Perspectives on Data Mining

Niall Adams

Department of Mathematics, Imperial College London
n.adams@imperial.ac.uk

April 2009

Objectives

- Give an introductory overview of data mining (DM) (or Knowledge Discovery in Databases (KDD)) - sweeping generalisations
- Provide statistical perspectives
- Set context for subsequent talks
- Provide controversial talking points!

Aside: *Not sure DM has role in statistics!*

Data Mining

Hand, Mannila, Smyth: " DM is 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"

Witten and Frank: "DM is the extraction of implicit, previously unknown, and potentially useful information from data"

Data mining is a changing discipline. A few years ago data mining was secondary analysis of large data sets

- Large data sets - collected automatically/opportunistically
- Secondary analysis - data usually collected for some primary purpose (eg. credit scoring). Analysis not directed toward this purpose is *secondary*.

Then, the pitch of DM was that such secondary analysis could

- reveal previously unknown and valuable information;
- yield added-value from large data warehouses;

An important business application was analytic CRM.

Early accusations of *data dredging* – still a possible problem in pharma?

Presently, the discipline of data mining seems to be concerned with analysis of *large* data sets, for both primary and secondary analysis.

Is data mining nothing more than statistics writ large? (Or machine learning?)

I don't think so - *intersecting* disciplines. A key distinction is mode of *data collection*

- Statistics *primarily* concerned with many types of data (experimental, survey, observational), but perhaps emphasis on smaller data sets
- Data mining concerned with automatically gathered observational data over a complete population (e.g. company's customer base)

Tasks

Modern DM addresses diverse data analysis tasks

1. classification/regression
2. density estimation/bump hunting/anomaly detection
3. clustering/segmentation
4. association rules (interesting relations between categorical variables)
5. combinations of these (DM practitioners fearless in this regard)

(1) nearly always about *prediction*; in contrast to much traditional statistical work, where inference about (interpretable) model parameters is central.

Importance of inference not always recognised in DM.

Large Data

But what is a large data set? It is customary when talking about DM to give an impressive list of giant data sources

- Particle Physics: LHC ATLAS experiment (1PB per year - 100Mb per second for retained events)
- Remote Sensing: NASA Earth Observing System (50GB per day)
- Credit card processing: Visa Europe (1.8×10^{11} transactions per year)
- Web search: Google (20 PB per day)

Much *Scientific* DM different in delivery to *commercial* DM.

"*Large*" may be relative to processing resources.

Some of these are examples are streaming data - high frequency, non-stationary, hard to store. Dimitris Tasoulis will talk about this subject.

These data sources need considerable pre-processing to fit the familiar statistical "tabular array of data".

Preprocessing is required for DM, but this task is often integrated into the DM process (more on this later).

Combining large tables from the data warehouse is an important issue. Mike Page will explore issues of data combination for brand surveys.

Modalities

Can distinguish two modes of DM

- Global models for prediction, classification, clustering
- Pattern discovery and detection find unusual local structures (unsupervised or otherwise).

Global Models

Global models for regression, classification, etc

- linear and logistic regression, discriminant analysis
- k-NN, kernel methods
- neural networks, projection pursuit
- support vector machines, Gaussian processes
- Trees, bagging, boosting, model combination

Note terminology. Colin Shearer will talk about data mining *algorithms*, the type of tasks to which they are suited and how they might be combined.

In a statistical view, an (e.g.) MLP neural network is a (nonparametric) regression model. Estimation of the weights requires an algorithm (either optimisation or MCMC).

The large data set aspect is the reason for the DM emphasis on algorithms – often impractical to do with routine tools, so need efficient procedures. Great emphasis on *scalability*.

Thus, some data mining research is concerned with convergence properties of algorithms.

Model criticism is very important in statistics - much effort to determine model adequacy.

Model criticism in data mining often less critical - perhaps because nonlinear prediction often required. Multiple models will provide effectively equivalent predictions via very different mappings.

Aside: visualisation

"Eyeball your data" – common piece of statistical advice.

Large data sets can be more difficult. Standard data displays saturate with points. Richly structured population data can also cause problems.

Useful tools for DM visualisation provide data linkage, and "drill-down". However, the sheer size of data dictates the use of tools which make the analyst more distant from the data.

Anthony Atkinson will demonstrate a novel clustering methodology that generates exploratory graphics.

Pattern discovery and detection

In DM a pattern is an unusual structure or relationship in the data set. Examples

- Model based outliers (cf. regression diagnostic)
- low probability objects
- shapes in time series (cf. technical analysis)
- structures in sequences
- associations rules

This task has to use all the data - statistical trickery that might be ok for modelling counterproductive here.

Detection - find matches to a specified structure. Discovery - unsupervised (mostly)

A variety of applications addressed by PDD tools

- fraud detection (e.g. plastic card transactions)
- fault detection
- market basket analysis
- bioinformatics

Main problem of PDD - familiar to statisticians:
can always find lots of patterns

DM tends to produce great algorithms for finding patterns. Good example is APRIORI algorithm for association rules (though note that vanilla implementations are flawed - objects of different dimension compared with same threshold).

Need to provide a scored ranking of discovered patterns. Statistical arguments might help here - but large non-iid samples invalidate most conventional hypothesis testing.

Interesting statistical work in controlling False Discovery Rates very relevant here – methods to control expected proportion of incorrectly rejected null hypotheses.

Discovered patterns delivered to domain expert. May results in further searches – iterative process.

Just like statistical modelling....

Process

DM has evolved to the point where experts are starting to take a view of the whole DM process. Frans Coenen will explore this aspect.

In general, just as with statistics, domain expertise is crucial in DM. Deep understanding of the meaning of data, the collection process, and processes involved in the environment of the data, are essential.

Otherwise, the analyst returns results that are already known!

In addition to the DM process, Will Thompson will look at how the process can be simplified so domain experts (who are less capable with data analysis) can administer the process directly.

Data quality

Data collection in DM, while automated, is still prone to various types of error. Examples

- measurement device (noise, signal saturation, failure)
- incorrect data entry (accidental or otherwise, e.g. fraud)

DM should pick up structures related to data quality issues, but they are usually not of primary interest.

Database bias also a *serious* problem. If database refers to a specific population,it is biased toward this population. Worse yet, entry to the population may have required selection (cf. reject inference in credit scoring).

About time

Already mentioned streaming data, analysis of which has to be incremental and adaptive.

Databases collected over time. For many objects (especially related to people), likely to be changes over time: stochastic drift, policy, environment (consider a database of mortgage applications over the last year!), etc

Some applications have more pernicious changes. For example, in fraud detection problems, known that fraudsters attempt to adapt tactics to beat existing detection systems. Leads to arms race.

Darcy Norman will talk around this evolutionary aspect of DM with emphasis to surveys.

Ethics

Are there *ethical issues* related to DM?

In experimental studies involving humans, careful *controls* are put in place concerning the storage, use and disposal of data.

In business DM, data can be put to more *diverse* purposes, without constraint. Perhaps no problem?

However, government is keen to monitor us more closely for our own protection (it is claimed). Such data could be mined for inappropriate and possibly sinister purposes. Should there be ethical controls *?

Note that RSS has recently set up a working group to explore this issue.

(* Yes, I am a conspiracy theorist!)

One last point:

To support the panel discussion, in the following talks see if you can

- identify what DM ideas are based on, or basically, statistics.
- determine the role DM can make in analysing market surveys?

Hand, D.J., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining*. MIT press.

Witten I.H. and E. Frank (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Hand, D.J., Kelly, M.J., Blunt, G. and Adams, N.M. (2000) Data mining for fun and profit. *Statistical Science*, **15**(2), 111–126.