Data Mining: An AI Perspective

Xindong Wu[1], *Senior Member, IEEE*

Abstract--**Data mining, or knowledge discovery in databases (KDD), is an interdisciplinary area that integrates techniques from several fields including machine learning, statistics, and database systems, for the analysis of large volumes of data. This paper reviews the topics of interest from the IEEE International Conference on Data Mining (ICDM) from an AI perspective. We discuss common topics in data mining and AI, including key AI ideas that have been used in both data mining and machine learning.**

Index Terms—**Data Mining, Artificial Intelligence, Machine Learning.**

I. The IEEE International Conference on Data Mining

DATA mining is a fast-growing area. The first Knowledge Discovery in Databases Workshop was held in August 1989, in conjunction with the 1989 International Joint Conference on Artificial Intelligence, and this workshop series became the International Conference on Knowledge Discovery and Data Mining in 1995. In 2003, there were a total of 15 data mining conferences, most of which are listed at http://www.kdnuggets.com/meetings/meetings-2003-past.html:

- Data Warehousing and Data Mining in Drug Development (January 13-14, 2003, Philadelphia, PA, USA)
- First Annual Forum on Data Mining Technology for Military and Government Applications (February 25-26, 2003, Washington DC, USA)
- SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology V (21-22 April 2003, http://www.spie.org/Conferences/Programs/03/or/conferences/index.cfm?fuseaction=5098)
- PAKDD-03: 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (April 30 - May 2, 2003, Seoul, Korea)
- SDM 03: 3rd SIAM International Conference on Data Mining (May 1-3, 2003, San Francisco, CA, USA)
- MLDM 2003: Machine Learning and Data Mining (July 5-7, 2003, Leipzig, Germany)
- KDD-2003, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (August 24-27, 2003, Washington DC, USA)
- IDA-2003, 5th International Symposium on Intelligent Data Analysis (August 28-30, 2003, Berlin, Germany)
- DaWaK 2003: 5th International Conference on Data Warehousing and Knowledge Discovery (September 3-5, 2003, Prague, Czech Repblic)
- PKDD-2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (September 22-26, 2003, Cavtat-Dubrovnik, Croatia)
- SAS M2003: 6th Annual Data Mining Technology Conference (October 13-14, 2003, Las Vegas, NV, USA)
- Data Warehousing & Data Mining for Energy Companies (October 16-17, 2003, Houston, TX, USA)
- CAMDA 2003: Critical Assessment of Microarray Data Analysis (November 12-14, 2003, Durham, NC, USA)
- ICDM-2003: 3rd IEEE International Conference on Data Mining (November 19 - 22, 2003, Melbourne, FL, USA)
- The Australasian Data Mining Workshop (December 8, 2003, Canberra, Australia, http://datamining.csiro.au/adm03/)

These 15 conferences do not include various artificial intelligence (AI), statistics and database conferences (and their workshops) that also solicited and accepted data mining related papers, such as IJCAI, ICML, ICTAI, COMPSTAT, AI & Statistics, SIGMOD, VLDB, ICDE, and CIKM.

Among various data mining conferences, KDD and ICDM are arguably (or unarguably) the two premier ones in the field. ICDM was established in 2000, sponsored by the IEEE Computer Society, and had its first annual meeting in 2001. Figure 1 shows the number of paper submissions to each KDD and ICDM conference.

Topics of interest from the ICDM 2003 call for papers [http://www.cs.uvm.edu/~xwu/icdm-03.shtml] are listed here:

1. Foundations of data mining
2. Data mining and machine learning algorithms and methods in traditional areas (such as classification, regression, clustering, probabilistic modeling, and association analysis), and in new areas
3. Mining text and semi-structured data, and mining temporal, spatial and multimedia data
4. Data and knowledge representation for data mining
5. Complexity, efficiency, and scalability issues in data mining

[1] Xindong Wu is with the Department of Computer Science, University of Vermont Burlington, VT 05405, USA (e-mail: xwu@cs.uvm.edu).

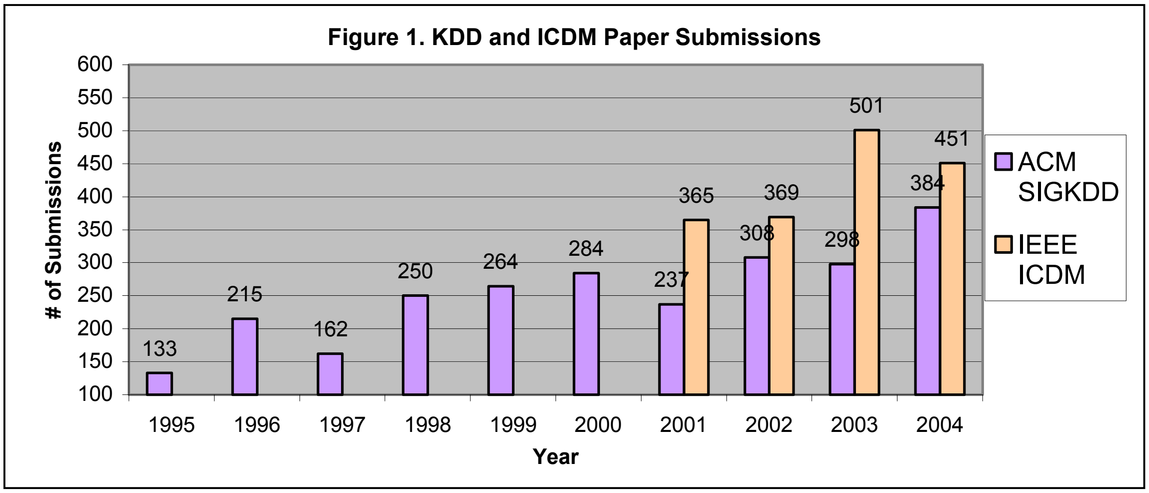

Figure 1. KDD and ICDM Paper Submissions

6. Data pre-processing, data reduction, feature selection and feature transformation
7. Post-processing of data mining results
8. Statistics and probability in large-scale data mining
9. Soft computing (including neural networks, fuzzy logic, evolutionary computation, and rough sets) and uncertainty management for data mining
10. Integration of data warehousing, OLAP and data mining
11. Human-machine interaction and visualization in data mining, and visual data mining
12. High performance and distributed data mining
13. Pattern recognition and scientific discovery
14. Quality assessment and interestingness metrics of data mining results
15. Process-centric data mining and models of data mining process
16. Security, privacy and social impact of data mining
17. Data mining applications in electronic commerce, bioinformatics, computer security, Web intelligence, intelligent learning database systems, finance, marketing, healthcare, telecommunications, and other fields

Clearly, some of the above topics are of interest from the database and statistics perspectives [Chen, Han and Yu 1996; Elder and Pregibon 1996; Zhou 2003]. Since the database perspective [Chen, Han and Yu 1996] and statistical perspective [Elder and Pregibon 1996] have been discussed and reviewed in detail in the literature, this paper concentrates on an AI perspective. We list the best papers selected from ICDM '01, '02, and '03 in Section 2, and discuss common topics in data mining and AI in Section 3.

II. Best Papers Selected from ICDM 2001, 2002, and 2003

Below are the best papers selected from ICDM 2001, 2002 and 2003, which have been expanded and revised for publication in Knowledge and Information Systems (http://www.cs.uvm.edu/~kais/), a peer-reviewed archival journal published by Springer-Verlag. The reference number before each paper, such as S336, M557 and R281, is the original submission number to each year's ICDM conference. We will see in Section III.A that these papers are all relevant to machine learning topics in AI.

ICDM 2001:

1. [S336] Discovering Similar Patterns for Characterising Time Series in a Medical Domain, by Fernando Alonso, Juan P. Caraça-Valente, Loïc Martínez, and Cesar Montes
2. [S409] Preprocessing Opportunities in Optimal Numerical Range Partitioning, by Tapio Elomaa and Juho Rousu
3. [S430] Using Artitificial Anomalies to Detect Known and Unknown Network Intrusions, by Wei Fan, Matthew Miller, Salvatore J. Stolfo, and Wenke Lee
4. [S457] Meta-Patterns: Revealing Hidden Periodic Patterns, by Wei Wang, Jiong Yang, and Philip Yu
5. [S516] Closing the Loop: an Agenda- and Justification-Based Framework for Selecting the Next Discovery Task to Perform, by Gary R. Livingston, John M. Rosenberg, and Bruce G. Buchanan

ICDM 2002:

1. [M557] Convex Hull Ensemble Machine, by Yongdai Kim
2. [M572] Phrase-based Document Similarity Based on an Index Graph Model, by Khaled Hammouda and Mohamed Kamel
3. [M632] High Performance Data Mining Using the Nearest Neighbor Join, by Christian Bohm and Florian Krebs
4. [M741] Efficient Discovery of Common Substructures in Macromolecules, by Srinivasan Parthasarathy and Matt Coatney
5. [M782] On the Mining of Substitution Rules for Statistically Dependent Items, by Wei-Guang Teng, Ming-Jyh Hsieh, and Ming-Syan Chen

ICDM 2003:

1. [R281] Clustering of Streaming Time Series is Meaningless: Implications for Previous and Future Research, by Jessica Lin, Eamonn Keogh, and Wagner Truppel
2. [R405] A High-Performance Distributed Algorithm for Mining Association Rules, by Ran Wolff, Assaf Schuster, and Dan Trock
3. [R493] TSP: Mining Top-K Closed Sequential Patterns, by Petre Tzvetkov, Xifeng Yan, and Jiawei Han
4. [R528] ExAMiner: Optimized Level-wise Frequent Pattern Mining with Monotone Constraints, by Francesco Bonchi, Fosca Giannotti, Alessio Mazzanti, and Dino Pedreschi
5. [R565] Reliable Detection of Episodes in Event Sequences, by Robert Gwadera, Mikhail Atallah, and Wojciech Szpankowski
6. [R620] On the Privacy Preserving Properties of Random Data Perturbation Techniques, by Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar

III. Common Topics in Data Mining and AI

A. Data Mining Papers on Machine Learning Topics

Machine learning in AI is the most relevant area to data mining, from the AI perspective. ICML 2003 [http://www.hpl.hp.com/conferences/icml03/] especially invited paper submissions on the following topics:

1. Applications of machine learning, particularly:
    a. exploratory research that describes novel learning tasks;
    b. applications that require non-standard techniques or shed light on limitations of existing learning techniques; and
    c. work that investigates the effect of the developers' decisions about problem formulation, representation or data quality on the learning process.
2. Analysis of learning algorithms that demonstrate generalization ability and also lead to better understanding of the computational complexity of learning.
3. The role of learning in spatial reasoning, motor control, and more generally in the performance of intelligent autonomous agents.
4. The discovery of scientific laws and taxonomies, and the induction of structured models from data.
5. Computational models of human learning.
6. Novel formulations of and insights into data clustering.
7. Learning from non-static data sources: incremental induction, on-line learning and learning from data streams.

Apart from Topic 5, all other topics above are relevant in significant ways to the topics of the 2003 IEEE International Conference on Data Mining listed in Section 1. Topic 2 is relevant to topics 2 and 5 in Section 1, Topic 3 overlaps with topics 3 and 1 in Section 1, and Topic 1 above and topic 17 in Section 1 both deal with applications. In practice, it is rather difficult to clearly distinguish a data mining application from a machine learning application, as long as an induction/learning task in involved. In fact, data mining and machine learning share the emphases on efficiency, effectiveness, and validity [Zhou 2003].

Meanwhile, every best paper from ICDM 2001, 2002 and 2003 in Section 2 can fit in the above ICML 2003 topics. With the exception of data pre-processing and post-processing, which might not involve any particular mining task, a data mining paper can generally find its relevance to a machine learning conference.

B. Three Fundamental AI Techniques in Data Mining

AI is a broader area than machine learning. AI systems are knowledge processing systems. Knowledge representation, knowledge acquisition, and inference including search and control, are three fundamental techniques in AI.

- ❖ **Knowledge representation**. Data mining seeks to discover interesting patterns from large volumes of data. These patterns can take various forms, such as association rules, classification rules, and decision trees, and therefore, knowledge representation (Topic 4 of ICDM 2003 in Section 1) becomes an issue of interest in data mining.
- ❖ **Knowledge acquisition**. The discovery process shares various algorithms and methods (Topics 2 and 6) with machine learning for the same purpose of knowledge acquisition from data [Wu 1995] or learning from examples.
- ❖ **Knowledge inference**. The patterns discovered from data need to be verified in various applications (Topics 7 and 17) and so deduction of mining results is an essential technique in data mining applications.

Therefore, knowledge representation, knowledge acquisition and knowledge inference, the three fundamental techniques in AI are all relevant to data mining.

Meanwhile, data mining was explicitly listed in the IJCAI 2003 call for papers [http://www.ijcai-03.org/1024/index.html] as an area keyword.

C. Key Methods Shared in AI and Data Mining

AI research is concerned with the principles and design of rational agents [Russell and Norvig 2003], and data mining systems can be good examples of such rational agents. Most AI research areas (such as reasoning, planning, natural language processing, game playing and robotics) have concentrated on the development of symbolic and heuristic methods to solve complex problems efficiently. These methods have also found extensive use in data mining.

- **Symbolic computation**. Many data mining algorithms deal with symbolic values. As a matter of fact, since a large number of data mining algorithms were developed to primarily deal with symbolic values, discretization of continuous attributes has been a popular and important topic in data mining for many years, so that those algorithms can be extended to handle both symbolic and real-valued attributes.
- **Heuristic search**. As in AI, many data mining problems are NP-hard, such as constructing the best decision tree from a given data set, and clustering a given number of data objects into an optimal number of groups. Therefore, heuristic search, divide and conquer, and knowledge acquisition from multiple sources [Zhang, Zhang and Wu 2004] have been common techniques in both data mining and machine learning.

For example, Ross Quinlan's information gain and gain ratio methods for decision tree construction, which uses a greedy search with divide and conquer, is introduced in both [Russell and Norvig 2003] and [Han and Kamber 2000], which are probably the most popular textbooks in AI and data mining respectively. Decision tree construction can make use of both symbolic and real-valued attributes.

Neural networks and evolutional algorithms (including genetic algorithms) are also covered in various AI and data mining references.

IV. Conclusion

Knowledge discovery from large volumes of data is a research frontier for both data mining and AI, and has seen sustained research in recent years. From the analysis of their common topics, this sustained research also acts as a link between the two fields, thus offering a dual benefit. First, because data mining is finding wide application in many fields, AI research obviously stands to gain from this greater exposure. Second, AI techniques can further augment the ability of existing data mining systems to represent, acquire, and process various types of knowledge and patterns that can be integrated into many large, advanced applications, such as computational biology, Web mining, and fraud detection.

Acknowledgements

The author would like to express his appreciation for the anonymous feedback from the Editorial Board on an earlier version of this paper, which has helped improve the paper.

References

[1] Ming-Syan Chen, Jiawei Han, and Philip Yu, Data Mining: An Overview from a Database Perspective, *IEEE Transactions on Knowledge and Data Engineering*, **8** (1996), 6: 866-883.

[2] John Elder IV and Daryl Pregibon, A Statistical Perspective on Knowledge Discovery in Databases, in *Advances in Knowledge Discovery and Data Mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.), AAAI Press, 1996, 83-113.

[3] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.

[4] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach, Second Edition*, Prentice-Hall, 2003.

[5] X. Wu, *Knowledge Acquisition from Databases*, Ablex Publishing Corp., U.S.A., 1995.

[6] S Zhang, C Zhang, and X Wu, *Knowledge Discovery in Multiple Databases*, Springer-Verlag, 2004.

[7] Zhi-Hua Zhou, Three Perspectives of Data Mining, *Artificial Intelligence*, **143**(2003), 1: 139-146.