CONFERENCE PAPER

Drug Safety 2007; 30 (7): 621-622
0114-5916/07/0007-0621/$44.95/0



# Principles of Data Mining

*David J. Hand*

Department of Mathematics, Imperial College London, London, UK

## Abstract

Data mining is the discovery of interesting, unexpected or valuable structures in large datasets. As such, it has two rather different aspects. One of these concerns large-scale, 'global' structures, and the aim is to model the shapes, or features of the shapes, of distributions. The other concerns small-scale, 'local' structures, and the aim is to detect these anomalies and decide if they are real or chance occurrences. In the context of signal detection in the pharmaceutical sector, most interest lies in the second of the above two aspects; however, signal detection occurs relative to an assumed background model, therefore, some discussion of the first aspect is also necessary. This paper gives a lightning overview of data mining and its relation to statistics, with particular emphasis on tools for the detection of adverse drug reactions.

"*As with all data analysis, data mining is a process; it is rarely a case of one step being sufficient*."

A working definition of data mining is "the discovery of interesting, unexpected, or valuable structures in large datasets".[1]

Drivers behind data mining are of two kinds: scientific and commercial. Problems and intellectual challenges stem from both contexts; tools and software development are especially driven by commercial considerations.

Data mining is not a recent discipline. Previously it has been called, in a derogatory way, 'data dredging', 'trawling' and 'fishing through data'.

Modern data mining combines statistics with ideas, tools and methods from computer science, machine learning, database technology and other classical data analytical technologies. Classic statistics and data mining differ in several aspects. Statistics is a discipline originally developed throughout the twentieth century based on relatively small datasets. Computers only came on stream towards the end of that century allowing for the manipulation of larger datasets.

Big datasets often have a problem of selectivity bias. The classical statistical perspective on data mining is that with large enough datasets you are bound to find some structures in non-random samples of the population. The computer science perspective is that even if that is true, it is also certain that there is valuable information in there. Poor quality data does not deter data miners from tackling the problems they are presented with. The popularity of data mining is expected to raise awareness of the importance of collecting good quality data.

Traditionally, data mining is a secondary analysis over data collected for some other purpose. For example, in supermarket data, the information is collected to work out the bill the customer is charged. That stored data can then subsequently be submitted to analysis looking for customers' transaction patterns. However, the discipline is evolving and in areas like micro array data and proteomic research, large datasets are collected primarily for discovering patterns, relationships and structures.

Data mining can be divided into two main classes of tools: model building and pattern discovery. Model building is a high level global descriptive summary of datasets, which in modern statistics include: regression models, cluster decomposition and Bayesian networks. Models describe the overall shape of the data. Pattern discovery is what one does

when detecting signals of adverse drug reactions (ADRs). A 'pattern' is a local structure, in a possibly vast search space, describing data with an anomalously high density compared with that expected in a baseline model. Patterns are usually embedded in a mass of irrelevant data.[2]

Deciding whether a pattern is 'interesting' should be done in the operational context using, for example, knowledge from experts to understand exactly what is being described. Interesting patterns, irregularly high counts of adverse events, can be of two main types: artefacts of the data recording process (e.g. digit preference) or genuine discoveries about the underlying mechanism.

In the particular area of therapeutics, the ultimate aim of pattern discovery is to predict people's risk of being affected in the future. To support the inferential statement, it is fundamental to assess whether the patterns are real or due to chance.

The results from the model are influenced by the problem and the context, so generalised statements are difficult to make and require that 'the anomaly' be presented in relation to a baseline model.

Finally, the question 'Do the patterns matter?' is of fundamental importance. Their value in a commercial context, or their interest in a scientific environment and, in terms of ADRs, their public health impact, needs to be assessed.

## 1. Multiplicity in Data Mining

The classical approach to multiple statistical testing, the Bonferroni approach, should not be used in this context. An alternative approach for data mining is to control for the proportion of patterns flagged as significant that are actually false, the false discovery rate or, in other words, the probability that a data configuration that is flagged as suspicious does not actually represent any real underlying structure.

## 2. Data Quality

Data quality is a fundamental issue in pattern discovery. After all, small local clusters, as well as anomalous and unusual values, are exactly the sort of things produced by data contamination and measurement error.

When working with millions of data points, manual checking is impractical and data cleaning requires automatic procedures. Various solutions have been proposed. It is possible to develop cleaning and imputation procedures similar to those developed for surveys and censuses. When looking for patterns, there is a risk that automatic cleaning procedures will erase or smooth out the anomaly of interest.

Most apparent patterns in data are due to chance and data distortion, and many more are well known or uninteresting. Interestingness is a function of the context and the question asked. Total data accuracy is impossible and it is vital to handle 'messy' data carefully.

The increasing number of large data sets that are available provide an opportunity for some novel explorations. There is no doubt that data-mining tools provide rich opportunities for discovering unsuspected structures and patterns. However, this is not to say that it will be easy. Discovering interesting, valuable and novel structures necessarily involves an interplay between the data analyst and the data expert.

## Acknowledgements

No sources of funding were used to assist in the preparation of this paper. The author has no conflicts of interest that are directly relevant to the content of this paper.

## References

1. Hand DJ, Manila H, Smyth P. Principles of data mining. Cambridge (MA): The MIT Press, 2001
2. Hand DJ, Blunt G, Kelly MG, et al. Data mining for fun and profit. Stat Sci 2000; 15 (2): 111-31

Correspondence: Professor *David J. Hand*, Department of Mathematics, South Kensington Campus, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK.
E-mail: d.j.hand@imperial.ac.uk