Thoughts on Foundations of Data Mining

Zhengxin Chen

Data Mining Research Lab

College of Information Science and Technology

University of Nebraska at Omaha, USA

Data mining research lab at University of Nebraska at Omaha

- College of Information Science and Technology: "Information science and technology is a broad and expanding field of basic and applied science that encompasses parts of many disciplines that relate to information processing and management. Information science and technology addresses the information itself, the manipulation and management of the information, and the application of the information to management and decision-making processes. The field encompasses all aspects of the generation, transmission, storage, retrieval, analysis, presentation, and use of information."
- Departments (Computer science, MIS) and programs (bioinformatics, information assurance, …)
- Data mining research lab: Yong Shi (director), Zhengxin Chen (co-director) and other participants (http://dm.ist.unomaha.edu/)
- Some recent research topics:
  - Multiple-Criteria Linear/Quadratic Programming
  - Text mining for bioinformatics
  - Industry/business applications
  - Etc.

Abstract of this talk

As noted in recent IEEE Foundations of Data Mining (FDM) workshops, data Mining has been developed, though vigorously, under rather ad hoc and vague concepts. There is a need to explore various fundamental issues of data mining.

In this talk, we discuss the reasons of studying FDM, aspects need to be addressed, approaches to FDM, as well as related issues. Materials presented in this talk are based on publications by researchers working in this area, as well as the speaker’s personal comments and observations on the recent development of FDM.

Objectives of this talk

- "Advertisement": hopefully it will draw more attentions from both researchers and practitioners (Canadians are among most active researchers in FDM)
- Explain motivation behind FDM
- Encourage more people interested in FDM
- Show basic ideas behind different approaches; trying to show the *variety* of theories and to be *representative*…
- But: **Not** a *comprehensive* coverage
- Warning: Due to short time of preparation, this talk is not well-organized!
- Also no technical details will be given in this talk

Outline (sample issues to be addressed in this talk)

- What is foundations of data mining (FDM)?
- Why should we study foundation of data mining?
- What is the criteria for FDM?
- What is the current status of this field? (What are the current research topics, who are involved, etc.)
- What we can do about this?

What is FDM (foundations of data mining)?

- **What is FDM (foundations of FDM)? Foundation is NOT introduction of DM! (compare with: Foundation of mathematics)**
- **It is NOT theoretical background needed for studying data mining.**
- Retrospective: bottom-up examination; generalization and abstraction
- Second order of data mining? "Mining" of mining methods/approaches
- Note: Different researchers working in FDM may have somewhat different opinions about FDM

Why study foundations of data mining?

- Question: There are endless methods useful for data mining. So why should not spend more time in learning data mining methods themselves?
- A: ICDM Workshop Foundations of data mining + Personal observation:
- Data Mining has been developed, though vigorously, under rather ad hoc and vague concepts.
  - Personal observation 1: Popular data mining tasks include classification, clustering, association rule mining, outlier analysis, sequence analysis, time series analysis, etc. New DM tasks may emerge in the future.
  - Observation 2: Each task has developed its own methods, in a somewhat ad-hoc manner.
  - Observation 3: Very different methods are used in each data mining task. For example, association rule mining is quite independent to cluster analysis.
  - We can list more observations…

Why to study Foundations of Data Mining (Cont.)?

- For further development of DM, a close examination on its foundations seems necessary. [Although not necessarily everybody agrees on this.]
- There is a need to explore various fundamental issues of data mining.
- The study of foundations of data mining is in its infancy, and there are probably more questions than answers. (Mannila 2000)

Criteria for a good theory for DM (wishlist)

- Mannila (2000): A good theory for data mining should consider
  - the process of data mining
  - Have a probabilitsitc nature
  - Be able to descibe different data mining tasks
  - Be able to allow for the presense of background knowledge
  - Etc

More recent discussion on Criteria for FDM

Need to find a theoretical framework with a common of language Be simple and easy to apply (Yao).

1. Provide useful results that we could apply to the development of data mining algorithmss and methods.
2. Be able to model typical data mining tasks:

   Clustering, Classification, association, etc.

4. Be able to discuss the probability nature of discovered patterns and models.
5. Be able to talk about data and inductive generalizations of data.
6. Be able to accept different forms of data (relational, sequential etc..)
7. Recognize that data is an interactive and iterative process, where comprehending discovered results is important

ICDM FDM Workshop info

- ICDM2004 WORKSHOP Foundations of Data Mining (Brighton, UK, November 01 - 04, 2004) http://www.cs.sjsu.edu/faculty/tylin/icdm04_workshop.html
- ICDM2005 WORKSHOP Foundations of Semantic Oriented Data and Web Mining (Houston, Texas, USA, November 27, 2005) http://www.cs.sjsu.edu/faculty/tylin/ICDM05/

Topics of interest for FDM

2004 ICDM FDM Workshop (with my comments)

- 1. Theory of Data Mining and Discovery
  - Develop theory (or theories) for data mining (and/or for discovery) as a whole
  - Develop theories for each data mining tasks
  - Data mining vs. discovery
- 2. Similarity and Dissimilarity of Learning and Discovery
  - Concerned with nature of data mining; need to investigate two underlying concepts
- 3. Logical Foundation of Data Mining
  - Logic can be used as the starting point; logic can unify various tasks of DM and contribute to formalization of individual DM task
- 4. Modeling of Data Mining
  - to model the entire activity of DM as well as individual tasks
- (To be continued…)

Topics of interest for FDM (cont.) 2004 ICDM FDM Workshop (with my comments)

- 5. Theory of Patterns (Patterns could be data, analytic functions, Turing machines)
  - This is trying to develop theory (or theories) on the shared properties of mined results.
- 6. Sampling and Complexity Reduction
  - Note this is not just concerned with how to do the DM, but also related how to view the nature of DM (for some people, the result of DM is just a condensed or simplified representation of the original data)
- (To be continued…)

Topics of interest for FDM (cont.)

2004 ICDM FDM Workshop (with my comments)

- 7. Uncertainty in Data Mining and Discovery
  - Uncertainty in DM can have several different meanings. For example, since soft computing provides various useful techniques (including fuzzy set theory, rough set theory, etc.) to deal with uncertainty of data for data analysis, they can be used as the starting point for data mining theory.
  - Different perspective: Mining the data by removing the uncertainty so that the nugget is revealed. (Discussion of this aspect in my book: *Data Mining and Uncertain Reasoning: An Integrated Approach*, Wiley, New York, 2001.)

Topics of *2005* workshop

Foundations of ***Semantic Oriented Data and Web*** Mining

- 1. Theory of ***High Dimensional*** Data Mining and Discovery
- 2. Similarity and Dissimilarity of Learning and Discovery
- 3. Logical Foundation of Data Mining and Text Mining
- 4. Modeling of Data Mining and ***Text Mining***
- 5. Theory of Patterns (Patterns could be data, analytic functions, Turing machines)
- 6. Sampling and Complexity Reduction
- 7. Uncertainty in Data Mining and Discovery
- 8. ***Theory of Web Intelligence***
- 9. ***Foundation of Unstructured Data Clustering***
- 10. Modeling of Data Mining ***in Grid Computing Environments***
- 11. Modeling of ***Data Stream Mining***
- 13. Modeling of Data Mining in ***Bioinformatics***
- 14. Other New and Novel Approaches

Aspects need to be explored in FDM: A personal perspective

- General aspects:
  - Nature of data mining/KDD
- DM task aspects:
  - Logical/mathematical foundations for individual DM tasks
  - Common (L/M) foundations for different DM tasks

    (Note: the above two aspects are closely related to each other)
  - Difference and commonality of DM tasks
  - Interplay/integration of DM tasks
- Algorithm aspects:
  - Common features of different DM algorithms
- Related issues (such as relationship with uncertainty) and infrastructure of DM

Establish the *logical* foundation by exploring the nature of DM/KDD

- Knowledge discovery as translation (S. Ohsuga, in T. Y. Lin et al. eds., Foundations of Data Mining and Knowledge Discovery, pp. 3-20, 2005):
- Knowledge discovery can be viewed as a translation from non-symbolic data to symbolic (as discussed in AI literature) representation (knowledge patterns).
- The reason to have this translation is that there is a large gap between symbolic and non-symbolic representations: While characteristic of symbolic representation is to eliminate quantitative measure and also to inhibit mutual dependency between elements, non-symbolic processing has opposite characteristics.

Knowledge discovery as translation (cont.)

- A quantitative measure is introduced in the syntax of predicate to measure the distance between symbolic and non-symbolic representations quantitatively.
- Example of transition matrix for logical expression: $F \wedge$ [Fà G] $\Rightarrow G$ can be written as an equivalent math form, **P**g = **P** f $\times$ T, if we can find a transition matrix $T = |T_{kj}|$ satisfies certain condition can be found.
- Translation can be done in approaches other than logic, such as using an artificial neural network (ANN). Although this is to represent an ANN by means of special intuitive logic, the approach loses some advantages of classic logic such as expandability and completeness of inference.

Mathematical foundation of a *specific* DM task: Association rule mining

- T. Y. Lin: Mathematical foundation of association rules– Mining association by solving integral linear inequalities, in the same book containing the Ohsuga paper, pp. 21-42
- Basic idea: Data is a table of "symbols" (tokens?) and a pattern is any algebraic/logic expressions derived from this table that have high supports. A pattern (containing generalized associations) of a relational table can be found by solving a finite set of linear inequalities within a polynomial time of the table size.
- Some other interesting points:
  - Patterns are properties of the isomorphic class, not individual relations
  - Un-interpreted attributes (attributes) are partitions; they can be enumerated.

Approaches to FDM

- Frameworks: Focusing on nature of data mining, common features and/or overall process of data mining activities
- More specific approaches: based on specific logical or mathematical methods
- Examples follow

(Earlier) frameworks of data mining

(Mannila 2000)

- Probabilistic approach: Data mining is viewed as the task of finding the underlying joint distribution of the variables in the data. This approach is closely related to the reductionist approach of viewing data mining as statistics.
- Data compression approach: Data mining is aimed to compress the data set by finding some structure for it. E.g., assocaition rules can be viewed as ways of providing compression of parts of the data, a decision tree is considered as a compression method for the target attribute, and a clustering is a way of compressing the data set.

(Earlier) frameworks of data mining

(Mannila 2000) (Cont.)

- Microeconomic view: Data mining is about finding actionalble patterns; the only interest is in patterns thant can somehow be used to increase utility.
- Inductive databases: Query concept should be extended to data mining; there is no such thing as discovery, it is all in the power of the query language. The term inductive databases refers to a normal database plus the set of all sentences from a specified class of sentences that are true in the data (inductive DB = data + theory of data).
- Note: None of the existing theories satisfy all the requirements of FDM theory mentioned earlier

Frameworks:

more recent progress

- Data mining foundation as a study of three related dimensions (Chen 2002):
  - The philosophical dimension deals with the nature and scope of data mining.
  - The technical dimension covers data mining methods and techniques.
  - The social dimension concern the social impact and consequences of data mining.

A Multi-level Framework for Modeling Data Mining (Yao et al.)

- The kernel focuses on the study of knowledge without reference to data mining algorithms.
- The technique levels focus on data mining algorithms without reference to particular application.
- The application levels focus on the utility of discovered knowledge with respect to particular domains of applications.

More specific approaches: Logical foundation for FDM

- Logical foundation of data mining based on Bacchus' probability logic (Xie and Raghavan 2002):
  - Precise definition of intuitive notions, such as “pattern”, “previously unknown knowledge,” “potentially useful knowledge,” etc.
  - A logic induction operator is defined for discovering “previously unknown and potentially useful knowledge".

More specific approaches: Granular computing (GrC) for FDM

- Granular computing as a basis for data mining (Lin 2002, Tsumoto 2002, and Yao 2001 – JT Yao's description below)
  - A concept consists of two parts, the intension and extension of the concept.
  - The intension of a concept consists of properties objects.
  - The extension of a concept is the set of instances.
  - A rule can be expressed in the form, φ=>ψ where φ and ψ are intensions of two concepts.
  - Rules are interpreted using extensions of the two concepts.

Granular Computing and Rough Set Theory

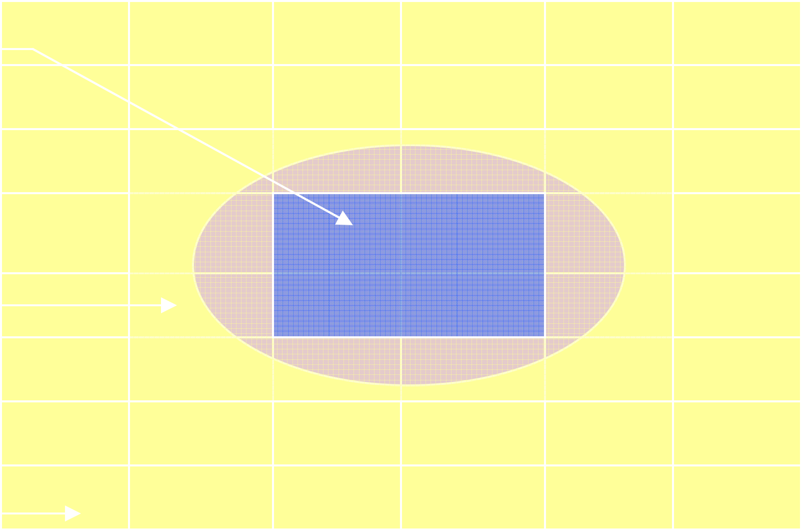
- "Granular Computing (GrC) is an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules in problem solving."
- GrC is a superset of rough set theory.
- Rough set theory is one of the main directions of GrC.

Rough set theory and association rule mining

- In association rule mining, we are interested in finding all one-way associations whose support and confidence are above specific threshold.
- Explanation Oriented Data Mining – ideas behind using rough set theory for association rule mining
  - Finding associations is only the first step. We still need to understand their meanings and implications.
  - Instead of finding phenomenon only, the reason for the existence of phenomenon should be discovered.
  - Measures like support and confidence are NOT enough. Explanation for "Why?" requires other data source or domain information

A rough introduction on Rough Set

- 1. $R_*(X)$
- 2. $POS_R(X)=R^*(X)$
- 3. $R^*(X)$
- 4. $NEG_R(X)=S-R^*(X)$



- The measure of approximation can be achieved by the distance between $R_*(X)$ and $R^*(X)$, namely Lower Approximation and Upper Approximation respectively.

Foundation of classification from a GrC perspective

- Classification Rule Induction Method
- Information table (this term is from rough set theory) is used to represent a set of objects
- Use granule centered strategies instead of attribute centered strategy
- A Family of Granules forms a partition/ granulated view of the universe.
- View Point of GrC on DM:
  Characterizing individual granules and finding relationships between the granules
  - Represent the relationship in if-then rules
- Different granulated views → Different levels of rules

Classification Rules

Object	height	hair	eyes	class
o1	short	blond	blue	+
o2	short	blond	brown	-
o3	tall	red	blue	+
o4	tall	dark	blue	-
o5	tall	dark	blue	-
o6	tall	blond	blue	+
o7	tall	dark	brown	-
o8	short	blond	brown	-

- Example:
  - If we are interested in (Class, +) concept, the rows involved are {o1,o3,o6}. Using granules produced based on Height, Hair, and Eyes, we can get these rules:
    - IF (Hair, red) THEN (Class, +)
    - IF (Hair, blond) $\wedge$ (Eyes, blue) THEN (Class,+)
    - IF (Hair, dark) THEN $\neg$(Class, +)
    - IF (Hair,red) $\vee$ (Hair, blond) $\wedge$ (Eyes, blue) THEN (Class,+)

Data mining as generalization: Another GrC perspective

- E. Menasalvas and A. Wasilewska, Data preprocessing and data mining as generalization process, Proc. IEEE FDM 2005, pp. 133-137.
- A generalization model is defined as a system (a 4-tuple) consisting of universe, generalization states, generalization operators and generalization relation.
- A knowledge generalization system is defined based on the notion of information system in rough set theory (in which a set is described in terms of its upper and lower approximation).
- A granular view of data mining is then formalized in terms of knowledge generalization system.

MRDM for FDM

- Multi-Relational Data Mining (MRDM) is the multi-disciplinary field dealing with knowledge discovery from relational databases consisting of multiple tables. http://www-ai.ijs.si/SasoDzeroski/MRDM2004/
- Inductive inference: it generalizes from individual instances/observations/objects in the presence of background knowledge, finding regularities /hypothesis about yet unseen instances.
- **Why study MRDM in the context of FDM? Inductive logic programming (ILP) that underlying MRDM can serve as a common ground for different DM tasks, thus contributing to FDM (never noted before).**
- Example: **WARMR** algorithm for association rule mining is based on Apriori algorithm for mining association rules in multiple relations. Used Datalog for representing data and background info.

MRDM for FDM (cont.)

- MRDM is studied by researchers outside of FDM rcamp(and may not be willing to label their work as for FDM), and no attention has been paid by FDM researchers. But MRDM is relevant to FDM.
- Problems (or at least potential problems) of MRDM for FDM: efficiency, scalability, etc.

Other candidate theories for FDM

- Genetic algorithms (GA): GA for classification, clustering, association rule and other types of mining
- Artificial neural networks (ANN): ANN for classification, association…
- Note: Although there has been significant amount of research work on using GA, ANN, etc. for data mining, focus has been on developing efficient algorithms. In order to use them for FDM, additional aspects should be examined.

Additional aspects and related issues

- “Borders” of DM tasks could be blurred…
  - Is this classification or association?
- Interplay of DM tasks
  - E.g. Classification + association
- Parameter-free data mining:
  - Danger of Parameter-laden algorithms: Incorrect settings may cause an algoirthm to fail in finding true patterns, the algorithm may report spurious patterns that do not really exist or overestimate the significance of reported patterns
  - Compression-based DM paradigem has been proposed to allow Parameter-free or parameter-light solutions to various DM tasks including clustering, classification and anomaly detection(E. Keogh, S. Lonardi and C. A. Ratanamahatana, Towards parameter-free data mining, KDD 2004).

Side note:
Data mining as An AI powered tool

http://www.aaai.org/AITopics/html/mining.html

- Data mining is an AI powered tool that can discover useful information within a database that can then be used to improve actions.
- **Data mining: "The Rebirth Of Artificial Intelligence"**
- -- Can these perspectives be used as starting point of FDM?

A non-conclusive conclusion

- Advantages of studying FDM: Better understand the nature of DM, more systematic way of conducing DM activities
- Shared goal (with somewhat different emphasis), diverse approaches, different opinions
- Will we succeed – or, What would be the possible outcome?
  - Can we achieve this ultimate goal: A “final” theory (or theories) of data mining?
  - … But at least along the way there will be many interesting *theoretical* results *useful* in practice

General references for FDM

- ICDM Foundation of Data Mining Workshop proceedings (Proceedings 2002, 2003, 2004, and 2005) Free 2004 online proceedings is available at: http://www.cs.sjsu.edu/faculty/tylin/icdm04_workshop.html (Previous years proceedings available as hard copies only)
- Lin, T.Y., Ohsuga, S., Liau, C.-J., Hu, X. (Eds.), *Foundations and Novel Approaches in Data Mining*, Springer, 2006 (to appear)
- Lin, T.Y., Ohsuga, S., Liau, C.-J., Hu, X.: Tsumoto, S. (Eds.), *Foundations of Data Mining and Knowledge Discovery*, Springer, Nov. 2005.
- Chu, Wesley; Lin, Tsau Young (Eds.), *Foundations and Advances in Data Mining, Springer, Nov. 2005.*

Basic readings (and background info)

- J. T. Yao, Panel discussion of FDM at RSCTC2004 http://www2.cs.uregina.ca/~jtyao/Conf/Penel_FDM.ppt
- Heikki Mannila, Theoretical Frameworks for Data Mining, 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD); also in SIGKDD Explore, 1(2), pp. 30-32, 2000. http://portal.acm.org/citation.cfm?id=846191&coll=portal&dl=ACM&CFID=55698756&CFTOKEN=95581561
- Z. Chen, The three dimensions of data mining foundation, Proc. ICDM FDM Workshop, 2002.
- Y. Yao, N. Zhong and Y. Zhao, A conceptual framework of data mining, Foundations on Data Mining (to appear).

Granular computing (GrC) for FDM

- Yao, Y.Y., Perspectives of Granular Computing , *Proc. 2005 IEEE Conference on Granular Computing.*
- J.T. Yao, Y.Y. Yao, and Y. Zhao, "Foundations of Classification", in T.Y. Lin, S. Ohsuga, C.J. Liau, and X. Hu (Eds), *Foundations and Novel Approaches in Data Mining*, Springer-Verlag, 2005, pp75-97.

MRDM references

- S. Dzeroski, Multi-Relational Data Mining: An Introduction, SIGKDD Exploration, 5(1), 2003.
- P. Domingos, Prospects and Challenges for Multi-Relational Data Mining, SIGKDD Exploration, 5(1), 2003.
- (These papers and related papers can be downloaded at http://www.acm.org/sigs/sigkdd/explorations/issue.php?issue=current)

Reference on Foundation of association rules

- Mohammed J. Zaki and Mitsunori Ogihara, Theoretical Foundations of Association Rules, in Proceedings of 3 rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'98), 1998. (Foundation of association based on formal concept analysis)

Other references

- Lin, T.Y. Issues in modeling for data mining, COMPSAC’02, 1152-1157, 2002.
- Tsumoto, S.,T.Y Lin, J.F. Peters. Foundations of Data Mining via Granular and Rough Computing. COMPSAC’02, 1123-1124, 2002
- Yao, Y.Y. Modeling data mining with granular computing, COMPSAC’01, 638-643, 2001.
- Yao, Y.Y., A step towards the foundations of data mining, SPIE Vol. 5098, 254-263, 2003.

Acknowledgements

- This talk has incorporated materials from various references, plus personal perspective
- Thanks to CSCI 8390 (Special topics– Logical and mathematical foundation of data mining) students at Univ. Neb. At Omaha, Fall 2005 (to rejuvenate my interest in FDM):
  - E. Beyenne
  - D. Djajakusli
  - A. Kaplun
  - M. Liu
  - R. Zhang

Thank you!

- Questions?