Frontiers in applications of data mining

David J. Hand
Imperial College London

Industry Day on Complex High-Dimensional Data
Isaac Newton Institute for Mathematical Sciences
Knowledge Transfer Network for Industrial Mathematics

19 May 2008

There are no frontiers in applications of data mining

.....

if a frontier is a boundary or limit beyond which the tools are not applied

Two Crows survey of business uses:

- Ad revenue forecasting
- Churn (turnover) management
- Claims processing
- Credit risk analysis
- Cross-marketing
- Customer profiling
- Customer retention
- Electronic commerce
- Exception reports
- Food-service menu analysis
- Fraud detection
- Government policy setting
- Hiring profiles
- Market basket analysis
- Medical management
- Member enrollment
- New product development
- Pharmaceutical research
- Process control
- Quality control
- Shelf management/store management
- Student recruiting and retention
- Targeted marketing
- Warranty analysis

Plus countless scientific applications

But if a frontier is *the limit of (current) knowledge*

then there are frontiers in applications of DM

There are frontiers because

- new problems arise all the time
- for which existing methods and tools won't work

These problems come from the interweaving of two strands:

- developments in computers

- new application areas springing up

1) Developments in computers

For example:

- vast new datasets
- creative and imaginative people seeing new possibilities of the computer power
- the web
- . . . . .

2) New application areas springing up

For example:

- biometrics, face recognition
- bioinformatics, genomics, proteomics, microarrays
- terrorism
- text mining
- . . . . .

Also old areas are changed unrecognisably by developments in computers:

- astronomy: out with photog plates and in with CCDs
- particle physics: terabytes of data
- post marketing surveillance in medicine
- epidemiology
- customer value management
- fraud detection
- . . . .

Where are the current gaps in our knowledge?
What are the important directions for future work?

or

Data mining comes of age

Data mining has evolved over the past 15 years

Data mining: its prehistory:

The statistical perspective

Fit many models to a limited data set

- trawling, fishing, snooping
- bound to find something interesting in the data
- but will it generalise?

Hence bad odour of DM amongst statisticians

The computer science perspective

DM is about analysing the databases
e.g. association analysis

No formal inference or generalisation to new cases

Implicit generalisation

without recognising its statistical risks

Basic inferential risks hidden by large data sets
- but problems if small subgroups
- and worse

Some major directions for the future

Business Intelligence

Business analytics

Customer value management

e.g. Revenue management

- pricing hotel rooms: predict highest prices people will pay without reducing occupancy
- airline tickets
- predicting likely outcome of blind auctions, so you can bid only just above the max

1) Experimentation

Classically DM has been about observational data

More recently there has been a move towards experimental data

The power of the computer enables products to be customised to each person

Example: Capital One

Capital One has over 6000 different credit card products worldwide)

But you don’t have to choose from 6000 products

By a process of elaborate experimentation of up to 90,000 experiments per year, CapOne has worked out what you will want

products, characteristics of products, prices, how the products are described, how they are marketed, account management, style of reacting to customers, etc etc

“We test every change in procedure, every job applicant, as well as every product offering, even down to the colour of our monthly statements.

...

We record every customer interaction, every card purchase. We collect data patiently from various lists, and then with the patience of a good scientist run experiment after experiment. For every action we have taken, we know what the reaction has been. If we have sent you a blue envelope or a pink one, we know which one you received and how you reacted to that. Every product that we develop is rolled out slowly and carefully. We track whether people buy something or not, and whether they ultimately use it.”

A specific example: their phone bill

"*We pay for incoming customer calls. But the calls were taking too long and our analysis demonstrated the problem was firmly with the way we managed the relationship, not our customers. Calls were constantly being transferred, all the time adding to our telephone costs and frustrating customers with the length of the transaction.*

Obvious solutions such as more training, and customer-selected menus were investigated but didn't work for us. Ultimately our technology team came up with the solution: predict the call reason and send the call automatically to the associate best skilled to deal with the problem."

Example: Web experiments

e.g. Google's *Adwords*, runs experiments on the web so you can see which ads are most effective.

Problems:

The constant danger of selectivity bias

Data miners have been slow to understand this

In experimentation
- beware use of different server fleets for control and treatment
- differences between when experiment run and now

The Red Queen effect

No experimental innovation in CVM stays ahead of the competition for long

But it is incremental

And if you don’t do it, you will soon be way behind

2) Anomaly detection

An area of special interest to me

e.g. fraud detection

- credit card
- telecoms
- benefit claimants
- money laundering
- etc

- fraud in scientific research

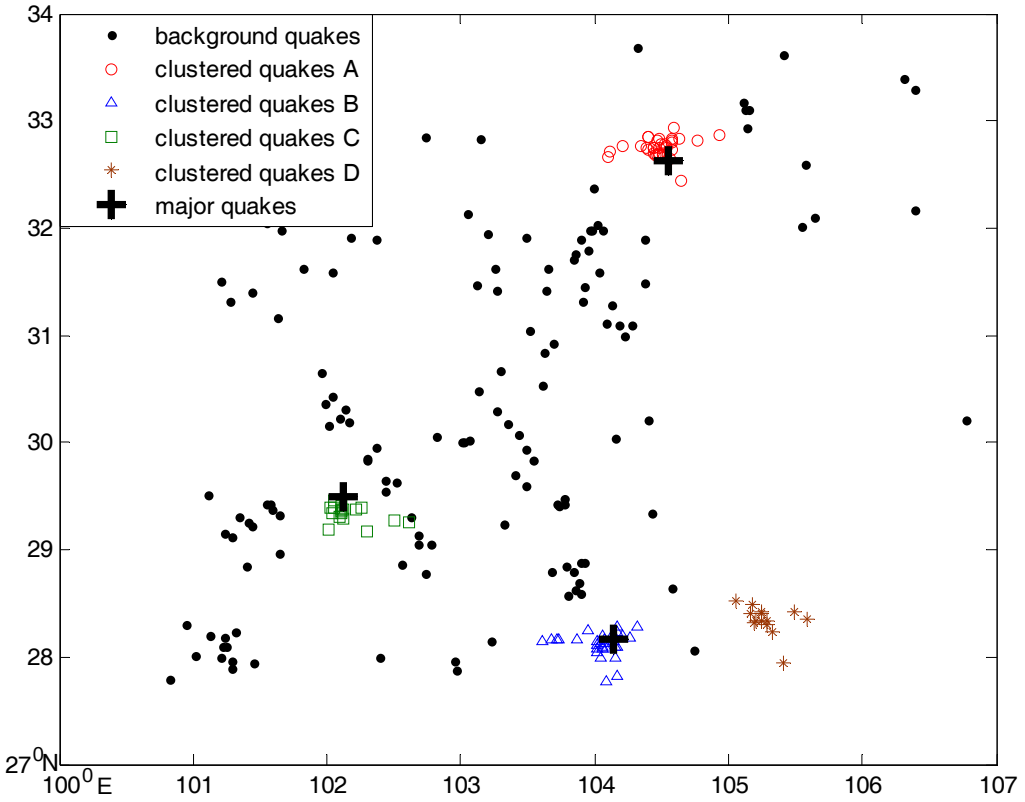
e.g. Hospital patient incidents (with James Bentham)

Patient was due to go to theatre , following an EPAU appointment for ? ectopic pregnancy . Patient was supposed to be on emergency list . Contacted theatre @ 16.30 to see what time patient would be going . Patient not listed for theatre . Reg on call informed

PHENOBARBITAL DOSE CHECKED ON DRUG CHART NO CD IN PHARMACY BOX DRUG BOXES CHAECK POM GIVEN AS PRESCRIBED TO PATIENT IMMEDIATELY CHECKED IN BNF AS STILL UNSURE STATES CD BUT PATIENT HAD ALREADY TAKEN TABLETS . 3 OPENED BOXES IN BAG FROM KINGFISHER S / N HT INFORMED

Bit arm during seizure

e.g. earthquakes (with Tao Pei)



e.g. detecting unusual astronomical objects (with Marc Henrion)

e.g. mine detection

e.g. detecting unusual events in ICU monitoring

e.g. detecting cheating students

e.g. detecting abnormal behaviour in crowds

. . . . . .

3) Display and visualisation?

Virtual reality for exploring databases

dynamic, interactive, exploration of data space - not simply looking at pictures on a screen

The next thing after Wii ?

4) Getting to grips with inference

Data miners have been slow about this

They often risk errors such as selectivity bias

5) Merging data sources

Data fusion

- official data (ONS, Treasury, local gov’t etc)
- transactional, corporate
- self supplied

Particular kinds of data

- text mining

- image mining

- (social) network analysis

- streaming data, dynamic data, tracking data

- the web

Data quality

Privacy

A final example: *credit scoring*

Every credit card transaction has 70-80 items of information associated with it (e.g. amount, nature of purchase, time, POS type, location, currency, ...)

100 transactions per year $\Rightarrow$ 7000-8000 dimensions

Summarise these down to descriptive characteristics of the use of that ***credit line***

But people have multiple credit lines: credit cards, cash cards, mortgages, car finance, overdrafts, season ticket loans, etc

Aggregate descriptors of credit lines $\Rightarrow$ description of ***customer***

Experian-Scorex provides over 400 *standardised aggregated attributes (STAGGs)*

Which need regular updating: regulations, new information,..

Build credit risk model for the customer using these

Needs regular updating: changes in economy, competition, technology, ...

END

www.imperial.ac.uk/people/d.j.hand