

Unleashing Data Mining's Full Potential

By Rob Meyer, D.Sc. and Stephen Langdell, Ph.D. Numerical Algorithms Group

Data mining techniques hold great promise for competitive advantage. But beyond this promise, data mining is still a fledgling concept to most in the business world. Among the keys to a successful business application of data mining are a worthy problem or need, a significant source of appropriate data, and knowledge of the appropriate data mining techniques. Even meeting these requirements is not enough if the problem hasn't been posed properly for a decision or action. Here's an example:

A major manufacturer of industrial process control hardware has completed a successful implementation of a vendor's CRM software. Multiple databases of customer information, product information, and support call activity are now warehoused together. Various analytical tools allow the user to find out instantaneously how many "PID temperature controllers" a customer has bought by model, territory, time period, sales rep, ad nauseum. One can also see all of the trouble reports and the history of resolutions.

Now seat yourself in the customer support call center. You take a call from another customer who is having a problem getting his controller to work properly. Intuitively, you suspect that the problem has been solved before and you would like to give the customer the "high percentage" solution. The question is how? Traditional systems are good at doing queries but what do you query? Do you read through 200 trouble reports for that model? Expert systems might help but they are complex and expensive to build and can't keep up easily with current information.

This problem is essentially a classification problem for data mining. You've prepared data from call reports, cleaned it and built a model (e.g., a decision tree, logistic regression, linear regression, nearest neighbors or neural network model). The customer gives you basic data about the model and installation. You use the model to classify this new observation. It returns to you the situations most like the current one and the "fix" most likely to solve the customer's problem. You might even be able to do it without dragging him through an hour-long diagnostic process and save both of you both some time!

You could easily construct the example above for circumstances as diverse as an Internet Service Provider selling DSL services or a consumer product. Once the problem to be solved is posed properly, the power of data mining can be understood and enthusiasm for these techniques will follow. To some degree, time itself will cure some of the misperceptions about data mining in commercial applications because one successful practitioner gaining competitive advantage will compel others to follow suit.

However, there are other impediments to the growth of data mining that are unfortunately inherent in the ways in which data mining software has been crafted until recently.

To begin, confusion often reigns on the name, purpose, and capabilities of a particular data mining technique. For the novice, confusion is created simply by the wide range of data mining techniques that are potentially available. This problem can be addressed by adequate documentation, but several data mining software packages have not provided the type of documentation that quickly and easily guides novices to the most appropriate modeling methods possible. Even users who are well-versed in data mining often find that the learning curve for new software packages is too steep because of insufficient documentation.

Even if one has a handle on the range of data mining techniques to potentially employ, there are other hurdles to surmount. One might rightfully assume that functions called by the same name should give the same results on a given data set, no matter which software package is used. But that is often not the case and while there is much common ground in functionality between data mining systems, the exact implementation of the algorithm differs from product to product. For example, there is no industry standard for dealing with "ties" under classification functions. Consequently, although each classifier using a different method of dealing with ties is equally valid, the results of one type of classifier may differ from the results using other routines to handle data ties. Similarly, there has been a lack of clarity and consistency in the implementation of a particular decision tree method called CART (Classification and Regression Trees). The results produced can be very different depending upon on how the decision tree was implemented.

For more experienced users, the inflexibility of data mining software packages can soon become the predominant problem. Until the recent introduction of plug-in component systems that allow a broader choice of data mining algorithms, anyone seeking to add functionality to his or her data mining system typically had to buy and learn a new data mining software package to get the desired functionality. Often, the new data mining software would lack some of the functionality of the original software, the result being that extra time and effort would need to go into solving the problem. The alternative might be to wait for an update of the original software but a better approach for many might be to use data mining components which operate on a "plug-in" basis from firms such as The Numerical Algorithms Group (NAG), which sells individual functions individually that can then be wrapped as COM objects or with other interfaces for use with a user's existing system.

The existence of plug-in components has other practical effects that are expected to speed the development of data mining applications. When developers have access to statistical and machine-learning software that can be integrated into a broad range of applications, they are spared the weeks or months it takes to develop, debug, test and maintain new software. For commercial developers, this can mean new data mining product features and reduced time-to-market for a new product. For developers building enterprise applications, it can mean faster project results and more satisfied users.

Realizing the promise of data mining is intertwined with the creation of an "open" environment that permits the best data mining methods to be brought to bear on a particular business problem. If one can avoid being locked into one interface or one set of methods, one can bring the full power of data mining to solve the problem at hand.

Stephen Langdell is a member of the Data Analysis and Visualization Group at NAG Ltd. and Rob Meyer is President of the Numerical Algorithms Group (NAG), a worldwide organization dedicated to developing quality mathematical and statistical components and 3D visualization software including the recently released NAG Data Mining Components. Queries can be forwarded to sales@nag.co.uk.