Barry Leventhal *is a marketing statistician with more than 30 years of experience in market research, geodemographics and mining large databases. In 2009, he founded BarryAnalytics, an independent advanced analytics consultancy. He sits on the board of the IDM journal and chairs the MRS Census & Geodemographics Group. He is a fellow of the IDM, MRS and RSS.*

Papers

An introduction to data mining and other techniques for advanced analytics

Barry Leventhal

Received (in revised form): 13th August 2010

Abstract

This paper reviews the use of data mining (DM) for extracting patterns from large databases, held by companies such as banks, retailers and telco operators. The DM process is discussed, together with the ideal architecture, for applying this approach in a data warehouse environment. Some related techniques are identified — advanced data visualization tools for converting large volumes of data into pictorial form together with text mining and social network analysis for extracting structured data from unstructured text and relationships. The role of contact optimization is highlighted, as a method for optimizing the business value that a company can achieve from its DM activities. Finally, the paper suggests some initial steps in selecting a DM software product and offers the author's personal guidelines for the types of product that are likely to be most useful in different situations.

Journal of Direct, Data and Digital Marketing Practice (2010) **12,** 137–153. doi:10.1057/dddmp.2010.35

Keywords: data mining, advanced analytics, data visualization, text mining, social network analysis, contact optimization

Introduction

Over recent years, companies have been capturing increasing volumes of raw data from their operational systems, holding them in data warehouses and using them to forecast trends and support decision making. For example, mobile phone operators now commonly load their call detail records into a database for a period of time, while supermarkets store every shopping transaction with full purchase details.

Data mining (DM) systems provide the intelligence to analyse this vast quantity of raw records, extract patterns and convert the data into actionable information.

According to Berry and Linoff,[1] commercial DM has really 'taken off', over the last decade, due to several factors:

- Large volumes of data are being produced, as illustrated above, via automated data capture systems, and more sophisticated analytical software is needed in order to extract the patterns from these datasets.

Barry Leventhal
BarryAnalytics Ltd
9 Markham Close
Borehamwood
Hertfordshire WD6 4PQ, UK
Tel: +44 7803 231870
E-mail: Barry@barryanalytics.com

- Computing power has become more affordable, enabling companies to invest in powerful data warehouse systems that provide excellent environments for DM.
- Interest in customer relationship management (CRM) is strong, and companies are realizing the central importance of their customers and the value of their data.
- Commercial DM products have become available, drawing on techniques from statistics, artificial intelligence and machine learning.

The main focus of this paper is the use of DM software products to discover hidden patterns within large databases and harness these for business advantage. We shall be considering the DM process, its relationship with other toolsets for advanced analytics and some of the best practices in selecting and deploying this technology.

This paper has two main aims:

- to discuss the approaches commonly followed in DM on large databases;
- to identify some of the tools available, together with other closely related techniques, and provide some guidelines on selecting an appropriate solution.

What is DM?

DM can be defined in different ways — Hand *et al.*[2] define it as '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'. The author's previous employer, a large data warehousing company (Teradata Corporation, internal communication), describes it as '*A process of discovering and interpreting patterns in data to solve business problems*'. This definition contains three key stages — finding patterns, interpreting them in order to check their usefulness and finally using the patterns to solve business problems.

DM is associated with large volumes of data containing many attributes

Although not mentioned in this definition, DM has come to be associated with large amounts of data, usually more than 100,000 records — and many contain millions — which often have thousands of associated attributes or variables. Such datasets typically occur in sectors such as financial services, retail, manufacturing, telecoms, travel, transportation and the public sector — where organizations have large customer bases and customers make multiple transactions, sometimes on a minute-by-minute basis.

Owing to the size of such databases, sophisticated tools are required to discover useful patterns from the vast number of potential relationships — hence the role for DM in order to help with this task.

Over the last few years, a wider set of activities known as 'Advanced Analytics' has emerged. Advanced analytics is defined by the independent research firm, Forrester Research, Inc.,[3] as '*Any solution that supports the identification of meaningful patterns and* *correlations among variables in complex, structured and unstructured, historical, and potential future data sets for the purposes of predicting future events and assessing the attractiveness of various courses of action. Advanced Analytics typically incorporate such functionality as data mining, descriptive modelling, econometrics, forecasting, operations research optimisation, predictive modelling, simulations, statistics and text analytics*'.

Such methods broaden the use of sophisticated analytics beyond conventional DM into areas such as deriving new data by analysing unstructured text (text mining) or by extracting relationships between records (social network analysis (SNA)). We touch on some of these techniques later in this paper, under 'Other Tools for Advanced Analytics', as they directly help to increase the value of data and the benefits that can be gained through DM.

The main types of DM models

'Analytical modelling' is the process of pattern discovery in DM

The process of pattern discovery, when mining a dataset, is known as 'analytical modelling' in order to create a DM model. This activity involves identifying meaningful relationships between variables in the data, and employing those relationships to create predictive or descriptive models. The outcome is expressed as a formula or algorithm that can calculate a score (predicted value or probability) for each individual record, for instance of response, defection or repeat sales, according to the data values for that record.

There are two main types of DM models:

1. Predictive model — a model constructed to predict a particular outcome or target variable. Commonly used predictive modelling techniques include multiple regression (for predicting value data), logistic regression (for response prediction) and decision trees (for rule-based value or response models).
2. Descriptive model — a model that gives a better understanding of the data, without any single specific target variable. Commonly used descriptive techniques include factor analysis (to extract underlying dimensions from multivariate data), cluster analysis (for grouping a customer database into segments) and association analysis (for discovering relationships between items such as retail products).

A wide range of analytical techniques are available for predictive and descriptive modelling, drawn from the worlds of statistics and machine learning. In their 2009 survey,[4] Rexer Analytics identified that the core techniques used by most data miners are regression analysis, decision trees and cluster analysis. Explanations of these and other techniques may be found in a variety of sources.[1,5–9]

The relationship between DM and statistical models

As data miners often employ statistical techniques, such as regression analysis, it may be thought that DM is simply a modern term for 'statistical analysis' — however, this is not the case, for a number of reasons.

There are key differences between statistical analysis and DM

The development of statistical theory has its roots in the late 19th and early 20th centuries, before the advent of computer technology. Methods were required for making inferences based on relatively small samples drawn from the corresponding populations. The theory was developed for testing hypotheses and measuring significance of results, taking sample size into account, since analysis at population level was not a viable possibility. At the same time, the number of records and the number of attributes for which measurements were recorded were sufficiently small to enable each variable to be examined individually and transformed as appropriate for analysis and modelling purposes.

DM, on the other hand, is applied to databases that typically hold an entire population of customers, together with thousands of variables that summarize their transactional behaviour, payments history, campaign responses and so on.

Therefore, in any project, the data miner is no longer restricted to working with small samples — the full customer base is available if desired. However, this requires some differences in approach from traditional statistical methods — statistical techniques may give misleading results if applied to a vast sample size, which carries risks of over-fitting the model or producing unhelpful results in which every variable appears to be statistically significant.

Furthermore, in DM, the dataset is liable to contain a huge number of candidate predictor attributes (variables), for example, volumes and values of transactions by product, channel, brand or period — far too many to be individually assessed and transformed manually. DM solutions ideally provide automated tools for selecting relevant attributes and recoding them in the form of variables for use in analysis.

A further key difference is that statistical analysis will aim to identify a model that is statistically significant — that is, outperforms a random prediction — based on a set of significant predictor variables. However, this provides no guarantee that the model will perform sufficiently well to be of business value.

DM goes further than that, by including diagnostic results to indicate likely business benefits from the model. The assessment is produced by using two methods in combination:

(a) Prior to modelling, a random subset of data is excluded from the analysis, for use in evaluating the power of the model developed on the remainder of the data — this excluded subset is known as a 'hold-out' sample, and is more likely to give a fair indication of model performance than if the model development sample were used.

(b) Various types of tables and charts are produced in order to assess the predictive power of the model, using the hold-out sample. For example, if a model has been built to predict campaign response, then the lift chart will show how response rate varies by the probabilities predicted from the model (often grouped into deciles). This will help the users to decide whether the model is likely to deliver enough benefit to justify its deployment and select the model deciles that should be targeted.

Lastly, having built and evaluated a DM model on a sample dataset, the model will be deployed by applying the scoring algorithm to all 'X' million records in the customer database. Therefore, facilities for large-scale model deployment are essential — the form that this takes will vary from package to package, as we will see below.

Both DM and statistical analysis require that the data are organized as a simple rectangular table, where the rows (or records) represent individuals (eg customers) and the columns contain structured variables (eg demographics, usage or purchasing behaviour). Often, much effort is required in order to assemble this analytic dataset, as we discuss in the following section.

The variables in this dataset are 'structured', in the sense that each column contains either numeric or character (categorical) values coded in a consistent format. However, an increasing amount of information is captured nowadays in an unstructured form, for example customer comments, accident reports and e-mail requests. A technique known as 'text mining' may be used to read unstructured data and derive facts that can be represented by structured variables, and included in analytic datasets — this approach is discussed under 'Other Tools for Advanced Analytics' below.

How are DM models built and deployed?

In 1996, a consortium of companies jointly agreed on a standard process for DM — this process, known as CRISP-DM (http://www.crisp-dm.org/index.htm), describes the life cycle of a DM project as a set of six phases that are independent of software products and techniques. The high-level process is illustrated in Figure 1, while the phases in the process are summarized in Table 1.

Data preparation typically takes 60 to 70 per cent of the work effort in a DM project

While all six phases are essential to the delivery of a DM project, the Data Preparation phase usually takes most time — it will typically account for 60 to 70 per cent of the work effort and is critical to the project's success. The following tasks are carried out as part of the Data Preparation phase:

(i) Data are extracted from one or more sources (eg from tables in a relational database), manipulated into a consistent format (eg summarized or aggregated), and joined together into a single file or table.
(ii) Variables are profiled, examined for usefulness and transformed or recoded.
(iii) New variables (such as average spend per visit) are derived in order to improve model performance.

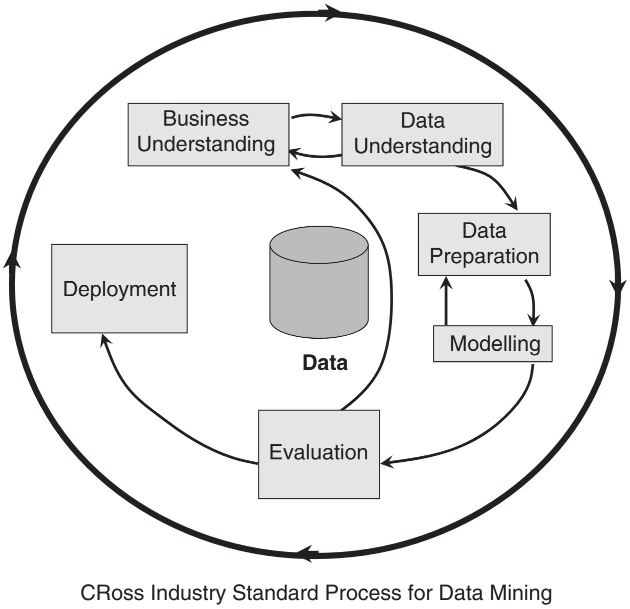

Figure 1: The data mining process (CRISP-DM).
Source: http://www.crisp-dm.org/Process/index.htm.

The resulting 'analytic dataset' is a file or table containing all the data to be used for analysis and modelling.

The analytic dataset contains all the data required for a DM project

Another key feature of CRISP-DM is the separation of the modelling and deployment phases — this has several implications:

(a) The modelling phase is likely to involve analysis of a representative sample from the database (typically up to 100,000 records). There are several reasons for preferring to use a sample, the most common being that analytical techniques such as logistic regression are computationally intensive, and thus, while model precision increases with sample size, there is little to be gained once the size exceeds 100,000 cases. At the deployment phase, the model algorithm is applied to score the entire database, or an appropriate subset of this population.

(b) In many business applications of DM, the modelling phase takes place only once, (until the model next needs to be updated); however, the deployment phase is carried out often on a regular basis, for example, to update customer scores each month, for example, on predictors of spend, credit worthiness or cross-sell opportunity, based on their most recent usage or behaviour.

(c) Given these differences between the modelling and deployment phases, the user may decide to deploy separate software solutions for these parts of the process, and adopt some method of transferring the model algorithm between them.

The issues of model deployment and model transfer are discussed below, under 'Data Mining Architecture'.

Table 1: Phases in the data mining process

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record and attribute selection, as well as transformation and cleaning of data for modelling tools.

Modelling

In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Evaluation

At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to the final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine whether there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the end-client can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases, it will be the end-client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will carry out the deployment, it is important for the end-client to understand up front what actions will need to be carried out in order to actually make use of the created models.

Source: Modified from http://www.crisp-dm.org/Process/index.htm.

Predictive and descriptive models may be applied separately or in combination

Predictive and descriptive models — How do they fit together?

As we identified earlier, there are two main types of DM model — predictive and descriptive.

Predictive models may be used to identify which individual customers are more likely to respond to marketing offers, use particular channels, pay off loans early (or late) and so on.

Descriptive models, such as segmentation exercises, may be used to identify the existence of different groups of customers, differentiable perhaps by motivations and needs, for which different marketing strategies and communication plans may be appropriate.

Both predictive and descriptive models may be developed either at individual customer level or, on an area basis, for example, geodemographic segmentation systems such as ACORN and MOSAIC. The understanding of the business problem and intended use for the solution will drive the choice of appropriate analysis units.

It is further possible that the eventual business use will require each customer to be scored on multiple models, and then a decision made according to their set of scores. Use of multiple models could occur in a number of situations, for example:

(a) A company that sends offers to mail order shoppers may wish to target customers based on a combination of propensity to respond and predicted order value, in order to maximize the value of goods ordered.
(b) A phone company that uses predictive models to identify likely churners may overlay future customer value forecasts, in order to focus its retention resources on 'at-risk' customers who would be valuable if they stayed.
(c) A bank may have different alternative product offers that could be sent to each existing current account customer, and will need to analyse the customer's propensities to take each of those products in order to decide on an 'optimized' action with greatest predicted return. At the same time, the analysis may depend on the customer's segment — which may determine the types of products that should be offered to someone with a given set of likely needs.

The third example situation is clearly more complex than the others — it requires a 'higher level' analytical solution that receives predictive and descriptive DM model scores as its inputs, and will generate an 'optimal' product offer decision for each customer. This type of solution — known as 'contact optimization' — is discussed below under 'Other Tools for Advanced Analytics'.

The 'shelf life' of a model

The shelf life or likely lifetime of a model depends primarily on the extent of change over time in the relationship between the characteristics of the model and the behaviour being predicted. Where the characteristics are fairly stable, a shelf life of several years may be achieved; however, any significant shift in this relationship would probably imply that the model should be redeveloped. For example, suppose that a mobile phone operator has been using a model developed 4 years ago, in order to target 'smart phones' to its subscribers. The model would have performed well for the first 2 or 3 years, when smart phones were primarily aimed at business customers; however, their market has been broadened to domestic customers during the last year or so. In other words, the model now needs to be redeveloped in order to target the wider audience for this product.

Best practice is to monitor model effectiveness using a random control group

The 'best-practice' way to identify when a model has reached the end of its shelf life is to include a randomly selected control group in all campaigns targeted using the model. This will enable model effectiveness — and hence the return on investment from DM — to be calculated by comparing response rates (or the appropriate metrics) between the target and control groups. Any major shift or trend in model performance may therefore be identified and should be investigated — causes could range from operational errors (eg unexpected changes to model input data) through to market shifts, as in the smart phone example. Therefore, although including random, less well targeted, customers in the campaign might result in some loss of sales, this should be seen as a necessary investment and an essential part of the DM process.

Model redevelopment is desirable if a new significant data source becomes available

Another reason for model redevelopment is if a new data source becomes available that will significantly improve the discriminatory power of the model. For example, if details of customer Internet and e-mail usage were obtained, these would be excellent predictors of need for a smart phone and thus would justify rebuilding the model in order to include those variables.

In a market that is more dynamic, with constantly changing factors or effects, the model shelf life may be considerably shorter. This implies that the modelling phase may need to be considerably faster and linked to the deployment process in a more automated way, and therefore model automation would become a priority. Some DM software products are particularly well suited to such requirements, and to creating models that are ‘disposable’ and ‘replaceable’ without great cost. These tools also enable users to build ‘intermediate’ models that can be deployed quickly and act as a stop gap, while more sophisticated model development is under way.

Similarly, model automation is a desirable feature if an organization needs to develop a large number of models — for example, if a business decided that it required separate models for all combinations of its six products, by four channels and ten regions, then 240 separate models would have to be built.

Software packages for DM

DM software packages usually contain all or most of the following functionality:

(a) A wide range of data preparation functions to profile and transform data from sources such as relational database tables, and create analytic datasets.
(b) Storage and management of analytic datasets.
(c) A range of multivariate analysis and modelling techniques such as regression, decision trees and cluster analysis, for model development and validation.
(d) Tools to allow modellers to visualize and explore models and their results.
(e) Facilities for automating the tasks in a DM project, for example, for setting up a workflow of steps that can be executed as a single process.

(f) Mechanisms for saving a model algorithm ready for deployment and a model management option for organizing models that have been built previously — these features are discussed more fully in the next section.

Table 2 shows some examples of widely used software products that contain such functions.

In addition to DM toolsets, there are a number of statistics packages that are useful for data analysis, manipulation and complex modelling. These generally require greater statistical skills and tend not to include automation or model management features. Some examples of statistics packages are shown in Table 3.

Internet polls imply that users tend to apply several tools in combination

A number of consultancies carry out user polls over the Internet and publish rankings of the most widely used DM tools (http://www.kdnuggets.com/polls).[4] As Internet polls are based on self-selecting samples, the results should be treated with caution; however, they indicate the wide range of tools being used and suggest trends in the marketplace. For example, the latest KDnuggets poll implies that the strongest growth has taken place in open source tools such as RapidMiner and R, which top the current ranking. Poll results also imply that users tend not to rely solely on one product for all their needs — more often, several tools are used in combination.

Table 2: Some software products for data mining

Product	Supplier	Notes
FICO Model Builder	FICO	
IBM Smart Analytics System	IBM	For use with IBM databases
IBM SPSS Modeler	IBM (SPSS)	Formerly SPSS Clementine
KnowledgeSTUDIO	Angoss	
KXEN Analytic Framework	KXEN	Employs structured risk Minimization for model reliability and automation
Oracle Data Mining	Oracle	For use with Oracle databases
Portrait Customer Analytics	Portrait Software	
RapidMiner	Rapid-I	Open source
SAS Enterprise Miner	SAS	
SQL Server Analysis Services	Microsoft	
Teradata Warehouse Miner	Teradata	For use with Teradata databases
TIBCO Spotfire Miner	TIBCO Software	

Table 3: Some examples of statistics packages

Product	Supplier	Notes
IBM SPSS Statistics	IBM (SPSS)	Formerly SPSS Statistics
R	Free Software Foundation	Open source
SAS	SAS	
Statistica	StatSoft	
TIBCO Spotfire S+	TIBCO Software	

DM architecture

By 'data mining architecture', we mean the approach taken to link together the analytical steps in the DM cycle (Figure 1), in order to form a repeatable business process. Without a workable architecture, an organization will struggle to build and deploy analytical models, particularly across large data warehouses.

When data volumes are relatively small (typically under 100,000 records) and very few models need to be managed, the architecture is less critical. DM can be deployed by extracting the required data onto a powerful PC or server, applying the model algorithm and then loading the scores into an appropriate application, for example, the campaign management system.

The choice of DM architecture becomes important when data volumes are large and/or many models need to be deployed

However, when the data warehouse contains millions of records and/or there are large numbers of models to be deployed, the choice of DM architecture becomes more important. In these circumstances, the best-practice approach involves:

(a) undertaking data preparation functions, such as profiling, transforming and building analytic datasets, directly in the data warehouse;
(b) converting DM models into SQL code that can be directly executed in the data warehouse, removing the need to export large volumes of data into another system for model scoring.

Both of these steps speed up the process, by eliminating data migration and by harnessing the processing power of the data warehouse system.

Over the last few years, the major data warehouse vendors have been moving towards enabling in-database mining, in order to facilitate this best-practice approach, either through their own software developments or by partnering with software suppliers — some examples are:

- IBM Intelligent Miner (IM), Oracle Data Mining (ODM) and Teradata Warehouse Miner (TWM) DM applications for IBM, Oracle and Teradata data warehouses, respectively;
- SAS Scoring Accelerator products for deploying models created in SAS Enterprise Miner within IBM DB2, Netezza and Teradata databases;
- IBM's acquisition of SPSS, resulting in updated products such as IBM SPSS Modeler Professional (formerly SPSS Clementine);
- Fuzzy Logix in-database analytics library for Netezza.

Communicating models between development and deployment

One of the enabling factors for in-database mining is the ability to transfer model scoring algorithms, developed in the modelling phase of the DM process, over to the deployment phase and run those algorithms in the data warehouse. Historically, model communication was often carried out manually — that is, the analyst documented the model scoring logic and passed those instructions to the IT department for coding and running — however, this method had many drawbacks and was highly prone to errors. Needless to say, the smallest interpretation error would produce incorrect scores and lead to serious prediction errors when the model was deployed.

PMML is a model definition standard that enables models to be transferred between DM tools

This problem was solved, a little over 10 years ago, when the Predictive Model Markup Language (PMML) standard was devised by the Data Mining Group (http://www.dmg.org/). PMML uses XML to represent DM models and can handle all of the common modelling techniques — including regression, decision trees and clustering. A software package may either produce PMML (ie output PMML files representing models) or consume PMML (ie read in PMML files created by PMML producers), or do both.

A second approach, available in some DM products, is to output the model scoring logic as an SQL program that can be run in a data warehouse environment.

A third option, taken by the Scoring Accelerator product from SAS, is to publish the model scoring (for models created using SAS Enterprise Miner) as a user function that may be executed in-database by partner data warehouse systems.

Model management software is another emerging trend

Model management

Over the past few years, there has been a trend among software vendors towards providing functionality for managing models — typical tasks being:

(a) storing the instructions for creating and refreshing analytic datasets (eg as a series of linked SQL programs);
(b) storing and organizing model scoring code, including version control and basic descriptions;
(c) executing analytic dataset creation and/or model scoring runs automatically, according to a predefined schedule;
(d) monitoring and reporting on model performance, including user notification.

Currently, a number of tools are available for model management, including:

- IBM (SPSS) Predictive Enterprise Services
- KXEN Modelling Factory
- SAS Model Manager
- Teradata Model Manager

Overall process

Putting the above approaches together, we arrive at the ‘best-practice’ architecture for DM — in the opinion of the author — which is shown in Figure 2. All stages in the process are ideally run in the data warehouse, with the possible exception of steps 3 and 4 that may use the analyst’s preferred DM tool.

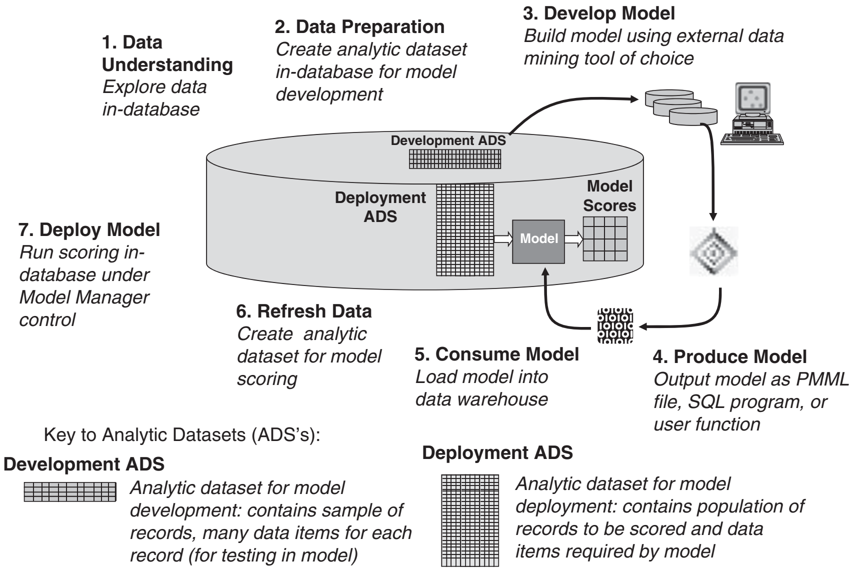
1. Data Understanding
Explore data in-database
2. Data Preparation
Create analytic dataset in-database for model development
3. Develop Model
Build model using external data mining tool of choice
Development ADS
Deployment ADS
Model
Model Scores
4. Produce Model
Output model as PMML file, SQL program, or user function
5. Consume Model
Load model into data warehouse
6. Refresh Data
Create analytic dataset for model scoring
7. Deploy Model
Run scoring in-database under Model Manager control
Key to Analytic Datasets (ADS's):
Development ADS
Analytic dataset for model development: contains sample of records, many data items for each record (for testing in model)
Deployment ADS
Analytic dataset for model deployment: contains population of records to be scored and data items required by model

Figure 2: Best-practice data mining architecture.

Other tools for advanced analytics

Business value is increased when multiple approaches are applied in combination

As we have discussed, DM is an important approach used in advanced analytics; however, increased business value can be generated by applying several approaches in combination. Some other techniques are outlined below, which can link with DM in order to form a more complete and powerful advanced analytics solution. Together with each technique, we identify a small number of solution providers — more complete lists of software products may be found elsewhere, for example on the KDnuggets website (http://www.kdnuggets.com).

Data visualization

Visualization makes it easier to interpret large amounts of data — most readers will be familiar with the charting options in packages such as Excel, which can display trends and patterns in spreadsheets. These tools are helpful, but they are only able to display relatively small numbers of data points within a static display.

Data visualisation tools enable massive amounts of data to be displayed and analysed graphically

Advanced data visualization tools are able to transform massive volumes of data into multidimensional pictures and animations that highlight areas of interest and help to understand complex data. As such, they can be applied either prior to engaging in DM — for example, to examine past behaviour and identify an appropriate subset to be mined — or after the analytical model has been created, in order to assess its performance. The latter use is particularly beneficial in combination with 'black box' DM techniques, such as neural networks and genetic algorithms, in order to identify the key variables in a model.

Some of the most popular visualization techniques are:

- Scatter plots, which enable interesting features to be discovered by overlaying other attributes onto a plot, combined with highlighting and interactive drill-down in order to 'dig more deeply'.

- Heat maps, which are used to locate 'hot spots' across a graph formed by interlacing two variables.
- Maps, which are used to examine data displayed geographically and search for possible relationships between the geographical objects. Maps may be relevant at any spatial scale — for example, national or regional maps can display the dispersion of a target market, while a map of a supermarket layout can display intensity of traffic or purchase patterns by department.

Visualizations may also be displayed as animations showing how patterns change over time, for example, the changes in activity levels during a day, within different parts of a city or retail outlet.

Advanced visualization techniques are available from a number of suppliers, including Advizor Solutions, IBM and SAS.

Text mining extracts structured variables from unstructured text files

Text mining

In many industries, valuable information is held in an unstructured format within text files, for example, the content of a customer complaint or an insurance claim description of an accident. As explained earlier, unstructured information cannot be directly entered into a DM tool — however, if it can be 'text mined' in order to derive structured variables, then these can be included. Text mining is the discovery of previously unknown information or concepts from text files by an automatic extraction process. For example, text mining (in conjunction with DM) could be used to identify that certain words used by insurance claimants had a particularly high association with a fraudulent claim.

Text mining solutions typically use linguistic analysis to extract facts from unstructured text. A number of different text mining tools are available, from suppliers such as Attensity, Clarabridge, IBM (SPSS), KXEN and SAS.

Social Network Analysis (SNA)

SNA extracts potentially useful variables about each customer, according to their relationships with others

Social networks are groups of people who are connected together in some way, for example, tend to interact or communicate with one another. SNA identifies such groups by applying network theory concepts such as 'nodes' and 'links' — 'nodes' are the individuals within the networks, while 'links' are the relationships between those individuals.

SNA has been gaining traction over the past few years, as analytics users have been starting to learn that SNA metrics are correlated with customer loyalty. For example, in the mobile phone sector, SNA can identify the members of each group or 'calling circle', determine the central communicator or 'key influencer' and extract various metrics about the strength of relationship within the group. If the mobile operator is concerned with spreading marketing offers by word of mouth, then these key influencers will be the best people to inform. Similarly, a good predictor of defection may be that a subscriber is in frequent contact with a person who themselves has recently defected.

Therefore, SNA can extract potentially useful new variables about the size, strength and composition of each customer's calling circle, for use in DM projects such as churn prediction.

SNA and link visualization are available from a number of suppliers, including Idiro and KXEN.

Contact optimization searches for the optimal allocation of the marketing budget over customers, products, channels and time

Contact optimization

Given that a business has created the DM models required in order to target products/services across its customer base, the next challenge is how to optimize the communications budget — in other words, what is the best offer to make to each customer and by which channel? — where, for example, only one of a number of offers can be made to an individual, by any one channel, at any one time. The technology that answers this question is known as 'contact optimization'.

Contact optimization 'sits above' DM, in the sense that it takes all of the analytical model predictions as inputs and searches for an optimal allocation of products and channels to customers over time. Furthermore, the allocation has to satisfy budget constraints, contact rules, and minimum/maximum volume limitations.

The differences between inbound and outbound communications imply a need for separate contact optimization approaches.

Inbound optimization is primarily concerned with delivering the 'best' solution for each individual customer who contacts the company's call centre or logs onto their website. This can be viewed as an extension of the customer management or CRM system — to supply the next best offer for each customer, based on a set of predicted propensities for the available products.

Outbound contact optimization aims to find the 'best' solution at an individual level and at the same time meet overall outbound marketing business targets and constraints. It enables the business to forward plan the communication mix and estimate the return over a future period, as well as compare alternative communication strategies.

Suppliers of contact optimization software include Experian, SAS, IBM (SPSS), TCP Marketing Solutions and Unica.

Initial steps in selecting a DM solution

No single DM solution is best for all purposes — users should 'mix and match'

With such a wide diversity of software products available in the DM marketplace, potential buyers can be faced with a bewildering choice. No single solution is 'best' for all purposes, and it is likely that users will 'mix and match' between several products, each with its own advantages for certain types of applications.

As Furness[10] advised, the selection of a DM tool should be approached in the same way as the selection of any IT product — by first defining clear business objectives and requirements, assessing the current and future state of use in the business and developing a set of selection criteria against which to rate the options. Furness also gave a checklist of steps and criteria, and advised that users should evaluate products on their own data, prior to making purchase decisions — looking at modelling performance, run times and scalability as sample sizes/model complexity increase.

Here are some guidelines on the types of DM tools that buyers would be advised to consider, based on the author's personal experience:

(a) If the requirement is to build a relatively small number of models based on a limited range of variables (as, for instance, in the automotive sector where suppliers have access only to occasional transaction events), then a 'hand-crafted' modelling solution should be adequate, using either a statistics package or a DM tool.
(b) On the other hand, if there are many models to be built using a large set of variables, (as, for instance, with a grocery retailer or a credit card provider, where customers generate a large amount of transactional information), then a more automated DM tool is likely to be more cost effective. Models created using automated tools are typically around 90 per cent as accurate as hand-crafted models, but should save 90 per cent of the analyst's time.
(c) However, model performance is critical in situations where correct decisions return a high benefit, or incorrect decisions incur a heavy loss — credit scoring models are an obvious example. In these cases, a non-automated DM tool that offers a choice of algorithms and in-depth analytics functions, in order to yield the best possible model, is likely to be more appropriate.
(d) If a large number of models will need to be managed, deployed and monitored, then consider a solution that provides model management functionality.
(e) If the models are going to be mainly used for targeting direct marketing campaigns, then look for a more 'marketing friendly' tool, (eg providing lift charts in preference to statistical diagnostics), that integrates well with the campaign management system.
(f) Finally, it is essential to plan the modelling architecture and consider how the models are going to be deployed — particularly when the data resides in a large data warehouse. Look for a DM tool that provides viable in-database deployment options, such as PMML files or SQL scoring code. Ensure that model deployment is fully tested for all the model types that are likely to be required.

Some risks of DM

The risks associated with DM need to be mitigated by good data quality, strong business focus and sound user training

There are some risks associated with DM — situations or scenarios that can result in serious problems or failure to deliver expected benefits. For example:

(a) Data quality issues — the data being mined must be of high quality, consistency and integrity. Failure to achieve this can be critical, both at the modelling and deployment stages in the process.

(b) Untrained users working with highly automated modelling tools can produce misleading or nonsensical results.
(c) Producing mountains of unusable or non-actionable results. Being able to identify patterns in a data warehouse is only useful when there can be a business application. Having lots of patterns without profit-generating applications can be a costly distraction.
(d) Poor evaluations of model efficiency, or lack of standards for evaluating descriptive results, can result in misuse of the findings and no gains from the process.
(e) Certain technical requirements apply in the modelling stage (such as not extrapolating outside the domain of the data), which is again why users have to be fully trained.

Conclusions

DM and related software products have been increasing in power and complexity over the past 10 years, due to developments such as in-database processing, mechanisms for communicating model algorithms between tools and systems for analysing non-structured data.

It is important to keep the business requirement and method of model deployment in mind, when selecting a DM toolset. Potential software buyers should evaluate the products that they shortlist before making purchase decisions.

No single product works ‘best’ in all scenarios and market sectors, and therefore users should be prepared to ‘mix and match’ between toolsets, ensuring that models may be communicated between them as required.

Business value is only generated when DM models are deployed as part of a continuous business process

Business value is only generated when models are deployed as part of a process, which includes continuous monitoring, evaluation, learning and refinement.

References and Notes

1. Berry, M.J.A. and Linoff, G.S. (2004) *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*, Wiley, Indianapolis, IN.
2. Hand, D.J., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining*, MIT Press, Cambridge, MA.
3. Kobielus, J. (2010) ‘The Forrester wave: Predictive analytics and data mining solutions, Q1 2010’, Forrester Research Inc. report, 4 February 2010.
4. Detailed report available from Rexer Analytics, http://www.rexeranalytics.com/Data-Miner-Survey-Results-2009.html.
5. Leventhal, B. (2006) ‘Data mining, analysis and modelling’, Chapter 2.4 of The IDM Marketing Guide: Best practice in direct, data and digital marketing, Institute of Direct Marketing.
6. Dillon, W.R. and Goldstein, M. (1984) *Multivariate Analysis Methods and Applications*, Wiley, New York.
7. Chiu, S. and Tavella, D. (2008) *Data Mining and Market Intelligence for Optimal Marketing Returns*, Butterworth-Heinemann, Burlington, MA.
8. Tsiptsis, K. and Chorianopoulos, A. (2009) *Data Mining Techniques in CRM*, Wiley, UK.
9. Nisbet, R., Elder, J. and Miner, G. (2009) *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press, London, UK.
10. Furness, P. (2010) ‘Marketing analytics software tools: How to choose and use them’ Henry Stewart ‘Introduction to Marketing Analytics’ conference, June 2010.