step determines the effectiveness of the entire KDD process. There are many challenges in this step, such as loosing the "laboratory conditions" under which we have operated. For instance, the knowledge was discovered from a certain static snapshot (usually sample) of the data, but now the data becomes dynamic. Data structures may change (certain attributes become unavailable), and the data domain may be modified (such as, an attribute may have a value that was not assumed before).

2. Taxonomy of Data Mining Methods

There are many methods of Data Mining used for different purposes and goals. Taxonomy is called for to help in understanding the variety of methods, their interrelation and grouping. It is useful to distinguish between two main types of Data Mining: verification-oriented (the system verifies the user's hypothesis) and discovery-oriented (the system finds new rules and patterns autonomously). Figure 1.2 presents this taxonomy.

Discovery methods are those that automatically identify patterns in the data. The discovery method branch consists of prediction methods versus description methods. Descriptive methods are oriented to data interpretation, which focuses on understanding (by visualization for example) the way the underlying data relates to its parts. Prediction-oriented methods aim to build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample. It also develops patterns which form the discovered knowledge in a way which is understandable and easy to operate upon. Some prediction-oriented methods can also help provide understanding of the data.

Most of the discovery-oriented Data Mining techniques (quantitative in particular) are based on inductive learning, where a model is constructed, explicitly or implicitly, by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future unseen examples.

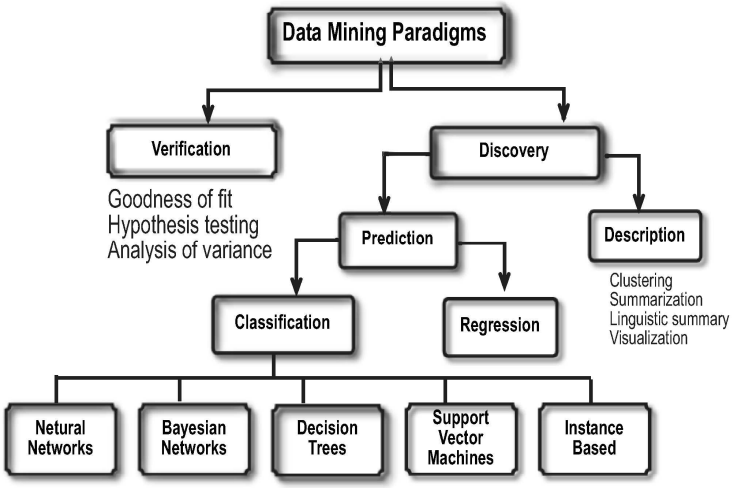

Figure 1.2. Data Mining Taxonomy.

Verification methods, on the other hand, deal with the evaluation of a hypothesis proposed by an external source (like an expert etc.). These methods include the most common methods of traditional statistics, like goodness of fit test, tests of hypotheses (e.g., t-test of means), and analysis of variance (ANOVA). These methods are less associated with Data Mining than their discovery-oriented counterparts, because most Data Mining problems are concerned with discovering an hypothesis (out of a large set of hypotheses), rather than testing a known one. Much of the focus of traditional statistical methods is on model estimation as opposed to one of the main objectives of Data Mining: model identification and construction, which is evidence based (though overlap occurs).

Another common terminology, used by the machine-learning community, refers to the prediction methods as supervised learning, as opposed to unsupervised learning. Unsupervised learning refers to modeling the distribution of instances in a typical, high-dimensional input space.

Unsupervised learning refers mostly to techniques that group instances without a prespecified, dependent attribute. Thus the term "unsupervised learning" covers only a portion of the description methods presented in Figure 1.2. For instance, it covers clustering methods but not visualization methods.

Supervised methods are methods that attempt to discover the relationship between input attributes (sometimes called independent variables) and a target attribute sometimes referred to as a dependent variable). The relationship discovered is represented in a structure referred to as a model. Usually models

describe and explain phenomena, which are hidden in the data set and can be used for predicting the value of the target attribute knowing the values of the input attributes. The supervised methods can be implemented on a variety of domains, such as marketing, finance and manufacturing.

It is useful to distinguish between two main supervised models: classification models and regression models. The latter map the input space into a real-valued domain. For instance, a regressor can predict the demand for a certain product given its characteristics. On the other hand, classifiers map the input space into predefined classes. For example, classifiers can be used to classify mortgage consumers as good (fully payback the mortgage on time) and bad (delayed payback), or as many target classes as needed. There are many alternatives to represent classifiers. Typical examples include, support vector machines, decision trees, probabilistic summaries, or algebraic function.

3. Data Mining within the Complete Decision Support System

Data Mining methods are becoming part of integrated Information Technology (IT) software packages. Figure 1.3 illustrates the three tiers of the decision support aspect of IT. Starting from the data sources (such as operational databases, semi- and non-structured data and reports, Internet sites etc.), the first tier is the data warehouse, followed by OLAP (On Line Analytical Processing) servers and concluding with analysis tools, where Data Mining tools are the most advanced.

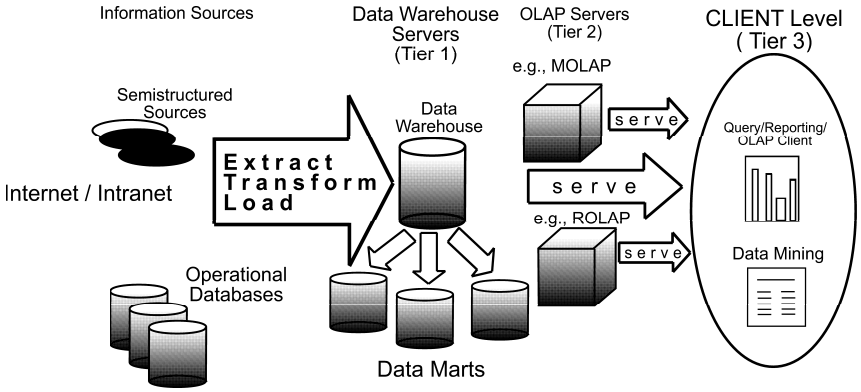

Figure 1.3. The IT Decision Support Tiers.

The main advantage of the integrated approach is that the preprocessing steps are much easier and more convenient. Since this part is the major burden for the KDD process (and often consumes most of the KDD project time), this

industry trend is very important for expanding the use and utilization of Data Mining. However, the risk of the integrated IT approach comes from the fact that those DM techniques are much more complex and intricate than OLAP, for example, so the users need to be trained appropriately. This handbook shows the variety of strategies, techniques and evaluation measurements.

We can naively distinguish among three levels of analysis. The simplest one is achieved by report generators (for example, presenting all claims that occurred because of a certain cause last year, such as car theft). We then proceed to OLAP multi-level analysis (for example presenting the ten towns where there was the highest increase of vehicle theft in the last month as compared to with the month before). Finally a complex analysis is carried out in discovering the patterns that predict car thefts in these cities, and what might occur if anti theft devices were installed. The latter is based on modeling of the phenomena, where the first two levels are ways of data aggregation and fast manipulation.

4. KDD & DM Research Opportunities and Challenges

Empirical comparison of the performance of different approaches and their variants in a wide range of application domains has shown that each performs best in some, but not all, domains. This phenomenon is known as the selective superiority problem, which means, in our case, that no induction algorithm can be the best in all possible domains. The reason is that each algorithm contains an explicit or implicit bias that leads it to prefer certain generalizations over others, and it will be successful only as long as this bias matches the characteristics of the application domain.

Results have demonstrated the existence and correctness of this “no free lunch theorem”. If one inducer is better than another in some domains, then there are necessarily other domains in which this relationship is reversed. This implies in KDD that for a given problem a certain approach can yield more knowledge from the same data than other approaches.

In many application domains, the generalization error (on the overall domain, not just the one spanned in the given data set) of even the best methods is far above the training set, and the question of whether it can be improved, and if so how, is an open and important one. Part of the answer to this question is to determine the minimum error achievable by any classifier in the application domain (known as the optimal Bayes error). If existing classifiers do not reach this level, new approaches are needed. Although this problem has received considerable attention, no generally reliable method has so far been demonstrated. This is one of the challenges of the DM research – not only to solve it, but even to quantify and understand it better. Heuristic methods can then be compared absolutely and not just against each other.

A subset of this generalized study is the question of which inducer to use for a given problem. To be even more specific, the performance measure needs to be defined appropriately for each problem. Though there are some commonly accepted measures it is not enough. For example, if the analyst is looking for accuracy only, one solution is to try each one in turn, and by estimating the generalization error, to choose the one that appears to perform best. Another approach, known as multi-strategy learning, attempts to combine two or more different paradigms in a single algorithm.

The dilemma of what method to choose becomes even greater if other factors such as comprehensibility are taken into consideration. For instance, for a specific domain, neural networks may outperform decision trees in accuracy. However from the comprehensibility aspect, decision trees are considered superior. In other words, in this case even if the researcher knows that neural network is more accurate, the dilemma of what methods to use still exists (or maybe to combine methods for their separate strength).

Induction is one of the central problems in many disciplines such as machine learning, pattern recognition, and statistics. However the feature that distinguishes Data Mining from traditional methods is its scalability to very large sets of varied types of input data. Scalability means working in an environment of high number of records, high dimensionality, and a high number of classes or heterogeneousness. Nevertheless, trying to discover knowledge in real life and large databases introduces time and memory problems.

As large databases have become the norm in many fields (including astronomy, molecular biology, finance, marketing, health care, and many others), the use of Data Mining to discover patterns in them has become potentially very beneficial for the enterprise. Many companies are staking a large part of their future on these "Data Mining" applications, and turn to the research community for solutions to the fundamental problems they encounter.

While a very large amount of available data used to be the dream of any data analyst, nowadays the synonym for "very large" has become "terabyte" or "pentabyte", a barely imaginable volume of information. Information-intensive organizations (like telecom companies and financial institutions) are expected to accumulate several terabytes of raw data every one to two years.

High dimensionality of the input (that is, the number of attributes) increases the size of the search space in an exponential manner (known as the "Curse of Dimensionality"), and thus increases the chance that the inducer will find spurious classifiers that in general are not valid.

There are several approaches for dealing with a high number of records including: sampling methods, aggregation, massively parallel processing, and efficient storage methods.

5. KDD & DM Trends

This handbook covers the current state-of-the-art status of Data Mining. The field is still in its early stages in the sense that further basic methods are being developed. The art expands but so does the understanding and the automation of the nine steps and their interrelation. For this to happen we need better characterization of the KDD problem spectrum and definition.

The terms KDD and DM are not well-defined in terms of what methods they contain, what types of problem are best solved by these methods, and what results to expect. How are KDD\DM compared to statistics, machine learning, operations research, etc.? If subset or superset of the above fields? Or an extension\adaptation of them? Or a separate field by itself? In addition to the methods – which are the most promising fields of application and what is the vision KDD\DM brings to these fields? Certainly we already see the great results and achievements of KDD\DM, but we cannot estimate their results with respect to the potential of this field. All these basic analysis have to be studied and we see several trends for future research and implementation, including:

- Active DM – closing the loop, as in control theory, where changes to the system are made according to the KDD results and the full cycle starts again. Stability and controllability which will be significantly different in these type of systems, need to be well-defined.
- Full taxonomy – for all the nine steps of the KDD process. We have shown a taxonomy for the DM methods, but a taxonomy is needed for each of the nine steps. Such a taxonomy will contain methods appropriate for each step (even the first one), and for the whole process as well.
- Meta-algorithms – algorithms that examine the characteristics of the data in order to determine the best methods, and parameters (including decompositions).
- Benefit analysis – to understand the effect of the potential KDD\DM results on the enterprise.
- Problem characteristics – analysis of the problem itself for its suitability to the KDD process.
- Expanding the database for Data Mining inference to include also data from pictures, voice, video, audio, etc. This will require adapting and developing new methods (for example, for comparing pictures using clustering and compression analysis).
- Distributed Data Mining – The ability to seamlessly and effectively employ Data Mining methods on databases that are located in various sites.

This problem is especially challenging when the data structures are heterogeneous rather than homogeneous.

- Expanding the knowledge base for the KDD process, including not only data but also extraction from known facts to principles (for example, extracting from a machine its principle, and thus being able to apply it in other situations).

- Expanding Data Mining reasoning to include creative solutions, not just the ones that appears in the data, but being able to combine solutions and generate another approach.

The last two are beyond the scope of KDD\DM definition here, and this is the last point, to define KDD\DM for the next phase of this science.

6. The Organization of the Handbook

This handbook is organized in eight parts. Starting with the KDD process, through to part six, the book presents a comprehensive but concise description of different methods used throughout the KDD process. Each part describes the classic methods as well as the extensions and novel methods developed recently. Along with the algorithmic description of each method, the reader is provided with an explanation of the circumstances in which this method is applicable and the consequences and the trade-offs of using the method including references for further readings. Part seven presents real-world case studies and how they can be solved. The last part surveys some software and tools available today.

The first part is about preprocessing methods, starting with data cleansing, followed by the handling of missing attributes (Chapters 2, 3). Following issues in feature extraction, selection and dimensional reductions are discussed (Chapters 4, 5). These chapters are followed by discretization methods and outlier detection (Chapters 6, 7). This covers the preprocessing methods (Steps 3, 4 of the KDD process).

The Data Mining methods starts in the second part with the introduction and the very often-used decision tree method (Chapters 8, 9), followed by other classical methods, such as Bayesian networks, regression (in the Data Mining framework), support vector machines and rule induction (Chapters 10-13).

The third part of the handbook considers the unsupervised methods, starting with visualization (suited for high dimensional data bases) in Chapter 14. Then the important methods of clustering, association rules and frequent set mining are treated (Chapters 15, 16 and 17). Finally in this part two more topics are presented for constraint-based Data Mining and link analysis, in Chapters 18 and 19.

The fourth part is about methods termed soft computing, which include fuzzy logic, evolutionary algorithms, reinforcement learning, neural networks and ending with granular computing and rough sets (covered in Chapters 20-24).

Having established the foundation, we now proceed with supporting methods needed for Data Mining in the fifth part, starting with statistical methods for Data Mining (Chapter 25) followed with logic, wavelets and fractals (Chapters 26, 27, 28). The topics of interestingness and quality assessment are discussed in Chapters 29 and 30. In this part, two more important issues are presented considering model comparison in Chapter 31 and Data Mining query language in Chapter 32.

Having covered the basics, we proceed with advanced methods in the sixth part, which covers topics like meta-learning in Chapter 33, bias vs. variance in Chapter 34 and rare cases in Chapter 35. Additional topics in the next sixteen chapters (from 36 to 51) include mining high dimensional data, text mining and information extraction, spatial methods, imbalanced data sets, relational Data Mining, web mining, causality, ensemble and decomposition methods, information fusion, parallel and grid-based, collaborative and organizational Data Mining.

With all the methods described so far, the next section, the seventh, is concerned with applications for medicine, biology, manufacturing, design, telecommunication and finance (Chapters 52 to 57). The next topic is about intrusion detection with Data Mining methods, followed by software testing, CRM application and target marketing the last application selected for presentation in this handbook (Chapters 58 and 61).

The last and final part of this handbook deals with software tools. This part is not a complete survey of the software available, but rather a selected representative from different types of software packages that exist in today's market. This section begins by public domain open source research-type software, Weka (Chapter 62), followed by two integrated tools (Data Mining tools integrated with database, data warehouse and the entire support software environment) represented by Oracle (Chapter 63) and Microsoft (Chapter 64). Then four specific software systems are presented (Chapters 65 to 68). These software systems employ various Data Mining methods discussed in detail previously in the book.

7. References Principles

This reference section presents major publications in knowledge discovery and Data Mining. In addition some major conferences and journals, where further information can be obtained, are listed.

7.1 Papers

1. Agrawal, R. and Srikant, R., Fast algorithms for mining association rules. In Bocca, J., Jarke, M., and Zaniolo, C., editors, Proceedings 20th International Conference on Very Large Data Bases, pages 487–499, 1994.

2. Agrawal, R., Faloutsos, C., Swami, A. Efficient Similarity Search in Sequence Data bases. International Conference on Foundations of Data Organization (FODO); Chicago, pages 69–84, 1993.

3. Aha, D., Kibler, W., Albert, M. K., Instance based learning algorithms. Machine Learning, 6:37–66, 1991.

4. Bauer, E., Kohavi, R., An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, 36:105-139, 1999.

5. Bayardo, J., Efficiently mining long patterns from databases. In In A. T. Laura M. Haas, editors, Proceedings of ACM SIGMOD'98, pages 85–93, Seattle, WA, USA, 1998.

6. Benjamini, Y. and Hochberg, Y., Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal Royal Statistical Society, Ser. B, 57:289-300, 1995.

7. Breiman, L., Bagging predictors, Machine Learning, 24 (2) : 123–140, 1996.

8. Brin, S., Motwani, R., Silverstein, C., Beyond market baskets: Generalizing association rules to correlations. In Valduriez P., Korth H. F., Proceedings of ACM SIGMOD'97, pages 265–276, Tucson, AZ, USA, 1997.

9. Calders, T., Goethals, B., Mining all non-derivable frequent itemsets. In Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02) Lecture Notes in Artificial Intelligence, volume 2431 of LNCS, pages 74–85. Springer-Verlag, 2002.

10. Clark, P., Niblett, T., The CN2 induction algorithm. Machine Learning, 3(4): 261–283, 1989.

11. Dehaspe, L., Toivonen, H., Discovery of frequent Datalog patterns. Data Mining and Knowledge Discovery, 3:7-36, 1999.

12. Dietterich, T., An Empirical Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. Machine Learning, 40(2):139–157, 2000.

13. Domingos, P., Hulten, G., Mining High-Speed Data Streams. Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, pages 71–80, 2000.

14. Fawcett, T., Provost, F., Adaptive fraud detection. Data-mining and Knowledge Discovery, 1(3):291–316, 1997.

15. Freund, Y., Schapire, R., A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55:119-139, 1997.

16. Holte, R. C., Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning, 11:63–90, 1993.

17. Maimon, O., Kandel A., Last M., Information–Theoretic Fuzzy Approach to Data Reliability and Data Mining. Fuzzy Sets and Systems, 117:183–194, 2001.

18. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., Efficient Mining of association rules using closed itemset lattices. Information Systems, 24(1):25–46, 1999.

19. Rokach, L., Maimon, O., Theory and Application of Attribute Decomposition, Proceedings of the First IEEE International Conference on Data Mining, IEEE Computer Society Press, pp. 473–480, 2001.

20. Shafer, J., Agrawal, R., Mehta, M., SPRINT: A Scalable Parallel Classifier for Data Mining. Proceedings of the 22nd International Conference on Very Large Databases; Bombay, pages 544–555, 1996.

21. Zaki, M., Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3):372–390, 2000.

7.2 Books

1. Backer, E., Computer-Assisted Reasoning in Cluster Analysis, Prentice Hall, 1995.

2. Barnett, V., Lewis, T., Outliers in Statistical Data. John Wiley, 1994.

3. Berry, M. J. A., Linoff, G., Data Mining Techniques: For Marketing, Sales, and Customer Support. New York: Wiley, 2004.

4. Bishop, M., Neural Networks for Pattern Recognition. Oxford: Oxford University Press, 1995.

5. Breiman, L., Friedman, J. H. , Olshen, R. A., Stone, C. J., Classification and Regression Trees. Wadsworth, Belmont, Ca., 1984.

6. Džeroski, S., Lavrač, N., editors, Relational Data Mining: Inductive Logic Programming for Knowledge Discovery in Databases. Springer-Verlag, 2001.

7. Dasu, T., and Johnson, T., Exploratory Data Mining and Data Cleaning. New York: John Wiley & Sons, 2003.

8. Duda, R. O., Hart, P. E., Pattern Classification and Scene Analysis. Wiley, New York, NY, 1973.

9. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., editors, Advances in Knowledge Discovery and Data Mining, MIT Press, 1996.

10. Freitas, A. A., Lavington, S. H., Mining Very Large Databases with Parallel Processing, Kluwer, 1998.

11. Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, San Diego, California, 1990.

12. Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

13. Hand, D. J., Construction and Assessment of Classification Rules, Wiley, New York, NY, 1997.

14. Maimon, O., Last, M., Knowledge Discovery and Data Mining: The Info-Fuzzy Network (IFN) Methodology, Kluwer Academic Publishers, 2001.

15. Maimon, O., Rokach, L., Decomposition Methodology for Knowledge Discovery and Data Mining Theory and Applications, World Scientific Press, 2005.

16. Mitchell, T., Machine Learning, McGraw-Hill, 1997

17. Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference, Morgan Kaufmann, San Francisco, CA, 1988.

18. Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.

19. Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

20. Witten, I. H., Frank E., Data Mining, Morgan Kaufmann, New York, 2000.

7.3 Main Conferences

1. ACM Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining (SIGKDD)
2. ACM Special Interest Group on Management of Data, International Conference on Management of Data (SIGMOD)
3. European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)
4. European Conference on Machine Learning (ECML)
5. IEEE International Conference on Data Mining (ICDM)
6. International Conference on Very Large Databases (VLDB)
7. International Conference on Machine Learning (ICML)

7.4 Main Journals

1. Data Mining and Knowledge Discovery
2. IEEE Transactions on Knowledge and Data Engineering
3. IEEE Transactions on Pattern Analysis and Machine Intelligence
4. Information Systems
5. International Journal of Pattern Recognitions and Applied Intelligence (IJPRAI),
6. Knowledge and Information Systems
7. Machine Learning