NAIC/NIPR E-REGULATION 2005

A Brief Introduction to Data Mining and Analysis

For

Insurance

Regulators

Brent Kabler
Research Supervisor
MO Dept of Ins
Brent.Kabler@insurance.mo.gov
(573) 526-2945

NAIC
making progress . . . together

Data Mining - Definition

Data mining inductively seeks to identify hidden patterns, sequences, and relationships in (typically) very large datasets.

Inductive analysis – analysis without theory-driven hypotheses

Automated algorithms – search functions, grouping and clustering methods, pattern recognition routines: automate search for patterns in very large datasets.

Data Mining - Critics

"Data mining is what IS staff call statistics."

"Data mining is a set of techniques designed to torture the data until it confesses."

"Data mining refers to the exaggerated claims of significance and/or forecasting precision generated by the selective reporting of results obtained when the structure of the model is determined experimentally by repeated applications of such procedures as regression to the same body of data…"

Data Mining – A Useful and Powerful Tool – with caveats

"There have always been a considerable number of people who busy themselves examining the last thousand numbers which have appeared on a roulette wheel, in search of some repeating pattern. Sadly enough, they have usually found it." -

Fred Schwed, 1940, *Where are the Customers' Yachts?*

Some statisticians argue that data mining fail to observe the strict rules of inference associated with the traditional hypothetical-deductive method, such as strict hypothesis testing and theory building, and can do more to lead researchers astray as to enlighten them.

Caveats (Cont)

In the prior example of finding patterns in a roulette wheel: a random process, such as a roulette wheel or a flip of a coin, can appear to produce meaningful or non-random outcomes – lead to false inferences (i.e. that the wheel is not random).

For example, it is very likely that 1 million spins of a roulette wheel will produce apparently non-random sequences, such as

"five red, five black," or "1, 2, 3, 4, 5, 6, 7, 8, 9, 10…..30"

Considerable substantive and methodological expertise is required to avoid mistakes – false inferences

Positives

My opinion: data mining can be, and has proven to be, an extremely valuable tool for analysts confronted with large volumes of data, in situations in which prior knowledge is low.

Advantages:

1. Can produce counterintuitive and unexpected results that more traditional methods might have missed
2. Largely automated, less labor intensive than traditional analysis

Data mining tools

Association –looking for correlations. Event or attribute A is typically present when event or attribute B is present

Sequence – temporal association. Event B usually follows Event A

Clustering or grouping – identifying classes that possess relatively unique characteristics; identifying "latent" or unobservable structures through observable patterns.

Prediction – assessing probability of a future outcome given known present conditions

A valuable regulatory tool

Data mining represents a potentially powerful tool that directly facilitates regulators' core mission:

1. **Predicting likelihood of non-compliance**
2. **Assessing likely consumer harm associated with non-compliance**
3. **Identifying appropriate interventions best suited to minimize risk of harm**
4. **Evaluation of the effectiveness of alternative intervention strategies**

Targeting Resources

Pareto Rule – A small number of causes are responsible for a great number of effects.

Examples – most companies generate greatest profits from a proportionately small percentage of customer bases (80/20 rule)

It is likely that the majority of consumer harm from market conduct issues come from a small percentage of firms, or a small number of particular industry practices.

Data mining can assist in identifying areas where regulatory intervention is most needed, and likely to have the most impact.

Examples of Regulatory Success

OSHA, a pilot project in Maine during 1990s developed sophisticated data analysis of the source of workplace injuries: "identify important concentrations of risk through systematic data analysis and then focus resources on those risks in an attempt to mitigate them."

Result: identified a significant concentration of risk associated with just 200 companies, representing just 1% of the market. Focused intervention on these high risk companies.

Regulatory Success –

Consumer Product Safety Council

Implemented extensive data collections, and procured analytical resources to analyze them. Results were counterintuitive, and suggested that resources were misdirected to less severe problems:

1. 411,689 Americans were injured by their beds, mattresses, or pillows seriously enough to warrant emergency room treatment, far outweighing the number injured by skateboards (48,186) or trampolines (82,722)

2. More people were hurt by sound recording equipment (38,956) than by chainsaws (29,684)

A Growing Practice

A recent GAO survey of 128 federal agencies found that 52 agencies are using or are planning to use data mining techniques to exploit vast amounts of data at their disposal.

Also, widespread in private sector – credit scoring

Problems

Lack of quality data

ETS and RIRS databases are subject to different reporting standards between the states, incomplete participation, subject to error, can't be integrated, and lacks elaborate coding mechanisms that would make the data suitable for statistical analysis.

These data could prove invaluable as a measure of regulatory outcomes, to assess differing targeting practices and interventions.

NAIC is currently devoting considerable attention to these issues.

Problems

Lack of sophisticated methodologies specific to market regulation.

Data mining and other analytical approaches could address this deficiency relatively quickly.

Example of Concrete Problem

Targeting Market Interventions

1. **Do current targeting techniques have a "hit rate" that is significantly better than one could do by chance?**
2. **Can more elaborate statistical profiling techniques and richer data perform better than current targeting techniques?**

Missouri Research

Missouri preliminary study to assess exam targeting in four states: can we predict which companies are most likely to exhibit serious market behavior problems?

Predicting Exam Outcomes

		Predicted Group	
Actual Group	***N***	***No Fine***	***Fine***
No Fine	122 100%	83 68.0%	39 32.0%
Fine	39 100%	9 23.1%	30 76.9%

Missouri Research

Percent of all cases correctly classified: 70.2%. If model held for future cases, could have increased the "hit rate" from 24% to 43%, or even higher depending on how the overall market compared with this subset.

It seems likely that percentage could probably increase significantly with better data and additional research: data mining can be a powerful technique to further this purpose.

Resources

Weka Toolkit – freeware downloadable software that implements many data mining algorithms.

See http://www.cs.waikato.ac.nz/~ml/weka/index.html

Also, see SAS and SPSS websites for two prominent statistical packages with recent additions of data mining modules.

SAS – www.sas.com

SPSS – www.spss.com