PhUSE 2010

Paper SP02

Using Data and Text Mining to drive Innovation

Dave Smith, SAS, Marlow, UK

INTRODUCTION

Dr Diane Jorkasky's keynote presentation at PhUSE 2009 challenged the audience to exploit the data held by pharmaceutical organisations to drive their businesses forward. This presentation will discuss the possible practical steps that programmers and statisticians can take to learn more from our previous trials. The areas covered will include:

- Text analytics - Using techniques such as Text Mining or Sentiment Analysis to exploit unstructured sources. These can range from verbatim adverse event descriptions to blog entries and tweets.
- Data mining - Using advanced analytical techniques to identify patterns and associations in the data that indicate good or bad implications for the company.
- Simulation – Using existing data to generate the parameters for simulations of clinical trials.
- Forecasting – Predicting the most likely usage of clinical trial supplies.

ANALYTICAL DECISION MAKING

Decisions in life are frequently made on the basis of gut feel or instinct, and business decision making is no exception. This has many dangers, as what seems instinctively right may be completely misguided. For example, if you were able to fold a piece of paper in half forty times, what would you expect the combined thickness to be? Gut feel might tell you that the result would be a few inches or feet thick, whereas the correct answer for a piece of paper 0.193mm thick is 131,862 miles (or about halfway to the moon).

The application of numerical methods to real world problems is the only way to eliminate bias in decision making and make full use of the available facts. This paper will cover some of the common techniques that can be applied to data available to the pharmaceutical industry and how this can take the industry forward.

DATA MINING

WHAT IS DATA MINING?

Data Mining is the analysis of large volumes of data to uncover previously unsuspected patterns for business benefit. Data mining covers a wide variety of techniques such as Predictive modelling, Variable Selection, Association and Segmentation.

The Data Mining Process

Data Mining is best conducted using a repeatable process that can be institutionalised, ensuring repeatability and efficiency. The SAS approach to this is through the SEMMA process, and SAS Enterprise Miner has been built to exploit this. SEMMA stands for Sample, Explore, Modify, Model, Assess, reflecting the main steps in the data mining process:

Sample – select the subsets of data to build, train and test the model as it is created.

Explore – use visualisation techniques to identify potential patterns and relationships in the data.

Modify – change the structure of data using techniques such as log transformations or binning the data into groups in order to make it analysis ready.

Model – apply a sequence of modelling techniques to attempt to explain the target variable of choice.

Assess – compare the results of the modelling exercise in business terms (e.g. Lift, profit) to select the best model to deploy.

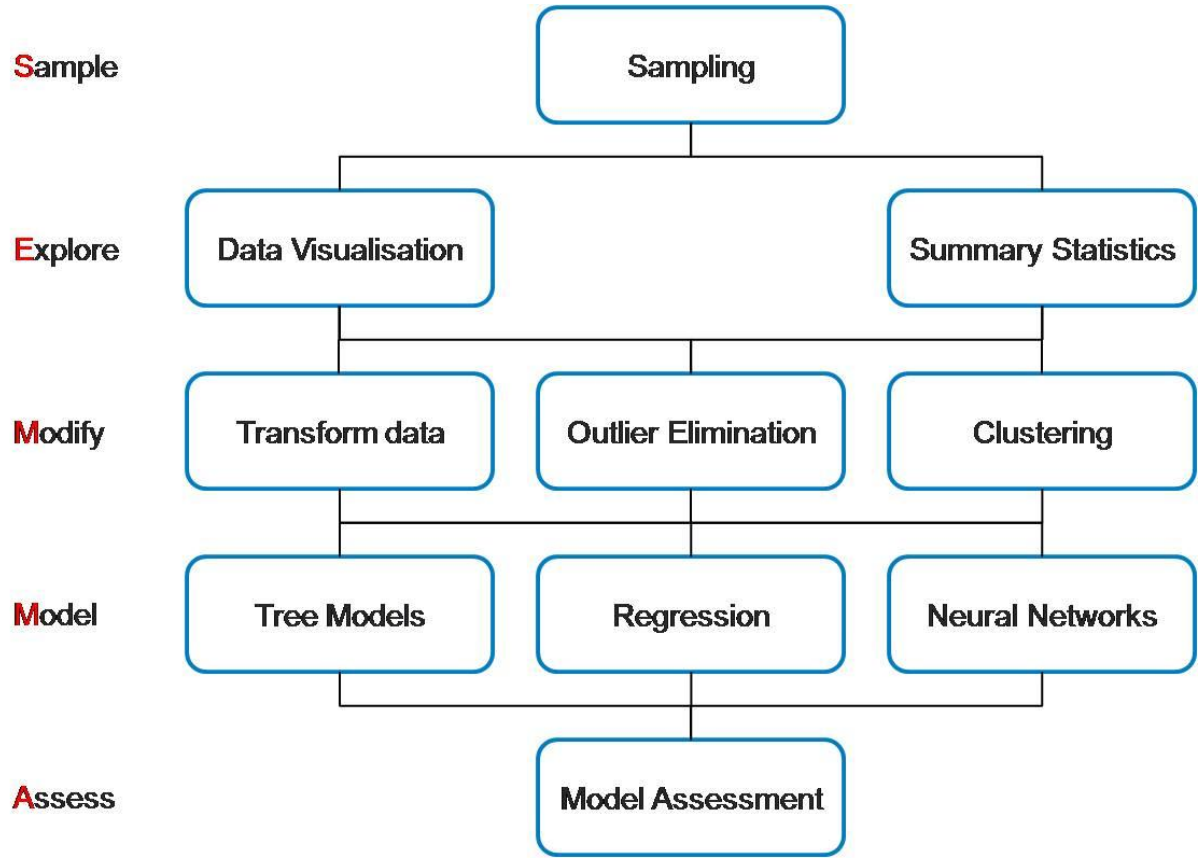

Figure 1 - The SEMMA Process

The data mining process generally needs to be a cycle of continuous improvement. This is because in the real world a model will typically have a lifespan of six months before it ceases to accurately reflect the situation and needs to be modified.

COMMON DATA MINING TECHNIQUES AND THEIR APPLICATION TO THE PHARMACEUTICAL INDUSTRY

Predictive Modelling

Predictive modelling is a technique used to determine the likelihood that something will happen at some point in the future. Typically this means regression analysis, where a linear or logistic model is developed from real world data to determine the causes of an event or observed result.

Predictive modelling can be used to answer questions such as

- What is the probability that a trial will need to re-open recruitment due to excessive numbers of withdrawals?
- Will intervention with a particular drug prevent hospitalisation?
- How much pharmaceutical product will be wasted in the supply chain due to stability issues?

Case Study: Preventing Hospitalisation

A US healthcare provider created a predictive model for hospitalisations in their Cardiac and Asthma patients. They collated the information that they had about the patients to see what factors led to hospitalisations. When similar conditions occurred in other patients they took early intervention and reduced hospitalisations across these groups by over 80%.

Variable Selection

Variable Selection is a technique used to reduce the number of possible variables that are used in the modelling step. This is commonly used to avoid over-fitting the data to the model, where you use so much information in building the model that the model is only useful for describing the sample data and not for the general population. It is also more generally true that a model with fewer parameters that adequately predicts the result is more desirable than a more complex model. This is because there is often a cost associated with collecting the data needed to feed the model and it is generally easier to explain.

Variable Selection can be used to answer questions such as

- Which genetic markers determine whether a patient will respond well to a treatment?
- Which process parameters predict whether a manufacturing batch will fail QC?

Case Study: Reducing a Problem in Manufacturing

A large pharmaceutical manufacturer was having problems with excessive numbers of batch failures on one of their biggest products. They had over 700 variables describing the characteristics of each production run; they were able to reduce this to the important 60 or so variables and use these to build a predictive model for batch failures. The model was so successful that they have almost eliminated batch failures at a considerable saving to the company.

Association Analysis

Association Analysis is a technique used to identify events that occur together, possibly in a particular sequence or order. It can also be used to identify disassociations (i.e. events that do not happen after a particular one has already occurred).

Association Analysis can be used to answer questions such as

- Which drug combinations are associated with adverse events?
- Which of these associations are strong enough to need investigating?
- Does the drug cause an adverse event or is the drug a common treatment for the condition?

Case Study: Identifying the Causal Links for Diseases

The Institute for Clinical Evaluative Sciences (ICES) produces scientific research that influences healthcare policy in Ontario. Using SAS Enterprise Miner researchers at ICES were able to identify safety issues with medicines and have had some taken from the market. They were also able to distinguish the causal factors of diabetes. It found that in outlying parts of Toronto, high immigration and low income equals high diabetes rates. Yet in the downtown core there was high immigration and low income, but surprisingly low diabetes rates; ICES was able to relate this to environmental factors such as access to public transport, walking trails and high-quality food sources.

Segmentation

Segmentation is a technique used to create groups that have similar characteristics. It can also give a measure of how ‘different’ each group is from the others.

Segmentation can be used to answer questions such as

- What are the attributes of a typical Brand A prescriber compared to other physicians?
- Which group of physicians are good prospects for investigators for my new trial?

Case Study: Identifying Prescribing Potential

A large pharmaceutical manufacturer was making very large numbers of calls on prescribers in the US market. The market was very saturated, and they were calling on nearly half of all US prescribers. The normal methodology was to select the top deciles of prescribing in a particular market and call on them, the theory being that these had the highest potential to prescribe more. Using SAS Enterprise Miner, the team were able to identify the high potential segments and describe their characteristics. As a result they were able to reduce the

numbers of prescribers that they called on by 40% for the same level of potential. An additional benefit was that they were calling on prescribers outside the top deciles, which gave them a competitive advantage over companies using the standard top decile methodology.

TEXT ANALYTICS

WHAT IS TEXT ANALYTICS?

Text analytics is the application of analytical techniques to unstructured sources such as comments or abstracts and includes techniques such as Text Mining and content categorisation.

Text Mining is the process of discovering and extracting meaningful patterns and relationships from text collections. Typically this involves reducing words to their root components (for example run, running, and ran are all translated to run) and then reducing documents by analysing the frequency with which words occur. The resulting reduced documents are then clustered into similar groups, which often indicate unsuspected patterns in the data.

Content Categorisation is a technique where similar documents are automatically categorised together, increasing the speed at which themes can be identified and documents of interest can be located.

APPLYING TEXT ANALYTICS TECHNIQUES IN THE PHARMACEUTICAL INDUSTRY

Text mining can be used to answer such questions as

- What group of adverse events might suggest a syndrome effect of a particular drug?
- Who is the Key Opinion Leader in a particular Therapeutic Area?

Content Categorisation can be used to answer such questions as

- Which abstracts should my researchers study first?
- Which tweets or blog entries might suggest a safety issue with one of my companies' drugs?

Case Study: Key Opinion Leader Mapping

A large pharmaceutical organisation was looking to understand the key influencers in the Rheumatoid Arthritis market to help increase sales of their product. They used a combination of Text Mining and Content Categorisation to identify abstracts of interest related to the Rheumatoid Arthritis market. They then categorised the abstracts into positive, neutral or negative towards particular brands. They then further identified the authors and their sphere of influence, using the so-called friends of friends measure where the number of connections with other authors is analysed. The key market opinion leaders could then be identified from their sphere of influence scores.

An interesting adjunct to this analysis was that the textual analysis of the abstracts identified a number of safety signals related to Vioxx; this was of only historical interest at the time that this work was done, but it did prove that text mining of abstracts could be used to identify drug safety issues.

SIMULATION

WHAT IS SIMULATION?

Simulation involves generating random variables fitting particular distributions to represent a real world situation. This is done many times to build the most likely outcome for a set of actions.

APPLYING SIMULATION IN THE PHARMACEUTICAL INDUSTRY

Simulation can be used to answer questions such as

- Given what we know already, what is the likelihood of a positive study outcome?
- What are the optimum starting dose regimens for an adaptive design study?

Case Study: Clinical Trial Simulation

A US Biotechnology company needed to ensure that its clinical trials had the best chance of a definite answer. The final study power can often be difficult to calculate using exact methods, for example if there are a lot of dropouts expected or if the distributions are difficult to model. Simulation gives a good way to manage this variability and assist decision making. The problem is that simulation generates massive amounts of data to be analysed and so computing resources are quickly stretched. The solution was to build a Linux cluster with 120 processors, which improved the number of replicates that could be processed from 1,000/day to 10,000/hour. The simulation application has now been made publicly available on the Biopharmaceutical Network; see the reference below for further details.

FORECASTING

WHAT IS FORECASTING?

Forecasting is the use of historical data to predict future events. There are many different types of forecasting, and each will have its place in solving a particular problem. The less sophisticated methods such as exponential smoothing tend to analyse the underlying trend and any seasonal components and leave any other factors as unexplained; this can be a gross oversimplification. More capable methods that can incorporate causal variables allow external factors to be included in the model, reducing the unexplained components and thus increasing the accuracy of the model. Another consideration is the way in which hierarchies are managed. Unless the hierarchy is considered as part of the model, the constituent parts will not add up correctly, and details at the lower levels will often be incorrect. For example if the total sales of Aspirin were being forecast the number of large size and small size packs would not add up to the total number of packs.

APPLYING FORECASTING IN THE PHARMACEUTICAL INDUSTRY

Forecasting can be used to answer questions such as

- What are the likely drug supply needs in a study?
- How much of each product do I need to keep in my supply chain?

Case Study: Optimising Inventory Levels

Alcon Inc. is the world's leading eye care company, and produces pharmaceutical, surgical and vision care products. One of the most difficult product lines to manage is the Custom Pak® surgical procedure tray, which contains all the elements a surgeon will need to perform a single surgery customised for the individual surgeon. The SAS forecasting system has been in place for nearly 20 years, and allows accurate planning and reduces inventory while delivering nearly 100% customer service levels. Alcon acknowledges that SAS has helped save the company more than $6 million in inventory costs alone.

CONCLUSIONS

There are a great many opportunities to drive innovation in pharmaceuticals through analytics. As stewards of information assets within our organisations we have an opportunity to make things better for all the key stakeholders (patients, payers, and sponsor companies), for example

- Reducing waste through accurate prediction of the requirements for study drug during a trial.
- Improving clinical decision making through trial simulation which may avoid conducting unnecessary trials.
- Reducing the length of trials through accurate prediction of study withdrawals.
- Identifying new indications through text mining of verbatim comments.

If we as programmers, statisticians and data managers take these opportunities then the pharmaceutical industry has a great future in Europe.

REFERENCES

SAS Global Forum 2010 Paper **Clinical Trials Simulation: A Publicly Available, Grid-Enabled, GUI-Driven SAS® System** by Peter Westfall, Alex Dmitrienko, Cynthia DeSouza, and Alin Tomoiaga

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dave Smith
SAS UK
Wittington House
Henley Road
Medmenham
Marlow
Bucks
SL7 2EB

Work Phone: 01628 404379
Fax: 01628 490550
Email: david.smith@suk.sas.com
Web: http://www.sas.com