KDD-99 Panel Report: Data Mining into Vertical Solutions

Ronny Kohavi
Blue Martini, Inc.
ronnyk@bluemartini.com

Mehran Sahami
E.Piphany, Inc.
Sahami@epiphany.com

Panelists:
Steve Belcher, Jim Bozik, Rob Gerritsen, Ken Ono, and Dorian Pyle

ABSTRACT

At KDD-99, the panel on Integrating Data Mining into Vertical Solutions addressed a series of questions regarding future trends in industrial applications. Panelists were chosen to represent different viewpoints from a variety of industry segments, including data providers (Jim Bozik), horizontal and vertical tool providers (Ken Ono and Steve Belcher respectively), and data mining consultants (Rob Gerritsen and Dorian Pyle). Questions presented to the panelists included whether data mining companies should sell solutions or tools, who are the users of data mining, will data mining functionality be integrated into databases, do models need to be interpretable, what is the future of horizontal and vertical tool providers, and will industry-standard APIs be adopted?

Keywords

Panel report, Inegration of Data Mining, Vertical Markets.

1. INTRODUCTION

The members of the panel were selected by the chairs to encourage diverse and conflicting viewpoints on a wide variety of topics related to the integration of data mining into vertical solutions. The chosen panelists were Steve Belcher (Unica Technologies Inc.), Jim Bozik (Acxiom Corporation), Rob Gerritsen (Exclusive Ore, Inc.), Ken Ono (ANGOSS Software Corporation), and Dorian Pyle (Data Miners). A more detailed description of the panelists can be found at the end of this article.

In order to focus on more controversial issues, each panelist was asked for his opinion on eight questions. The panel chairs chose six questions that we felt covered a broad set of interesting topics whose answers differed by at least two panelists.

Due to the linear nature of the panel presentations, each panelist was asked to respond to two different questions. Each question was answered by two consecutive panelists (except for first and last) whose responses differed significantly.

The structure of this article mirrors the structure of the panel, with the addition of a synthesis section by the panel chairs after each question. The panel chairs edited the text and slides from the authors. The synthesis includes the panel chairs' views on the topic.

2. The Questions

Question 1: Solutions versus Tools: what should companies sell?

Jim Bozik: As a prospective user of data mining software, we are not interested in purchasing tools. Software is the most important mechanism that helps us address the needs of our clients. However, the software itself is only important in that it helps us provide answers to our clients in a timely, cost effective, and perhaps visually appealing way, and that it satisfies needs that we could not address before. For the past year, we at Acxiom have been involved in evaluating software packages that will permit us to enhance our 'analyst toolkit.' We offer the following suggestions: (i) listen to the customer, then offer solutions using your tools that answer the customer's needs, and (ii) provide guidance to the user on appropriate configuration of the testing platform, or file sizes for test data.

From a user's perspective, purchasing data mining software is not that different than purchasing any consumer product – we have a need, and how do the software tools address that need?

Synthesis:

We believe that most companies should concentrate on creating solutions with appropriate domain terminology and guidance. However, there is room for a few tool vendors that can succeed in OEMing their technology (see Question 5).

Question 2: Who are the users of the data mining software? Business users or analysts?

Jim Bozik: We define Analysts as people that "think analytically." They need not be statisticians. We contend that we can train someone on software, or model building, but that it is much harder to train someone to think analytically. There is an art to analysis that does not simply come from the mechanics of analysis. Analysts must validate that the data matches the expected data, spot problems with data collection, create new variables, select variables and models, and make sure the models appeal to the clients.

Business users lack the time and practical experience needed to build useful models.

Dorian Pyle: In an ideal world, the software would do what the user wants, not what the user said they want. The answer to date has been to encapsulate particular domain knowledge with the power of data mining tools and give them a business face. Data mining, the core technology, is invisibly wrapped, thus isolating the business user from specialist technology (data mining). The users of data mining are those who needs answers to questions based on available data, including business managers, planners, plant managers, marketers, and investors.

Synthesis:

Most data mining software today requires significant expertise in knowledge discovery, modeling methods, databases, and statistics. The use of such tools requires an analyst. In vertical domains, it

is possible to make data mining more invisible [John 99] by incorporating additional infrastructure appropriate for specific applications. Such applications can then be used effectively by business users.

Question 3: Will data mining functionality be successfully integrated into databases?

Dorian Pyle: Data mining functionality will not be successfully integrated into databases for the following reasons:

1. All of the data to be accessed is not necessarily in a database. Multiple formats will require specifically tailored search and collation criteria. Technically, it would be an enormous challenge to determine what to incorporate.
2. The questions asked require multiple methods of inquiry. One size does not fit all.
3. Performance and concurrency. Data mining activity can overload any corporate data repository (for now).

Rob Gerritsen: Data mining functionality will be successfully integrated for reasons that fall into two categories: it's natural and it's inevitable. It's natural because data mining is data centric and can take advantage of a number of the features of DBMS. Integration is also inevitable, primarily because of marketplace factors.

1. In every sense a model is a compression or distillation of the data. The model itself is also data. It is pretty clear that the model itself, as data, belongs in the database. The most glaring missing feature in the current set of data mining products is model management, which can be provided by a database.
2. When a predictive model is used to predict new values it is generating new data. Like all other data, the predictions belong in a database.
3. With the model in the database, it is easier to monitor model performance over time and update the model as required.
4. Few data mining tools provide a security model, while all DBMSs contain extensive provisions for security.
5. DBMS contain optimizers that can provide faster data access. DBMSs could also include pre-computed values, thereby eliminating large numbers of accesses when models are induced.
6. There is good evidence that marketplace factors are at work right now to make this integration. Compaq, Informix, Microsoft, Oracle, and Sybase are integrating data mining into their DBMSs.
7. Like an operating system, a DBMS is a platform. Expanding the platform provides more value to customers and keeps competitors out.

Synthesis:

We believe that more data mining primitives will be incorporated into databases, but there is room for advanced algorithms, specialized analysis, and front-end user interfaces built outside the database. While core data mining algorithms will be added to DBMSs, knowledge discovery environments will be built outside the DBMSs. These will include task-oriented functionality, specialized transformations, exploratory data analysis, and visualization capabilities.

Question 4: Do Models Need to be Interpretable?

Rob Gerritsen: Models need to be interpretable for both model builders and for business users. The latter because using the knowledge that comes out of data mining can entail significant risks – "just trust me" won't do it for most business executives. The former because models that are not interpretable can hide significant errors.

Models are abstractions of reality. It is very easy to make a mistake in a model. Neither the builders of the model nor the ones who are putting their business or career on the line by using the model can afford to treat a model as a "black box." Understanding the model is crucial for discovering data anomalies, detecting variables that "leak" the target value but won't be there when the model is deployed, and providing the needed trust in domains where it is needed (e.g., medicine). Moreover, in some situations, such as approving a loan, there are legal requirements to explain why someone may be denied a loan.

Steve Belcher: Foremost, models need to work. In order to measure this, models must be validated. Still, the interpretability of a model is subject to customer needs. If a customer is trying to build a model for lending, then interpretability may be a legal requirement. But this does not imply that models always need to be interpretable.

Synthesis:

In most cases, models need to be interpretable in order for business users to have confidence in them. Telling a customer that his loan was denied because he is on the wrong side of a hyper-plane is unthinkable and illegal without further explanations. Moreover, the interpretability of a model can allow for business users to gain new insights by inspecting them, providing new knowledge that can carry over to affect business actions outside the realm of target predictions. While there are some situations in which a well-validated "black box" model can be applicable (e.g., handwriting recognition), we believe that most business users are willing to sacrifice some amount of classification accuracy in order to have an interpretable model. Such understanding might also lead to better model development in the longer term.

Question 5: Is There a Future for Horizontal Data Mining Tool Providers?

Steve Belcher: There is a very limited future for horizontal data mining tool providers. Many vendors are currently consolidating. Data mining tool vendors are either being acquired by larger companies (such as the recent acquisition of Darwin from Thinking Machines by Oracle), or are moving into building vertical applications. Such vertical applications are often easier to use and are more specialized than the offerings of horizontal tool vendors. Furthermore, data mining models must be usable in business environments to solve real problems. As a result, such models must be integrated into other business applications, opening the door for data mining tools to be embedded into other systems.

Ken Ono: There is a strong future for horizontal data mining providers. The OEM business model is one of many approaches for data mining suppliers. A data mining component provider can deliver the tools and redistributable components so vertical solution providers need not concern themselves with the intricacies of data mining algorithms. By incorporating data mining into vertical applications, the complexities of data preparation, validation and the algorithm itself can be entirely hidden. When data mining becomes that easy-to-use, the market will widen and drive further sales of data mining components.

The analytic tools will continue to grow at a modest pace. It is vital that data mining component providers also deliver a leading workbench as many organizations have quantitative experts on staff and will want to leverage this asset. While an embedded automatic data mining process is infinitely better than not doing mining at all, an expert individual using a workbench will be better than an automated process.

In response to data mining being incorporated into the database, data mining vendors must leverage and enhance functionality of databases. Client side tools are still required for data exploration and discovery of new and interesting insights, a process that is distinct from server-based predictive modeling.

Synthesis:

The industry cannot sustain the number of horizontal data mining tool vendors that exist today. Basic data mining functionality will be available through databases, but more advanced features and client-side front-ends will be provided through several strong players who will develop a sustainable OEM model. The trend in the recent years shows that the market for horizontal data mining tools is limited as customers seek more integrated solutions.

Questions 6: Will Industry-standard APIs be adopted? Will They Help Horizontal Data Mining Companies?

Ken Ono: Standard APIs for data mining are already starting to emerge and will be adopted. The most significant emerging standard is OLE DB for Data Mining from Microsoft. The availability of OLE DB for DM will pave way for wide deployment of low risk predictive models. High-risk predictive models (such as “Should I give this person a loan?”) should probably remain a manually driven process left to experts. Another benefit of OLE DB for DM is that it creates infrastructure for deployment of models. Experts can create predictive models in their favorite tool and then use OLE DB for DM as a deployment vehicle.

Predictive Model Mark Up Language (PMML) is another emerging standard. It is an XML extension for describing the contents of a predictive model. PMML defines a way for a predictive model to be transferred between environments, persisted in a repository, searched and queried.

Such APIs will help data mining vendors by:

1. Reducing the cost of ownership of adopting and providing solutions that contain data mining.
2. Increasing the level of awareness about and the demand for data mining.
3. Increasing competition in the data mining space.

In conclusion, standards will help make data mining a widely deployed technology.

Synthesis:

While there are currently a few industry standard APIs for data mining emerging, it seems too early to tell how widely they will be adopted and deployed. The emergence of OLE DB for Data Mining from Microsoft certainly raises the awareness of data mining, and also helps to accentuate its importance. Still, this standard is still nascent, and it will likely take some time before it is adopted by a wide number of vendors.

The development of APIs for data mining is likely going to make it easier for tools providers to embed their technologies into other systems. Moreover, it will allow the models resulting from data mining to be more easily deployed. On the down side, however, the use of standard APIs will also make it easier for data mining tools from smaller horizontal vendors to be commoditized and displaced as larger companies make more significant in-roads into the data mining space. The big question is whether the young field of data mining is mature enough for standard APIs; early standardization may slow innovative development of new algorithms and techniques.

3. Summary

We summarized the processes of selecting the panelists, the process of selecting interesting questions, the responses from the panelists, and our own synthesis and viewpoints.

Data mining is an emerging field, trying to cross the chasm from a technology used by innovators into the mainstream [Kohavi 98, Agrawal 99]. Several efforts for data mining standards are being developed, most notably the Predictive Model Markup Language (PMML) [Cover 99] and Microsoft's OLE DB for data mining. Efforts to standardize the broader knowledge discovery process are also being developed [CRISP-DM 99].

There are no right or wrong answers to our questions, but different viewpoints and opinions create healthy discussions that will hopefully help the field cross the chasm. Alan Kay wrote that "The Best Way To Predict the Future is to Invent It." We are all working on creating the future, and we believe that data mining will be one of its important components.

4. Original Panel Presentation Slides

The original panel slides are available at:

http://robotics.Stanford.EDU/~ronnyk/kddPanel.zip

References

5. REFERENCES

[Agrawal 99] Rakesh Agrawal, Data Mining: Crossing the Chasm. KDD-99. http://www.almaden.ibm.com/cs/quest/papers/kdd99_chasm.ppt

[Cover 99] Robin Cover, Predictive Model Markup Language (PMML).

http://www.oasis-open.org/cover/pmml.html

[CRISP-DM 99] CRoss-Industry Standard Process for Data Mining. http://www.crisp-dm.org/

[John 99] George H John, Behind-the-Scenes Data Mining: A Report on the KDD-98 Panel, SIGKDD Explorations, June 1999, Volume 1, Issue 1, p. 6

[Kohavi 98] Kohavi Ron, Crossing the Chasm: From Academic Machine Learning to Commercial Data Mining. Invited talk at the International Conference on Machine Learning, 1998. http://robotics.Stanford.EDU/~ronnyk/ronnyk-bib.html

About the Panelists:

Dorian Pyle: Dorian is a consultant with Data Miners. He has mined and modeled data in many industries and areas over the last 25 years, including, Customer Relationship Management (CRM), Enterprise Resource planning (ERP), Web site, logistics, equity markets, health-care, insurance, investment, banking, telecomms, sales, marketing, fraud, simulation, criminal investigation, personnel profiling, expertise capture, industrial automation. Author of "Data Preparation for Data Mining" already published. Next book provisionally titled "Mining for Models" to be published about April 2000. Currently engaged in producing several training courses for national and international presentation in association SAS and other major players.

Jim Bozik: Acxiom Corp., Strategic Decision Services Division . Jim joined Acxiom in April 1997 from Signet Bank, where he was a Sr. Analyst in the Retail Marketing and Analysis Division. His work experience also includes analytical positions in the Business Research Division of Hallmark Cards and the Statistical Research Division of the U.S. Census Bureau. Jim has a B.A. in Mathematics and Computer Science from Defiance (OH) College and a M.A. in Statistics from Penn State University. His professional interests are in the use of statistical methods in applied settings, the proper use of graphs in communicating information, and quantitative literacy. At Acxiom, in addition to client responsibilities, Jim is leading the staff's Continuous Improvement effort, which includes researching analytical software tools.

Ken Ono: ANGOSS Software Corporation. Ken is VP of Technology at ANGOSS. Ken is the chief architect and head of development for the ANGOSS suite of data mining solutions, currently KnowledgeSEEKER, KnowledgeSTUDIO, KnowledgeExcelerator and KnowledgeAccess. Ken also focuses on OEM, embedding and other licensing transactions with partners involving ANGOSS data mining components.

Mehran Sahami: E.piphany, Inc. Mehran Sahami is a Systems Scientist at E.piphany leading their research and development effort in data mining. He is also the Senior Manager for E.piphany's Real-Time Products group. Prior to joining E.piphany, Dr. Sahami was involved in a number of machine learning and data mining research projects at Xerox PARC, SRI International, and Microsoft Research. He has also worked as a consultant on problems in text mining, clustering, and classification. He is a Visiting Lecturer at Stanford University, where he teaches classes on programming methodology, artificial intelligence, and the ethical implications of technology. He received his B.S., M.S., and Ph.D. in Computer Science all from Stanford University.

Rob Gerritsen: Exclusive Ore, Inc. Dr. Gerritsen has over 35 years experience in data processing, including more than 31 years in database management and data mining. Since founding Exclusive Ore Inc. in 1997, Dr. Gerritsen has focused on data mining consulting and technology, including research into effective integration of data mining technologies and RDBMS. His contributions span the spectrum from the application of AI to the theory of database design. In 1973 he was the first person to successfully apply the tools of artificial intelligence to database design. He was previously co-founder and Vice President of Technology at Two Crows Corporation, where he co-authored a hands-on study of more than 15 data mining products. Prior positions include President of Seed Software, Inc., and Associate Professor at The Wharton School. He co-designed and co-implemented the award winning client-server application, Leonardo, at the National Gallery of Art. He has a Ph.D. in System Science from Carnegie-Mellon University.

Ronny Kohavi: Blue Martini Software. Ronny Kohavi is the director of data mining at Blue Martini Software. Prior to joining Blue Martini, Kohavi managed the MineSet project, Silicon Graphics' award-winning product for data mining and visualization. He joined Silicon Graphics after getting a Ph.D. in Machine Learning from Stanford University, where he led the MLC++ project, the Machine Learning library in C++ now used in MineSet and for research at several Universities. Dr. Kohavi co-chaired KDD-99's industrial track with Jim Gray. He co-edited (with Dr. Provost) the special issue of the journal Machine Learning on Applications of Machine Learning. Kohavi and Provost are co-editors of the special issue of the Data Mining and Knowledge Discovery journal on E-Commerce and Data Mining (to appear in 2000).

Stephen D. Belcher: Unica Technologies Inc. Steve is a Consultant with Unica Technologies Inc., a leader in the marketing automation marketplace. Unica is known in data mining circles for its market-leading Model 1 and Pattern Recognition Workbench (PRW) products, and has recently released Impact!, the world's first predictive campaign management system. Stephen has worked in IT and data mining for over 16 years, in a wide variety of industries. He has taught at several colleges in both graduate and undergraduate programs, and wrote his doctoral dissertation on the application of neural networks in financial forecasting. Stephen is a member of the AAAI and the IEEE Computer Society.