DATA MINING: SHOULD IT BE INCLUDED IN THE 'STATISTICS' CURRICULUM?

Siva Ganesh
Massey University
New Zealand

Teaching of statistics involves developing and adapting robust procedures for understanding statistical concepts, and for the management and analysis of statistical data. The field of statistics is constantly challenged problems that arise from science, industry and business. Traditionally, the statistics curriculum deals with data often collected to answer specific questions. However, in the modern 'information' age, vast amounts of data are collected, often automatically, with the advent of powerful computers. Data Mining is the process of extracting knowledge from large volumes of data. Since 'computation' plays a major role in this process, computer scientists have a significant claim over the ownership of data mining. Nevertheless, data mining techniques, in general, have a statistical base; and statisticians are beginning to show a significant interest in the area, including offering tertiary courses in 'statistical' data mining.

INTRODUCTION

Data Mining is the process of extracting knowledge from large volumes of data. In other words, data mining is the science (or art) of discovering unexpected patterns, valuable structures and interesting relationships in large and complex data, with an emphasis on large observational databases. The combination of fast computers and cheap storage makes it easier to extract "useful" information out of everything from supermarket buying patterns through Banking and Stock market trades to medical diagnostics and remote sensing. Data mining is widely used for achieving organisational goals and allowing investigators to go beyond simple data queries and reporting, in order to understand why things happen and to manipulate what happens in the future.

What distinguishes data mining from conventional statistical data analysis is that data mining is usually done for the purpose of 'secondary analysis' aimed at finding unsuspected relationships, perhaps, unrelated to the purposes for which the data were originally collected. In other words, data mining is very much an *inductive* exercise, as opposed to the traditional *hypothetico-deductive* approach of statistics. Data mining sits at the common frontiers of fields such as Information Systems (Database management & Data warehousing), Computer Science (Artificial Intelligence, Machine Learning & Pattern Recognition), and Statistics (Data Visualisation & Modelling).

The last decade or so have seen hundreds of computer software manufacturers jumping onto the data mining bandwagon. Major statistical software packages such as SAS, S-plus, SPSS and Statistica are being marketed as data mining tools rather than 'statistical'. With the sophistication and competitive edge of software packages, some would have you believe that, armed with large data and software tool, (correct) answers will come pouring out! Despite the above *myth*, the field of data mining is having a major impact in business, industry and science, and it affords research opportunities for new analytical and methodological developments.

DATA MINING VS STATISTICS

From a statistical perspective, data mining can be viewed as computer automated exploration and analysis of large and complex volumes of data. It may even be regarded as 'statistically intellectual'! The common feature of data mining and statistics is "learning from data".

Recall that, statistics, as taught in traditional curriculum, may be described as being characterised by data that are small, clean, static and randomly sampled, and often collected to answer a specific question. None of these apply in the data mining context. To a classical statistician, a data set with few thousand observations may be large, but to a 'data miner' (loosely defined as the user of data mining techniques) this is small! Modern databases often contain millions of records (megabytes, gigabytes or even terabytes in size), and data of such magnitude clearly put into context the futility of standard statistical techniques. This means, if all (or even a sample) of such large data is to be processed during an analysis, adaptive or sequential techniques

need to be utilised. These techniques have been the concern of computer scientists working in the areas of machine learning and pattern recognition.

However, it doesn't mean that statistics has no roll in the exploration of large volumes of data. After all, statistics is regarded as one of the most successful information science offering excellent designs for data collection and techniques for data analysis and modelling. Nevertheless, academic statistics has yet to enlarge its purview to take in all aspects of data collection, data manipulation and data interpretation. Unless this happens other more strongly computer based disciplines will increasingly offer that broader perspective, taking what they need from statistics, and consigning the statistics discipline to a supporting role.

Leaving aside the data collection aspects, the most commonly utilised data mining techniques are, Association rules (and Market basket analysis), Segmentation (and Self-organising maps), Memory-based reasoning, Decision trees, Neural Networks and Visualization. Although these techniques have been developed and promoted in the fields other than statistics, they heavily rely on the traditional statistical ideas and theories. For example, decision-tree approach is used for 'predictive modelling' such as regression modelling, and discrimination and classification problems which usually fall into the area of 'multivariate statistics'. Neural networks technique, on the other hand, may also be used for predictive modelling, but does not fall into any traditional statistical analytic processes. While association rule induction has its roots in theory of probability, segmentation merely means 'clustering' of data, again a branch of multivariate statistics. Recent developments in data mining technologies include the 'bundling techniques' (bagging, bumping and boosting approaches, developed by statisticians), genetic algorithms and text-mining.

It is possible to argue that, while there is great deal in common between data mining and statistics, the two have their own unique identities. We may also argue that, the peculiarities of the problems they each tackle and the nature and constraints of the methods they utilise could lead to a fruitful synergy. In fact, there are deep theoretical issues arising from data mining problems which would benefit from a statistical perspective and understanding. It appears that, data mining may be put within the context of *greater statistics* that can be defined, at least loosely, as "every thing related to *learning from data*". Greater statistics tends to be inclusive, eclectic with respect to methodology, closely related to other disciplines, and practised by many outside of traditional professional statistics and outside of academia.

The majority of data miners appear to have relatively little formal statistical expertise, hence, they sometimes make errors which trained statisticians would avoid as obvious. This implies that data miners must take on board statistical insights regarding the potential for spurious associations and issues of substance versus statistical significance, leading to a requirement for training data miners in statistics or statistics graduates in data mining. The approach needs to be practical and example-based, and some re-focussing of traditional statistics curriculum is desirable emphasising changes in data collection and analysis which have emerged in the past ten or so years.

Computer scientists have beaten the statisticians in offering data mining courses, since the advent of substantial improvement in computing power, in the 1990s. However, most if not all of these offerings have concentrated on the implementation of efficient algorithms from a machine learning point of view. Making sense of data by means of understanding the techniques utilised and improving such methodologies must also be a significant component of the learning process. The latter requires the knowledge of statistical modelling.

Although, several service-oriented and problem-driven statistical research methods courses are being offered at tertiary level, often via statistics units or consulting centres, courses on statistics oriented data mining has not been widely available on the menu. However, such courses are beginning to emerge at institutions such as Stanford and Massey Universities. These courses provide a broad statistical perspective of data mining at undergraduate level, and are aimed at students majoring in statistics and at those majoring in fields such as computer science, database management and business studies. Such courses must provide a broad coverage of techniques often categorised as *supervised* or *unsupervised*, and provide *descriptive* or *predictive* modelling. A typical prescription may take the form, "… an introduction to data mining applications including data preparation and data warehousing; query, association, market basket

and rule induction methods; prediction using regression, decision-trees and neural nets; clustering using hierarchical methods and self-organising maps; classification using trees and neural nets; a visual approach with real examples and case studies; use of leading data mining software tools; ...". An advanced course in 'statistical' data mining, say at postgraduate level, may include 'statistical underpinning' of (above) commonly used techniques plus a wide range of new developments such as genetic algorithms, text-mining, bagging, bumping and boosting algorithms, and Bayesian belief networks. Assessing the performance of data mining software tools should also be part of such an advanced course.

CONCLUSIONS

The challenge of collecting, exploring and disentangling complicated interrelationships among various characteristics of data is what makes 'statistical analysis' a rewarding activity. Data mining, although oversold, emphasises the new problems and opportunities that arise from data warehousing, and from the creation of new, often large, databases. It may be seen as an attempt to automate the processes by which statistical analysts often encounter unexpected information that is aside from the main purpose of the analysis. In the past ten years or so there have been large changes in the methodology for data analysis, taking advantage of advances in computer technology.

There is no doubt that there is mutual ignorance between statisticians and (computational) data miners. One part of the reason for this mutual ignorance lies in a conservativeness in statistics (perhaps a consequence of its mathematical past, inducing an inclination towards rigour at the cost of adventurousness) versus a risk taking attitude in computing (program it and see if it works, without worrying too much about provable properties of the algorithm). Another part lies in the modern statistical concern with models and the modern computer science concern with algorithms. Both concerns are perfectly sound, and are natural consequences of the way the disciplines have developed. There is now wide acceptance that progress in data mining will demand a merging of the insights of computing specialists with those of (theoretical and applied) statisticians.

It is something of an indictment of the statistical profession that so few statisticians have become involved in a deep way with data mining. Statisticians have a lot to teach data miners, while data miners have many fascinating and new problems which statisticians have not even begun to look at. There is the opportunity for an immensely rewarding synergy between statisticians and data miners (or computing specialists). It would be nice to think that researchers from both communities would come together to pool their distinct perspectives and approaches, to tackle the really important problems which are facing us in this modern data-rich world.

As statisticians, both as teachers and researchers, we need to confront the 'black box' hype common in the marketing of data mining, and take the challenge that data mining techniques pose for statistics including the designing of data warehouses, and assert the role statisticians as ‘information scientists’.

There is no doubt that data mining is 'statistically intellectual'. Statistical exploration of data mining has been active in recent years by means of research articles appearing in statistical as well as non-statistical journals and conference proceedings, textbooks written, workshops presented at conferences and other venues, and university courses offered at both undergraduate and postgraduate levels. Finally, turning attention to the question: Should data mining be included in the Statistics curriculum?; the answer inevitably should be an enthusiastic "yes". After all, *Data Mining* essentially is *Statistical Data Analysis*!

ACKNOWLEDGEMENTS

Many thoughts and statements in this article are courtesy of authors of texts, journal papers and conference proceedings cited below.

REFERENCES

Berry, J.A.M., & Linoff, G. (1997). *Data mining techniques: for marketing, sales and customer support*. New York: Wiley.

Berry, J.A.M., & Linoff, G. (2000). *Mastering data mining: the art and science of customer relationship management*. New York: Wiley.

Chatfield, C. (1997). Data mining. *Royal Statistical Society News, 25,* 1-2.

Friedman, J.H. (1998). Data mining and statistics: what's the connection. *29th Symposium on the Interface*.

Hand, D.J. (1998). Data mining: statistics and more? *American Statistician*, *52,* 112-118.

Hand, D.J. (1999). Data mining: new challenges for statisticians. *Proceedings of the ASC (Association for Survey Computing) International Conference*, 21-26.

Hand, D.J. (1999). Statistics and data mining: intersecting disciplines. *SIGKDD Explorations*, *1*, 16-19.

Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M. (2000). Data mining for fun and profit. *Statistical Science*, *15,* 111-131.

Hand, D.J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). Elements of statistical learning: data mining, inference and prediction. New York: Springer Verlag.

Jorgensen, M., & Gentleman, R. (1998). Data mining, *Chance, 11,* 34-42.

Maindonald, J.H. (1999). Data Mining from a Statistical Perspective, Statistics Consulting Unit,
ANU, Australia, available at: http://www.maths.anu.edu.au/~johnm/dm/dmpaper.html

Witten, I.H., & Frank, E. (1999). Data mining: practical machine learning tools. Morgan Kaufmann.