8

Data Mining: an Introduction – Case Study

Objectives:

- To study the importance of data mining in health care.
- We discusses how data mining can be packaged in such a way that professional can take part directly in data mining and thus assess the implications of using data mining in safely-critical or service-critical applications.
- Data mining offers a new approach to data analysis using techniques based on machine learning alongside the conventional methods.
- To discuss how data mining can be packaged in such a way that professionals can take part directly in data mining.

Abstract. This section outlines a case study describing the importance of data mining in health care taken from Clementine, SPSS Inc. USA. Health care and pharmaceutical companies face an explosion of data arising from clinical, administrative, commercial and scientific activities. There are many traditional techniques for analyzing data including statistics, management reporting and data display. Data mining offers a new approach to data analysis using techniques based on machine learning (algorithms derived from research in artificial intelligence), alongside the conventional methods.

These techniques work by "learning" patterns in data. They find patterns and make predictions, which elude all but the most expert users of conventional methods. In addition, they generate decision or prediction models, based on the actual historical data. These models are synthesized- not programmed explicitly by programmer or physician. Thus, they represent true evidence-based decision support.

Health care and pharmaceutical professionals have a special duty of care, as their decisions may be a matter of life and death for their clients. The esoteric nature of data mining can distance these professionals from the models. We discusses how data mining can be packaged in such a way that professional can take part directly in data mining and thus assess the implications of using data mining in safely-critical or service-critical applications.

S. Sumathi and S.N. Sivanandam: *Data Mining: an Introduction – Case Study*, Studies in Computational Intelligence (SCI) **29**, 217–229 (2006)
www.springerlink.com

8.1 The Data Flood

The proceedings of the 1995 conference in Knowledge Discovery in Databases (KDD) opens with the following quotations:

> *"It is estimated that the amount of information in the world doubles every 20 months. What are we supposed to do with this flood of raw data? Clearly little of it will ever be seen by human eyes."*
>
> *"Computers promised fountains of wisdom but delivered floods of data."*

We are familiar with this phenomenon of data explosion. Computerized systems collect data about a myriad of every day transactions: at supermarket checkouts, bank cash machines, airline tickets, phone calls, buying gasoline, etc. (the list is endless).

Health care providers, insurers and suppliers, and the pharmaceutical industries contribute their share

- Administrative systems log patient admissions and discharges, resource utilization in hospitals and practices, delivery and use of supplies, staff shift patterns and hours worked, cost of procedures and times taken for procedures.
- Every clinical act and its outcome are recorded. Patient records are being transferred to electronic form.
- Pharmaceutical knowledge increases daily, with new compounds, new dosage patterns, etc. Clinical trails generate huge amounts of data which must be analyzed and which, because of the controlled environment in which they are collected, should be an invaluable resource for future study.
- Marketing and sales by pharmaceutical manufacturers and pharmacies provide the return on the huge investments required bringing new drugs to the market. Every such sale can be captured as data and analyzed to help make more sales and predict future demand.
- Scientists in companies and hospitals gather mountains of experimental and laboratory data.

Information Systems managers are only too aware of this data explosion. They continually have to upgrade computers with more disk storage. The IT industry bombards them with offers of the latest databases, data warehouses, data marts and a host of data translation, transformation and reporting tools all aiming to tame the data explosion.

8.2 Data Holds Knowledge

Data holds the record of the organization's performance in all of its business areas. A hospital, which has been admitting patients for many years, has data it can use to estimate accurately the likely cost of treatment and length of stay of patient. A health care insurer has data about all of its subscribers, including those who defaulted on payments and those who generated above average claims. Today these organizations use statistical techniques to estimate the overall average figures to help plan the future. If only they could understand the data for each individual patient (or group of similar patients) *in detail* – they could use patient history as recorded in the data, to help make management or clinical decisions. As these decisions would be based on actual outcomes, this would truly be evidence-based care planning.

8.2.1 Decisions From the Data

Look at aspects of business knowledge we can discover, or the *"decisions from data"* which can be made:

- *Marketing.* One can tell from results of past sales campaigns who is likely to buy what and when
- *Forecasts.* Can past demands for a resource be used to forecast the pattern of demand for upcoming weeks or months?
- *Customer loyalty.* Some people transferred to another health insurer last month; can one detect the "signature" of those who defected? Can it be determined which are profitable clients whom one wishes to recapture? Can one identify others like them who may be at risk of detecting and thus plan pre-emptive offerings to keep the profitable ones?
- *Fraud.* When did it occur and what were the key signs? Do other cases show the same signs?

Provided the relevant data has been collected, is available and reasonably accurate, the answers to these questions, as well as host of others, may be hidden in the organization's data. Yet it is the volume of data collected that makes it impossible for humans to understand.

8.3 Data Mining: A New Approach to Information Overload

Many vendors, consultants and analysts make data mining appear complex, difficult, mysterious and expensive. It may sometimes be complex (involving many parts), but it need not be mysterious or difficult.

Data Mining simply means:

8.3.1 Finding Patterns in Data, which we can use to Better, Conduct the Business

Some people, especially in the United States, use the term *knowledge discovery* instead of data mining. In this section, the terms knowledge discovery and data mining are used interchangeably. Both describe the process of discovering a non-obvious pattern in data that can be used to for making better business decisions.

It turns out that vast majority of applications boil down to finding a relatively small number of types of data patterns. Here are some examples:

- *Classification.* To which set of predefined *categories* does this case belong? In marketing, when planning a mail shot, the categories may simply be the people who will buy and who will not buy. In health care, they may be high-risk and low-risk patients.
- *Association.* Which things occur together? For example, looking at shopping baskets, one may find that people who buy beer tend also to buy nuts at the same time.
- *Sequence.* Is essentially a time-ordered association, although the associated events may be spread far apart in time. For example, one may find that after marriage, people buy insurance.
- *Clustering or Segmentation.* Is like classification except that the categories are not normally known beforehand. We might look at a collection of shopping baskets and discover that there are clusters corresponding to health food buyers, convenience food buyers, luxury food buyers and so on.

Data mining is not mysterious; it is simply applied common or business sense.

8.3.2 Data Mining can be Breakthrough Technology

For the first 25 years of computing, the emphasis was on automating existing business processes. Order entry systems became online transaction processing systems, and existing manufacturing, accounting and other systems were implemented in the computer. During the last 10 years much of the emphasis has been on managing these automated processes better. Executive Information Systems (EISs), management reporting tools and more recently Online Analytical Processing (OLAP) tools have allowed management to observe the performance of the business as recorded by these computerized processes.

Data mining is *evolutionary*, in the sense that is an extension of the EIS. It can be used simply to uncover more subtle patterns in the data that help optimize the existing processes.

However, data mining can also be *revolutionary;* it can enable new, proactive business processes. The main reason for this is that in the era of the EIS, and with many OLAP tools, only summary information is available. These reports might indicate that fraud is a problem, but unless the detailed transactions are retained, there is no possibility to identify individual frauds. Another example is that until recently, banks charged one standard rate of interest on loans, depending on the circumstances of the borrower. This means that there are higher and lower risk borrowers in each interest rate band, the low risk borrowers are paying more than they should, effectively subsidizing the higher risk borrowers. With modern data ware houses and data mining tools, it is possible to segment customers according to the risks involved ultimately to charge an interest rate based on individual cases, rather than average risk. This model of lending is fundamentally different process from that based on average risk.

8.3.3 Data Mining Process in an Information System

Many enterprises now have a multilevel-information hierarchy. Business transactions are captured in an online database. This is optimized for high throughput and high security of transactions. This usage is incompatible with the requirements for analysis and management reporting. Every CIO dreads the killer SQL statement that could lock up the entire transaction database for minutes and end annoying and possibly losing the customer and staff.

For this reason, many have implemented a *data warehouse.* A data warehouse can be a complete copy of all the transaction information together with customer and product details. The warehouse is, however, free of the requirement to complete thousands of online transactions a second. It is organized to support queries, report writing and deeper analysis of information. Typically a warehouse is topped up with the most recent transactions from the online database (say) hourly, or overnight.

Such a warehouse rapidly becomes very large, and different departments tend to want to look only at different subsets of the data. For this reason, some companies are implementing data marts, effective departmental subsets of the enterprise data warehouse.

The only real prerequisite for data mining is a business problem plus relevant data. So data mining can be carried out on any data source. However, pattern finding is very demanding of computer power so it is unusual to mine the transaction database directly. More typically, mining is carried out on a data warehouse or data mart.

It is also common for data mining to require, or benefit from, additional data. This is often brought – in-geo-demographic or customer lifestyle data, which is combined with the organization's data about their own customer's behavior.

Successful data mining requires both business knowledge and some analytical ability. In the past experience, the former has proven most crucial, as business knowledge and common sense can go a long way toward steering the user into reasonable use of data mining tools.

In theory, the process of data mining is straightforward. Once a business problem has been identified, data, which is likely to have a bearing on that problem (or all data), is assembled. In practice, this is the most time-consuming part of data mining; data is rarely complete, and there may be tedious, but not usually too difficult work to be done in combining separate data sets.

Ideally, there will be cases that we think may contain a pattern related to the goal. Say we are trying to reduce the frequency of cases where there is an unplanned re-admission of a patient. We study a reasonable and representative selection of recent cases where such re-admissions occurred and also select at least an equivalent number of broadly similar cases where there was no unplanned re-admission. The approach is to divide this data into at least two subsets. One or more will be used for *learning* – "training" the models to distinguish between the re-admitted cases and the others. Different subsets of the data will be used for testing the ability of the model to correctly predict whether or not there is a re-admission. The model is used only if it achieves a consistent, reasonably accurate, prediction on these test cases.

Decisions from Data

The result of data mining is one or more decision models. These must be tested rigorously. Once tested, these decision models can be deployed in many different places throughout the enterprise. For example, a model may be trained to recognize cases of bad credit risk; this would be deployed at the sales transaction (or possibly the sales quotation) process. Another model might alert management to falling sales (or rising sales); this might be run immediately after the data warehouse is updated with the most recent transaction information. Both these models are *intelligent agents*, whose intelligence lies in applying the business rules and patterns learned during the modeling process.

In the health care domain, one might use a model when a patient is to be admitted for a particular treatment. Such a model, trained with information obtained from similar past cases, might estimate the total cost of treatment, or the likely length of stay. If there is enough detail in the past histories, this model may be able to provide forecasts for that admission, which are more accurate than assuming the average for the standard procedure in each case. Whether this is technically possible depends on the richness, accuracy, and completeness of the available data. Whether or not it is desirable might be a management or an ethical decision.

Also, as indicated above, models may suggest new processes. In a company, which sells insurance policies, for example, customer and product segmentation may indicate success in development of a new product and the customer group to which it could be profitably promoted.

8.3.4 Characteristics of Data Mining

As indicated above, data mining is normally about detail data (not summary data). If we want to identify individuals who will respond to a particular medication we have to look at details of the individual, not some average across many patients. Of course, if the business question is about summary data – for example, predicting the total annual costs of a department – then it may be appropriate to learn from such total costs. Data mining considers data at the appropriate level of detail for the business problem.

Data mining, despite the impression given by some is not blundering blindly about in data hoping to find something interesting. Many years ago gold miners worked by, washing away mountains. Today, effective miners know what they are looking for. They take samples before mining in earnest, and deploy business knowledge (geology, mineralogy and so on) before starting work.

The best data mining is like this intelligent prospecting. It works best if we have known patterns to learn from. These might to known clinical successes and failures, known re-admissions, bad debts, already detected fraud cases or known of operational failure. In each case, the goal is to learn the *signature* in order to recognize these cases again and take action to ensure good outcomes, prevent bad outcomes or mitigate their effect.

Although this technique of learning in *supervised* mode (i.e. with known cases whose signature we are trying to learn) is usually the simplest approach; it is possible to learn unsupervised. For example, we can apply data mining techniques to learn normal behavior (without known cases of normal and abnormal behavior) and so create a model, which will warn of deviations from the norm, whenever they occur. This is the basis for detecting frauds in cash transactions, for example.

Understand the Business or Clinical Problem First

As with most projects, the first requirement is to have a clear idea of what problem we are trying to solve; the more precise this is better. Avoid projects, which say, "Let's explore the data and see what we find." Projects which say "let's try and segment our customers into different types" are possible, though they are often much more tractable if the preferred groupings are first identified on business grounds (e.g. find high profit and low profit groupings). The easiest project has a precise focus. "Can the profile of cases which led to unplanned re-admissions be identified?" Why are they different from other similar cases? Can better predictions be made for unplanned re-admission?"

This overall approach is summarized as, first define a business goal, then deploy the data mining technology on the pertinent data and models, then deploy the models to reap the benefit.

8.3.5 Data Mining Technology

The idea of trying to turn data into useful information is not new. Traditional techniques include statistical analysis and data visualization (i.e. plotting graphs and charts), and these are the heart of the most Executive Information Systems (EIS) or Business Intelligence (BI) products. OLAP (Online Analytical Processing) allows users to interactively select subsets of the total data and produce on-demand management reports.

With these techniques, however sophisticated, the computer is acting in a *passive* role. The user says what set of attributes is to be displayed, and the computer "merely" computes the relevant graphs and reports. However visually appealing the graphical displays may become, and no matter how fluently the user to the can drill into the data to discover more detail, it remains the *responsibility of the user to identify the interesting trends*, or request the fruitful cross-correlation's. When faced with mountains of data, even the most diligent user may miss profitable opportunities.

These passive techniques worked well when only 10 or 20 pieces of information were available for each case. Then it was possible to examine all the cross-correlation's or plot all the combinations of four, five or even six factors that can be displayed on a computer screen. However, today, hundreds of pieces of information about each case are available, thousands if we are including transaction or case histories. Trying to find the appropriate combination of field values, which best predict a particular set of outcomes is like searching for a needle in a haystack.

While these passive techniques are important parts of the data-mining toolkit, the excitement arises from new technologies that *actively* support the discovery of new patterns. Using *machine learning techniques* from artificial intelligence (for example, neural networks and rule induction), the computer takes an active role. The user says "devise a method for predicting these business goals (it could be store profitability, likelihood of a prospect purchasing a product, the best treatment for a medical condition) taking given factors as input." Typical factors might be customer demographics, geographical location of sites, medical history – anything about the situation for which we have historical data .The algorithms "discover" which factors contribute to each goal and can consider all factors if necessary. The result is a decision procedure which, given the values for all the relevant input factors provides forecast (s) for the business goal(s).

8.3.6 Technology Limitations

It is not algorithms that discover knowledge. The function of an algorithm is to find a pattern, perhaps a *coincidence*, in the data. Only human expertise in the form of knowledge of the meaning and context of the data, can decide both, how the data might reasonably be analyzed and how to interpret and evaluate any results produced. It uniquely remains the reasonability of the professional to rigorously test any model or production and crucially to decide whether to act on its "recommendations." This is especially true in the health care and pharmaceutical area where decisions are literally matter of life or death.

But the people with the business or medical or expertise – often the "data owners" – are seldom skilled in analysis technologies and even less in machine learning. Until recently, they required an additional "technology" expert usually an external consultant, to operate the analysis technology on their behalf. As the sophistication, and hence the complexity, of analysis techniques increases and more different techniques are used together, the gap between the business professional and the technical analyst widens.

This situation is not ideal in a business sense because:

- The cost of such expertise is high.
- When modeling is complete, the data owner is left dependent on what may be an external, and still expensive, resource.

Many banks and credit card companies in the United States are under dilemma because they rely on models built at some considerable expense by external modeling companies for credit scoring direct marketing. As the pace of business increases and competitors repeatedly introduce new products to address gaps and niches in the market, customer behavior changes. Unless models are constantly updated they become obsolete. But the cost and time scales of external modeling (and also the costs of deployment) inhibit the banks from updating their models frequently enough.

The gap between business knowledge and analytical knowledge is even more important during modeling.

- The data owner is removed from direct participation in the analyses, limiting the use of his business knowledge to a time-consuming feedback process.
- Yet data mining is really more of an art than science. It is very hard work to follow train of thought hunches, if the business professional has to keep telling the technology expert what to do, and then wait minutes or often hours or days for the result.

8.3.7 BBC Case Study: The Importance of Business Knowledge

SPSS Inc. was retained by the BBC to develop a model that would predict the audience share that a proposed new TV program would achieve given it was transmitted at a particular time. This appears to be an ideal data mining application; the BBC has years of historical data showing what audience share watched each program. However, the context of a TV program is quite complex. The data includes:

- The proposed TV program (time of showing, genre, target audience type, star presenter, etc.)
- Preceding and following programs on the same channel
- Programs at the same time or overlapping with proposed program, on competitor channels
- The weather
- Major public or sporting events
- Time of year (e.g. winter, summer, etc.)
- Whether or not the program is syndicated across many regions

The project was carried out with one year's viewing data for prime time, between 6:00 and 10:00 in the evening. A neural network model was able to predict audience share within plus or minus four percent. It was not assumed to be particularly accurate, but the BBC was pleased. The prediction had similar accuracy to their best program planner's estimates. It took two years for these planners to become experts; with data mining the same accuracy is achieved in a few seconds. And as viewing habits can be changed we could retrain in a matter of hours to provide an updated model.

But there were some problem programs where prediction was much more than four percent wrong. This was disappointing and without explanation. It turned out BBC was very excited about many of results, as the neural network prediction had matched their own human expert prediction. Armed with this "confirmation" of their own estimate, they could go back to management with renewed confidence and look for other reasons why the programs had under-performed.

While the neural network can make predictions, it is very much a black box and provides little insight into why the prediction is being made. By contrast, another machine learning technique, rule induction, does provide an "explanation" in the form of explicit rules. For the BBC, the induced rules gave insight into "reasons" for the predictions and how much each factor contributes to the predicted audience share.

But it is the time for a health warning about these so-called explanations. Rules induced by data mining tools are merely descriptions of patterns found in the data. They may be explanations "which causal dependency"; equally they may be coincidences, and they may suggest causality where none is present. In general, knowledge of the subject is required to decide.

In the BBC project, the rule induction tools generated the following rule:

> *"Any program (X) which follows the UK soap opera will achieve six percent less share that if X is put on at any other time."*

We might be tempted to draw the conclusion that UK soap operas are so bad that they cause the audience to turn off (or switch channels), leaving them so irritated that they wont turn on (or switch back) for the next program, no matter how good it is.

However, the expertise of BBC planners immediately provided a much better explanation. UK soap opera figures were dominated by one program, East Enders. This achieved the highest market share of any BBC TV program. So the commercial channels specifically targeted this large audience by putting on one of their best programs (a crime serial called The Bill) immediately after East Enders. The data mining had found a compact, accurate description of a pattern in the data, but it was not an explanation.

This case illustrates the importance of having business knowledge and experience intimately associated with the data mining and interpretation of the results. It was the desire to bridge the gap between professional and business data owners and the technical analysts and modelers, that led SPSS to develop Clementine – an advanced, interactive, data mining system for end-users.

Project Clementine

In 1992, Integral Solutions Ltd. initiated "Project Clementine", aiming to build a comprehensive data mining system accessible to business and professional end users such as doctors. To make the advanced analysis techniques available to data owners requires:

- a variety of techniques including visualization , statistics and machine learning
- packaging with technology details hidden from the user
- delivery in an intuitive, easy – to-learn tool

The main technique to achieve this is "visual programming". From palettes, the user selects icons representing data sources, manipulation, graphs, reports, learning algorithms, etc. The icons are connected to define data flows, and edited through pop-up dialogues to define the details of the processing involved. The stream of icons provides a natural representation of the screen of the data mining process.

Users are encouraged to interact with data. Data features and patterns – a cluster, for example – can be identified using the mouse; users then generate icons that select cases corresponding to the defined regions. These facilities allow rapid data exploration, and formulation and testing of hypotheses based on observed features.

The configuration of the machine learning engines is automatic. Users express only high-level preferences – such as "favor generality over accuracy"– and Clementine configures the tools by considering both user input and by examining the structure of the data. The user is protected from consideration of the details of the technologies involved. (Tools can also be used in "expert" mode.)

Feedback from users suggests that this approach does indeed make sophisticated analyses accessible to professional and business users. Clementine is now in use in such areas as: banking, insurance, telecom, electricity, real-time financial trading, dentistry, childcare, pharmaceuticals and toxicology; with users ranging from IT specialists and dedicated data analysts to financial managers and biochemists.

In health care, the duty of care which physicians and management owe to patients requires that responsible professionals are fully involved in, and fully comprehend the implications of the decision models, which influence the choice of patient care. A user interface such as that of Clementine makes this possible.

8.3.8 Some Medical and Pharmaceutical Applications of Data Mining

Described briefly are two applications of data mining in science and medicine:

- *Drug discovery.* Very large databases now exist describing the composition and properties of countless chemical compounds. Data mining is now being used by pharmaceutical companies to research for promising substances that may from the basis of new drugs.
- *Skin corrosively.* Univeler's Colworth laboratory in the UK is responsible for assessing the effect on the human skin of possible new cosmetic, skin care and household products. At present, there is great reliance on animal testing. Data mining is being used to learn the known effect on the skin of known chemicals and hence, predict the effect of new preparations. The effect is to allow much more the screening to be done using computers and reducing the need for animal testing.

The range of applications is limited only by the availability of systematic data.

8.3.9 Why Does Data Mining Work?

Data mining can be remarkably effective and can find patterns and relationships that elude other techniques. Some of the reasons for this are:

- Since it works simply by learning from data it can (with right tools) be easy-to-use. This means business experts can contribute directly. This contribution often provides the magic ingredient.
- Results of data mining are very easy to test. Since the goal is to "predict" the past, the results are known; accuracy of the models can easily be determined.
- Models are rooted directly in the business experience. We do not have to depend on the skills or intuition of a programmer to synthesize a model. They are created automatically from the data and represent an unbiased distillation of the business experience.
- Decision models are generated automatically. This is a fast, cost-effective procedure and there should be no (programming) errors in the model.
- Models are easily updated by re-learning.
- Many of the techniques can handle many input factors (Clementine has used 7000/record) while the same techniques can ignore input factors, which don't contribute to a particular decision.
- Some of the techniques, notably neural networks, can discover complex non-linear models. It takes considerable expertise in classical statistics to deal with non - linear relationships.
- Good data mining tools (Such as Clementine) allow users to mix and match many techniques to solve problems. User's confidence in the results is increased if multiple techniques are deployed, and they all provide similar predictions.

8.4 Summary

Data mining is a new approach to finding patterns in data. In addition to using traditional techniques of data visualization and statistics the modern data mining toolkit includes machine-learning algorithms. These algorithms synthesize the models directly from data. The key to successful data mining is to define the business or clinical problem to be solved. New knowledge is not discovered by the algorithms, but by the user. The user's business or clinical knowledge is essential to determine what factors to consider for particular applications, and how the various input factors may need to be combined. Clinical, business or scientific expertise is always required to determine whether models can safely or usefully be employed. Some data mining tools like Clementine package the techniques so they can be used directly by and the results can be understood by business or clinical professionals.

8.5 Review Questions

1. What is data flood?
2. What are the data mining technology limitations?
3. Explain about the mining in BBC – Case study