Mining Multiple Models

Graham J. Williams
Adjunct Professor
School of Information Sciences and Engineering
University of Canberra

Graham.Williams@togaware.com

Abstract

Data mining is much more than simply building statistical models from large collections of data. In particular, this paper records a core task of mining as exploring through the space of models that are built in a data mining project. The idea was first introduced through the concept of multiple inductive learning (MIL) (Williams, 1988, 1991) and further developed in practise as mining the data mine (Williams and Huang, 1997). Many data mining advances that have since emerged have further developed the idea: multiple modelling, ensemble learning, bagging, and boosting all help the data miner explore different ideas and look for different insights in modelling. In this paper we review these ideas and a number of data mining projects that highlight the significant role played by mining the data mine.

1 Introduction

A data miner is engaged in the activity of aggregating very large collections of data to explore for new insights and understandings that will provide improvements for some process of interest. Application areas include customer relationship management, fraud prevention and control, and risk rating, to list but a few. Data mining is commonly defined as the non-trivial extraction of novel, implicit, and actionable knowledge from large databases (Fayyad et al., 1996).

The tools deployed by a data miner include common statistical modelling approaches as well as modelling approaches developed from research into machine learning and artificial intelligence. Traditionally, this means building decision trees or logistic regression models or neural networks.

Data underlies data mining and comes in many shapes and sizes. For data mining, the data is generally characterised by its sheer size. Its size is one of the key differentiators from traditional research in machine learning and statistics. Each entity (which may also be referred to as a record in data base terminology, or a training instance in machine learning terminology, or samples in statistical terminology) might be described by anywhere from 10 to 20,000, or more, features, and we may have from 20 to 200 million, or more, entities. Generally, small sets of entities arise in situations where we have very many features describing such entities, as is typical in genomics research, image data, and text mining. Datasets with fewer features but many more entities are typical in industry and government describing clients or customers. Thus the data is often of at least megabytes in size, usually in the gigabytes, and less frequently, in the terabytes.

Traditional approaches to modelling and data mining tend to deal with flat data in the form of a single row of data representing a single entity, with no relational data explicitly allowed. That is, links between entities must be captured in some other way, and repeated data needs to be aggregated in some way so that all entities have the same signature (the same number of features describing each entity).

Complex relationships, then, are generally not mined in data mining. In the administrative medical domain, for example, the entities that exist include patients and doctors, but also receptionists, pathologists, specialists, insurance claims officers, etc. Complex relationships exist between all these entities but generally remain too complex to be handled by today's data mining technology. Instead, the complex relationships need to be re-represented in a simpler, flatter form.

Whilst statistics provides many of the traditional tools deployed for modelling in data mining, the data miner spends much more time through other phases of a data mining project, which includes business understanding, data understanding, data preparation and cleaning, modelling, evaluation, and deployment (Fayyad et al., 1996). Common wisdom, indeed, is that modelling is just a small portion of the overall task (often less than 10% of the overall effort in any data mining project).

More broadly, data mining is about deploying multiple technologies to enable data exploration, data analysis, and data visualisation of very large databases at a high level of abstraction, generally without a well defined, specific hypothesis in mind, to extract knowledge from the data in any, and in many, ways.

The technology deployed in data mining comes from a diverse arena of research disciplines, beginning with databases, quickly drawing in machine learning and statistics, and encompassing high performance computing, computational mathematics, intelligent systems, visualisation, and web services. Together, these technologies deliver a rich, if sometime diverse, toolbox that a data miner deploys to deliver knowledge from data by analysing relationships in information.

In this paper we explore what might happen *after* we have built a model. Indeed, we review the idea of building multiple models and then exploring these models for the insights we need to deliver from data mining. Projects deploying such an approach are briefly described.

2 Mining Models

Making the simple observation that, in building decision tree models, for example, choices between different splits sometimes only marginally differentiate variables, Williams (1988) introduced the idea of building multiple decision trees and combining them into a single model. This began the data mining approach of exploring through a much richer space of models to identify and extract more information than otherwise would have been. It also later eventuated, from theoretical studies by many others, that ensemble learning was a good approach to model building and data mining (Hastie et al., 2001).

The original work of Williams (1991) used the Australian Resources Information System (ARIS) database (Walker et al., 1985). This consisted of some 11,000 entities, each recording extensive geographical information about a 700 square kilometer region of Australia. In particular, the rangeland regions of Australia were used in the study (8,000 entities). For each region 40 features were selected describing dominant soil type, vegetation, moisture indicators, and distance to nearest seaport.

The output variable for the study was a measure of the viability of the pastoral use of the land (for sheep and cattle grazing). Some

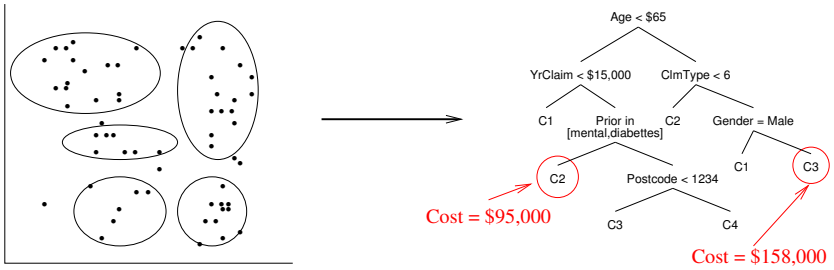
106 entities had been manually assessed by pastoral experts as to their viability, and this small dataset (although, at the time, regarded as reasonably sized) was used for model building. A version of the ID3 algorithm for decision tree induction (Quinlan, 1986), using the information-theoretic cost function, was used.

In building models in this domain, the decision tree algorithm only marginally chose one variable over another, to result in sometimes quite different looking trees. This fact was taken advantage of, rather than thought of as a problem, so as to produce multiple models, each giving different, but useful, insights into the domain. The final model developed for this domain consisted of multiple decision trees, with conflicts between the models being resolved logically. The system was called MIL for Multiple Inductive Learning.

The key outcome of this research was the idea of building multiple models and combining them. Follow on research took this idea further in the context of data mining with the realisation that model building was really the starting point to achieving the goals of data mining. Williams and Huang (1997) introduced the concept of mining the knowledge mine, and hot spots data mining.

The basic idea is that of building models that can be decomposed into smaller units that effectively describe different regions of a dataset (or population). Converting decision trees to rule sets is common practise, starting with the C4.5 tool (Quinlan, 1993). Rules, generally in the form of a conjunction of conditions, can then be used to symbolically describe these different regions of a dataset. Indeed, we can think of each set of conditions, each rule, as a nugget! A nugget captures some collection of entities, and our task in data mining is to determine which nuggets are the most interesting. The generation of the nuggets can be left to a variety of approaches, but a common one we introduced is to combine clustering with tree building where the cluster identifier becomes the target variable in the tree building.

We can picture the hot spots data mining process as follows:



Working from left to right, we start with a dataset (a 2 dimensional dataset in this case) that has no specific target variables. Thus we have an unsupervised learning problem. By clustering the data in some way, for example using traditional k-means, we can end up with a collection of clusters. For a very large dataset (e.g., millions of entities) we may indeed end up with quite a number of clusters (upwards of 100 or even 1000).

Each cluster can then be considered as a class, and entities will then belong to one class or another. This class can then be used as a target variable for a supervised learning problem. Using a decision tree builder we produce a set of rules (from a single decision tree each rule corresponds to a single path from the root node to a leaf node - providing a conjunction of conditions).

Each of these rules (or paths, or perhaps we can call them "nuggets") can then be considered independent of any other nuggets. The concept is to then explore through this space of nuggets searching for any that are interesting by some measure.

The nuggets might be as simple as the following:

Nugget 1	Age is between 28 and 35 **and** Weeks $\geq$ 10
Nugget 2	Weeks $<$ 10 **and** Benefits $>$ \$350

We note again that they are simple conjunctions of conditions.

For each nugget we collect together summary data about those entities that make up the subset of the dataset described by the nugget. This might include things like the size of the nugget, and average values of various features over those entities in the nugget, or measures of how far the nugget average is from the population average, and so on.

A simple example might be the following table where perhaps we have a total of just 280 nuggets and we might collect various summary items as in:

Nugget	Size	Age	Gender	Services	Benefits	Weeks	Hoard	Regular
1	9000	30	F	10	30	2	1	1
2	150	30	F	**24**	**841**	4	2	4
3	1200	65	M	7	220	20	1	1
4	80	45	F	**30**	**750**	10	1	1
5	90	10	M	12	**1125**	10	**5**	2
6	800	55	M	8	550	7	1	**9**
...								
280	30	25	F	15	450	15	2	6
All	40,000	45		8	30	3	1	1

Specific nuggets are then scored as to their interestingness based on a number of measures. For example, the bold entries in the table indicate values that are found to be more than two standard deviations from the population values. Thus we add scores to these nuggets. By this we produce a ranking of nuggets which can then be explored by domain experts in order, looking for new insights.

Over many different data mining projects, these ideas have repeatedly shown themselves to provide more insights into the relationships between entities and features, with respect to some target variable. In the following section we briefly review a number of these projects.

3 Applications in Health

Australia has a universal health care system the has been providing primary medical care to patients since 1975. For administrative purposes (i.e., to make payments to the doctors) data is collected for each transaction performed. Since the introduction of Medicare in 1975, over a terabyte of data has been collected, and mostly never analysed.

This tremendous resource, that can tell quite a story about the changing health of Australians, started being used with data mining in the early 1990's. Over the years it has been used for identifying inappropriate provider practices, and for public fraud committed against Medicare.

The mining of the knowledge mine approach was successfully deployed to identify a particular type of fraud being committed by a group of patients against Medicare. The particular group ranked highly on a number of disjoint features, and in combination this lead the domain experts to follow up on their insurance claims, and eventually determine that they were fraudulent.

Another major piece of health data mining was made possible with the creation of the Queensland Linked Dataset (Williams, Vickers, Baxter, Hawkins, Kelman, Solon, He and Gu, 2002). This was the culmination of a project between CSIRO Data Mining, the Commonwealth Department of Health and Ageing and the Queensland Department of Health, bringing together a large collection of health care data for the purpose of data mining.

The Medicare Benefits Scheme (MBS), as mentioned above, together with the Pharmaceutical Benefits Scheme (PBS), provide universal health care insurance for Australians. These schemes cost several tens of billions of dollars each year and data relating to virtually every non-hospital medical activity in Australia since 1975 is recorded. But a significant gap in this data was information relating to hospitalisations of patients, which was a state rather than a commonwealth responsibility. This project brought together these datasets for the first time in Australia.

The resulting dataset consisted of 5 years of MBS and PBS transactions and 4 years of Queensland hospital admissions data for all patients in Queensland. The data was carefully de-identified so as to preserve patient privacy and confidentiality. The dataset consisted of records for 1.1 million individuals who were hospitalised in Queensland between 1995 and 1999, and there were 3 million hospital records in the data. For these patients there are 100 million MBS transactions and 60 million PBS transactions. For hospital records there are nearly 60 variables recorded, nearly 20 for MBS and 15 for PBS. Overall these data account for 500MB, 8GB, and 4GB respectively.

We have deployed this dataset in a number of data mining tasks, but the early work explored building multiple models and from these models exploring for significant, if rare, relationships between interactions with the medical system, and, for example, episodes in hospital. Indeed, this early work lead to initial discoveries in the dataset of relationships between multiple drug prescriptions and hospitalisation for specific conditions (Williams, Baxter, Kelman, Rainsford, He, Gu, Vickers and Hawkins, 2002).

4 Summary

In this paper we review the idea of modelling as one step along the path to data mining, where the aim is to gain insights into the world we are modelling, and with these insights, to take action to improve our business processes or our understanding of how things work. In particular, we have presented the genesis of the idea of building multiple models and illustrated applications where this has demonstrated useful outcomes.

References

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., eds (1996), *Advances in knowledge discovery and data mining*, The MIT Press, Cambridge, MA.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The elements of statistical learning: Data mining, inference, and prediction*, Springer Series in Statistics, Springer-Verlag, New York.

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine Learning* **1**(1), 81–106.

Quinlan, J. R. (1993), *C4.5: Programs for machine learning*, Morgan Kaufmann.

Walker, P. A., Cocks, K. D. and Young, M. D. (1985), 'Regionalising continental data sets', *Cartography* **14**(1), 66–73.

Williams, G., Baxter, R., Kelman, C., Rainsford, C., He, H., Gu, L., Vickers, D. and Hawkins, S. (2002), Estimating episodes of care using linked medical claims data, in *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence (AI02) Canberra, Lecture Notes in Artificial Intelligence*, Vol. 2557, Springer-Verlag, Canberra, Australia, pp. 660–671.

Williams, G. J. (1988), Combining decision trees: Initial results from the MIL algorithm, *in* J. S. Gero and R. B. Stanton, eds, *Artificial Intelligence Developments and Applications*, Elsevier Science Publishers B.V. (North-Holland), pp. 273–289.

Williams, G. J. (1991), Inducing and combining decision structures for expert systems, PhD thesis, Australian National University. http://togaware.redirectme.net/papers/gjwthesis.pdf.

Williams, G. J. and Huang, Z. (1997), Mining the knowledge mine: The Hot Spots methodology for mining large, real world databases, *in* A. Sattar, ed., *Advanced Topics in Artificial Intelligence*, Vol. 1342 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 340–348. http://togaware.redirectme.net/papers/ai97.pdf.

Williams, G., Vickers, D., Baxter, R., Hawkins, S., Kelman, C., Solon, R., He, H. and Gu, L. (2002), The queensland linked data set, Technical report, CSIRO Mathematical and Information Systems.