Three Perspectives of Data Mining

Zhi-Hua Zhou*

National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China

Abstract

This paper reviews three recent books on data mining written from three different perspectives, i.e. databases, machine learning, and statistics. Although the exploration in this paper is suggestive instead of conclusive, it reveals that besides some common properties, different perspectives lay strong emphases on different aspects of data mining. The emphasis of the database perspective is on *efficiency* because this perspective strongly concerns the whole discovery process and huge data volume. The emphasis of the machine learning perspective is on *effectiveness* because this perspective is heavily attracted by substantive heuristics working well in data analysis although they may not always be useful. As for the statistics perspective, its emphasis is on *validity* because this perspective cares much for mathematical soundness behind mining methods.

Keywords: Data mining; Databases; Machine learning; Statistics

1. Introduction

The rapid progress in digital data acquisition and storage technology has led to the fast-growing tremendous and amount of data stored in databases, data warehouses, or other kinds of data repositories such as the World Wide Web. Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult, if not impossible, for human beings to extract them without powerful tools. In order to relieve such a *data rich but information poor* plight, during the late 1980s, a new discipline named data mining emerged, which devotes itself to extracting knowledge from huge volumes of data, with the help of the ubiquitous modern computing device, i.e. computer.

Due to its interdisciplinary nature, data mining has received contributions from a lot of disciplines such as databases, machine learning, statistics, information retrieval, data visualization, parallel and distributed computing, *etc*. The first three in the list, i.e. database, machine learning, and statistics, are undoubtedly the primary contributors. It is obvious that without the powerful data management techniques donated by the database community and the practical data analysis techniques donated by the machine learning community, data mining would be seeking a needle in the haystack. It is interesting that even recently, a leading statistician raised a cry that the statistics community should embrace data mining [4], which exposes that this community had not yet taken data mining seriously at least at that time. However, it is still clear that without the solid theoretical foundation donated by the statistics community, data mining will be building a castle in the air.

An interesting issue rises. That is, what are the differing roles of data mining in the view of these different disciplines, considering that different disciplines have their own backgrounds? Fortunately, this

* Tel.: +86-25-359-3163, fax: +86-25-330-0710.
E-mail addresses: zhouzh@nju.edu.cn (Z.-H. Zhou).

issue can be partially addressed now because we have three academic data mining books written by leading experts from three different perspectives.

The first one, J. Han and M. Kamber's *Data Mining: Concepts and Techniques* (Morgan Kaufmann, 2001) [5], as the authors admitted: "*we present the material in this book from a database perspective*" ([5], pp.xix), was written from a database perspective. The second one, I. H. Witten and E. Frank's *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Morgan Kaufmann, 2000) [9], as the title of the book suggests, was obviously written from a machine learning perspective. Finally, *Principles of Data Mining* (MIT Press, 2001) [6] by D. Hand, H. Mannila, and P. Smyth, was written from the statistics perspective, as the authors said: "*we take a fairly strong statistical view of data mining in this text*" ([6], pp.xxx).

These books provide good sources for exploring the different perspectives of data mining. However, before any concrete discussion, I will mention that although these books are excellent, they are not the only excellent books on data mining. Choosing them for this exploration is mainly due to my own taste. More importantly, the exploration in this paper should be regarded as suggestive instead of conclusive, because the outcome may only reveal biases of the authors instead of the opinion of the corresponding research communities; similarly some discussions may only reflect my own opinion. This is the reason why I said the exploration of the issue above could only be *partially* addressed.

The rest of this paper is organized as follows. Section 2 briefly introduces the contents of the reviewed books. Section 3 explores the database, machine learning, and statistics perspectives. Section 4 discusses some other traits of the books. Finally, Section 5 concludes.

2. Contents

J. Han and M. Kamber's book is organized in a taxonomic style where any given technique is presented just once. This style is very popular in academic books. The book contains 10 chapters. Chapter 1 provides an introduction to data mining, where categorizations of data mining tasks and data mining systems are presented. Chapter 2 focuses on data warehouse and on-line analytical processing. Chapter 3 presents techniques for preprocessing the data prior to mining, including cleaning, integration, transformation, and reduction. Chapter 4 introduces primitives for defining data mining tasks, data mining query language DMQL, and the architectures of data mining systems. Chapter 5 is devoted to descriptive data mining, including characterization and discrimination. Chapter 6 focuses on association rule mining. Chapter 7 presents techniques for classification and regression. Chapter 8 looks at cluster analysis. Chapter 9 focuses on data mining in advanced data repository systems, such as multimedia data mining and web mining. Finally, Chapter 10 discusses applications and challenges of data mining.

I. H. Witten and E. Frank's book is organized in a layered style where a technique may be presented at several places with the predecessor detailed by the successor. Such a style allows readers easily either to read about only the basics or the full details. However, it also makes the book look a little messy: readers who want to know all about a specific technique must jump between the chapters.

This book contains 9 chapters. Chapter 1 provides an introduction to data mining, where vivid examples are employed to illustrate the usefulness of machine learning in data mining. Chapter 2 focuses on the input knowledge representation, including concepts, examples, and attributes. Chapter 3 is devoted to the output knowledge representation, including decision trees, classification rules, *etc*. Chapter 4 briefly presents some popular prediction methods, plus a section on association mining. Chapter 5 focuses on evaluating or comparing prediction performance. Chapter 6 details most methods presented in Chapter 4, plus a section on clustering. Chapter 7 provides methods for attribute selection, discretization, data cleaning, and briefly

introduces ensemble learning methods such as Bagging and Boosting. Chapter 8 presents the WEKA software package attached in this book. Finally, Chapter 9 briefly discusses some advanced topics such as text mining and web mining.

D. Hand *et al.*'s book is organized in a component-based style where a technique may be presented at several places but at each place only one component (out of four) of the technique may be presented. Such a style is very novel. On one hand, it may help the readers understand that data mining algorithms are based on some very general and systematic principles. But on the other hand, the readers may be able to see full data mining algorithms only after reading all (or most) of the book.

This book contains 14 chapters. Chapter 1 provides an introduction to data mining, where four components of data mining algorithms are distinguished. Chapter 2 focuses on the measurement and data quality. Chapter 3 presents visualization and some dimensional reduction methods such as principal component analysis and multidimensional scaling. Chapter 4 provides some probability tools for dealing with uncertainty. Chapter 5 uses four different kinds of algorithms to show how data mining algorithms could be described as the combination of components. Chapter 6, Chapter 7, and Chapter 8 present three components of data mining algorithms, i.e. models and patterns, score functions, and search and optimization methods, respectively. Chapter 9 focuses on descriptive modeling, including density estimation and cluster analysis. Chapter 10 and Chapter 11 are devoted to classification and regression modeling, respectively. Chapter 12 presents the fourth component of data mining algorithms, i.e. data management strategy. Chapter 13 devotes to the discovery of patterns, mainly focuses on rules. Finally, Chapter 14 introduces content-based retrieval.

Two of the books have supplementary appendices: J. Han and M. Kamber's book has two appendixes introducing Microsoft's OLE DB for Data Mining and the authors' DBMiner, respectively, while D. Hand *et al.*'s book has an appendix briefly reviewing univariate random variables and common probability distributions.

3. Perspectives

A novice at data mining may panic at the tremendous variance of these three books: are they really for the same discipline? In order to explain the variance, we should examine what is meant by the term *data mining*.

In J. Han and M. Kamber's book, data mining was defined as "*the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories* " ([5], pp.7). In I. H. Witten and E. Frank's book, it was defined as "*the extraction of implicit, previously unknown, and potentially useful information from data*" ([9], pp.xix). While, in D. Hand *et al.*'s book, it was "*the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner*" ([6], pp.1).

To make the difference more evident, let's consider another term: KDD, i.e. knowledge discovery in databases. The classic definition of KDD is "*the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" [3]. With regard to the relationship between data mining and KDD, two popular usages exist. One is to regard them as synonyms, while the other is to regard the former as a step, although maybe the core step, in a sequence of steps of the latter.

Now let's look again at our three books. J. Han and M. Kamber's book definitely adopts the first usage, as it said: "… *the term data mining is becoming more popular than the longer term of knowledge discovery in databases. Therefore, in this book, we choose to use the term data mining.*" ([5], pp.7) D. Hand *et al.*'s book adopts the second usage, as it admitted: "*The KDD process involves several stages … In this text we*

will focus primarily on data mining algorithms rather than the overall process." ([6], pp.3) The usage adopted by I. H. Witten and E. Frank's book is quite blurry because it does not mention KDD at all! However, through examining the seven steps of KDD ([5], pp.7), we may find that this book adopts a trade-off between the above two usages. Its interpretation of data mining is broader than that of D. Hand *et al.*'s but narrower than that of J. Han and M. Kamber's in that although it does not pay much attention to data integration, selection, and knowledge presentation, it does involve most steps of KDD.

So, what is the consequence of above difference?

Since J. Han and M. Kamber's book concerns itself with the whole discovery process, it contains a separate chapter for data preprocessing ([5], Chapter 3) and addresses the issue of how to present the mining results (e.g. [5], pp.190). Moreover, it uses a whole chapter to discuss primitives, query language, and the architectures of data mining systems ([5], Chapter 4), which is important for building systems supporting the whole process efficiently. Since its definition of data mining ([5], pp.7) explicitly stresses that the processing object is "*large amounts of data*", a whole chapter is devoted to efficient management of huge volumes of data ([5], Chapter 2). The AOI based characterization and discrimination ([5], Chapter 5) also utilize powerful data management techniques ([5], pp.187). Such an interest in data management is distinctive, as D. Hand *et al.* indicated: "*very few algorithms published outside the database literature provide any explicit guidance on data management for large data sets.*" ([6], pp.150)

Instead of covering the whole discovery process, the objective of I. H. Witten and E. Frank's book is "*to introduce the tools and techniques for machine learning that are used in data mining*" ([9], pp.xx). Since data mining is a non-trivial process, the effectiveness of the tools and techniques may be the prime consideration. Therefore, this book devotes a whole chapter ([9], Chapter 5) to the evaluation and comparison of the effectiveness of data mining algorithms. It also presents a very simple but quite effective rule induction method, i.e. 1R ([9], pp.78). Moreover, it spends almost one third of the whole length of the book ([9], Chapter 6 and 7) in presenting heuristics or tricks that may enhance the effectiveness of the basic algorithms, although the enhancements may not be always useful. Such an endeavor for improving practical effectiveness results in the WEKA software package attached in the book, whose description occupies a whole chapter ([9], Chapter 8).

Since D. Hand *et al.*'s book regards data mining as the core step of KDD and focuses on only this step, it is able to go very deep into the mining methods. Evidently, this book bears a strong taste of favoring mathematical validity, which could be easily recognized from the abundant mathematical materials it provides. For example, although most data mining books admit that data used in data mining is "*observational*" ([6], pp.1), few addresses the issue of distinguishing between the opportunity sample and random sample as this book does ([6], pp.21, pp.48). Another example is: although almost all data mining books regard sampling as a powerful tool to tackle the huge data volume, few present *sufficient statistics* behind sampling methods as this book does ([6], pp.19, pp.112, 425). Moreover, partitioning the structures to be discovered into models and patterns ([6], pp.9, Chapter 6), and categorizing data mining tasks according to such a partition ([6], pp.11), also demonstrate the mathematical rigor of this book.

Thus, from the difference in the coverage of these books, it could be perceived that the database, machine learning, and statistics perspectives of data mining put particular emphases on efficiency, effectiveness, and validity, respectively. Note that this does not mean the machine learning and statistics perspectives do not care for efficiency, or the database and statistics perspectives do not care for effectiveness, or the database and machine learning perspectives do not care for validity. What is claimed is that each of the three perspectives emphasizes some aspect much more than the other two perspectives do.

In fact, the emphases are so strong that even from the topics covered by all three books, their traces can be found. For example, on the topic of decision trees, J. Han and M. Kamber's book presents methods of

scaling basic algorithms for large datasets ([5], pp.292); I. H. Witten and E. Frank's book provides a trick to simplify decision trees through discarding misclassified examples and relearning ([9], pp.247); while D. Hand *et al.*'s book discusses the local piecewise model structure behind trees ([6], pp.174).

However, although those perspectives have different preferences, an important fact cannot be neglected. That is, they also share many common properties, as shown by the overlapped contents of the books.

It is worth mentioning that some other differences can also be found among the books. However, many of them do not particularly distinguish the perspectives, at least from my point of view. For example, different categorizations and diversified algorithms were presented for cluster analysis ([5], Chapter 8; [6], Section 9.3 to 9.6; [9], Section 6.6). But it may be more reasonable to attribute such a variance to the relative youth of the area of cluster analysis, as indicated in a recent paper [2]. Another example is the exclusion of neural networks in I. H. Witten and E. Frank's book due to the deficiency of neural networks in understandability ([9], pp.xxx). Since some machine learning researchers endeavor to improve the understandability of neural networks [8], and neural networks have already been applied to data mining [7], I incline to regarding the exclusion of neural networks as a preference of the authors instead of that of the machine learning community.

It is interesting to see that each of these books devotes a section to distinguishing between machine learning and data mining ([5], pp.218), machine learning and statistics ([9], pp.26), and statistics and data mining ([6], pp.18), respectively. (It seems that the difference between database and data mining/machine learning/statistics is so evident that there is no need to make a distinction.) From my own point of view, the fundamental difference between machine learning/statistics and data mining exists in the aspect of data volume being processed. While the fundamental difference between machine learning and statistics exists in the aspect of research methodology. That is, statisticians prefer theoretical soundness while machine learners require both theoretical soundness and experimental effectiveness. Note that, at least in my own opinion, the experimental aspect of the machine learning methodology owes much to the setup of the UCI Machine Learning Repository [1].

4. Miscellaneous

All three books are very excellent. They present such a comprehensive view of data mining that some very recently developed techniques, such as ensemble learning methods ([5], pp.324; [6], pp.358; [9], pp.250), are also discussed. The comprehensiveness is further enlarged because each book has a *References* (or *Bibliography*) section pointing to a great mass of literature. The *Further Reading* (or *Bibliographic Notes*) sections in the books may lead the readers to trace the literature eruditely although they are recommended from different perspectives.

As for the presentation, I. H. Witten and E. Frank's book is the most readable. Its narration is so lively that sometimes I even wonder that whether I am really reading an academic book. J. Han and M. Kamber's book presents many materials with a question-and-answer style, which makes it very easy to follow. Although D. Hand *et al.*'s book conveys tremendous mathematical materials, reading it is not dull because it employs many interesting or even impressive examples, e.g. zero blood pressure caused by missing values ([6], pp.57) and the explosion of the space shuttle Challenger related to the low temperature ([6], pp.386), for illustration. Moreover, D. Hand *et al.*'s book implicitly adopts a step-by-step style in presenting complicated materials, which may greatly help the readers master such material. For example, *bias* is at first informally involved when measurement is introduced ([6], pp.45), then formally presented when the properties of estimators are discussed ([6], pp.106), and further explored when the design of score functions is analyzed ([6], pp.221).

It is worth mentioning that some proverbs appearing in the books that are particularly apt for data miners, e.g. "*one person's noise could be another person's signal*" ([5], pp.381), "*simplicity-first*" ([9], pp.77), "*all models are wrong but some are useful*" ([6], pp.168).

All the books could be used as textbooks for classes. J. Han and M. Kamber provide a set of exercises for each chapter in the book, and a suite of slides at the book homepage (www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-489-8). I. H. Witten and E. Frank provide the software package WEKA, a suite of slides, exams, assignments, and even quizzes, at the book homepage (www.mkp.com/books_catalog/catalog.asp?ISBN =1-55860-552-5). Personally, for an introductory data mining course, I would suggest to use J. Han and M. Kamber's book as the main textbook, with I. H. Witten and E. Frank's book as the primary reading material and the source of course projects, and use D. Hand *et al.*'s book as the advanced reading material.

5. Concluding remarks

Data mining is not esoteric. In fact, the work of this paper is also a kind of data mining, that is, mining data mining books.

Interesting patterns come out from such a mining practice. That is, besides some common properties, different perspectives of data mining put strong emphases on different aspects. In particular, the database, machine learning, and statistics perspectives lay particular emphases on efficiency, effectiveness, and validity, of data mining, respectively.

We know the danger of overfitting the noise in data mining, especially when the training set is quite small. Since the data size used in my review is rather small (only three books), the chances of overfitting the individual preference of the authors (instead of really discovering the accent of the perspectives) may be high even though I have tried my best to be scrupulous. Moreover, some discussions in the review may only reflect my own inclination although I have tried my best to be unbiased. Therefore, it is better to regard the outcome of this exercise as suggestive instead of conclusive, and take them with great care.

Note that the mined patterns do not mean that the machine learning and statistics perspectives do not care for efficiency, or the database and statistics perspectives do not care for effectiveness, or the database and machine learning perspectives do not care for validity. What is claimed is that each of the three perspectives emphasizes some aspects more than the other two perspectives do. In fact, only when we simultaneously take all these three aspects (or more) into account, may we get successful data mining results.

Acknowledgements

The comments and suggestions from Prof. A. G. Cohn and Prof. D. Perlis greatly improved this review. The author was partially supported by the National Natural Science Foundation of China under grant no. 60105004, and the Natural Science Foundation of Jiangsu Province, China, under grant no. BK2001406.

References

[1] C. Blake, E. Keogh, C.J. Merz, UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.htm], Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[2] V. Estivill-Castro, Why so many clustering algorithms - a position paper, *SIGKDD Explorations* 4 (1) (2002) 65-75.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge discovery and data mining: towards a unifying

framework, in: *Proc. KDD-96*, Portland, OR, AAAI Press, Menlo Park, CA, 1996, pp.82-88.

[4] J. H. Friedman, Data mining and statistics: what is the connection? Keynote Speech of the *29th Symposium on the Interface: Computing Science and Statistics*, Houston, TX, 1997.

[5] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, 2001.

[6] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, CA, 2001.

[7] S. Mitra, S. K. Pal, P. Mitra, Data mining in soft computing framework: a survey, *IEEE Trans. Neural Networks* 13 (1) (2002) 3-14.

[8] A. B. Tickle, R. Andrews, M. Golea, J. Diederich, The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks, *IEEE Trans. Neural Networks* 9 (6) (1998) 1057-1068.

[9] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, CA, 2000.