

Data Mining and Science?

Knowledge discovery in science as opposed to business

Brian J Read

CLRC Rutherford Appleton Laboratory
Chilton, Didcot, Oxon OX11 0QX, UK

Brian.Read@rl.ac.uk

Keywords: Data Mining, Knowledge Discovery, Air Quality, Air Pollution

Abstract

The relatively new discipline of data mining is most often applied to extraction of useful knowledge from business data. However, it is also useful in some scientific applications where this more empirical approach complements traditional data analysis. The example of machine learning from air quality data illustrates this alternative.

1. Introduction

Data Mining is the essential ingredient in the more general process of Knowledge Discovery in Databases (KDD). The idea is that by automatically sifting through large quantities of data it should be possible to extract nuggets of knowledge.

Data mining has become fashionable, not just in computer science (journals & conferences), but particularly in business IT. (An example is its promotion by television advertising [1].) The emergence is due to the growth in data warehouses and the realisation that this mass of operational data has the potential to be exploited as an extension of *Business Intelligence*.

2. Why is Data Mining Different?

Data mining is more than just conventional data analysis. It uses traditional analysis tools (like statistics and graphics) plus those associated with artificial intelligence (such as rule induction and neural nets). It is all of these, but different. It is a distinctive approach or attitude to data analysis. The emphasis is not so much on extracting facts, but on generating hypotheses. The aim is more to yield questions rather than answers. Insights gained by data mining can then be verified by conventional analysis.

3. The Data Management Context

“Information Technology” was originally “Data Processing”. Computing in the past gave prominence to the processing algorithms - data were subservient. Typically, a program processed input data tapes (such as master and detail records) in batch to output a new data tape that incorporated the transactions. The structure of the data on the tapes reflected the requirements of the specific algorithm. It was the era of Jackson Structured Programming.

The concept of database broke away from this algorithm-centric view. Data assumed an existence independent of any programs. The data could be structured to reflect semantics

of relationships in the real world. One had successively hierarchical, network, relational and object data models in commercial database management systems, each motivated by the desire to model better the structure of actual entities and their relationships.

A database is extensional, storing many facts. Some information is intensional; that is, it manifests as rules. Some limited success was achieved with deductive databases that stored and manipulated rules, as for example in Prolog based systems. This encouraged *Expert Systems*. However, it was hard to achieve solid success. A main difficulty was the knowledge elicitation bottleneck: how to convert the thought processes of domain experts into formal rules in a computer.

Data mining offers a solution: automatic rule extraction. By searching through large amounts of data, one hopes to find sufficient instances of an association between data value occurrences to suggest a statistically significant rule. However, a domain expert is still needed to guide and evaluate the process and to apply the results.

4. Business Data Analysis

Popular commercial applications of data mining technology are, for example, in direct mail targeting, credit scoring, churn prediction, stock trading, fraud detection, and customer segmentation. It is closely allied to data warehousing in which large (gigabytes) corporate databases are constructed for decision support applications. Rather than relational databases with SQL, these are often multi-dimensional structures used for so-called *on-line analytical processing* (OLAP). Data mining is a step further from the directed questioning and reporting of OLAP in that the relevant results cannot be specified in advance.

5. Scientific Data Analysis

Rules generated by data mining are empirical - they are not physical laws. In most research in the sciences, one compares recorded data with a theory that is founded on an analytic expression of physical laws. The success or otherwise of the comparison is a test of the hypothesis of how nature works expressed as a mathematical formula. This might be something fundamental like an inverse square law. Alternatively, fitting a mathematical model to the data might determine physical parameters (such as a refractive index).

On the other hand, where there are no general theories, data mining techniques are valuable, especially where one has large quantities of data containing noisy patterns. This approach hopes to obtain a theoretical generalisation automatically from the data by means of induction, deriving empirical models and learning from examples. The resultant theory, while maybe not fundamental, can yield a good understanding of the physical process and can have great practical utility.

6. Scientific Applications

In a growing number of domains, the empirical or black box approach of data mining is good science. Three typical examples are:

1. Sequence analysis in bioinformatics [2]

Genetic data such as the nucleotide sequences in genomic DNA are digital. However, experimental data are inherently noisy, making the search for patterns and the matching of sub-sequences difficult. Machine learning algorithms such as artificial neural nets and hidden Markov chains are a very attractive way to tackle this computationally demanding problem.

2. Classification of astronomical objects [3]

The thousands of photographic plates that comprise a large survey of the night sky contain around a billion faint objects. Having measured the attributes of each object, the problem is to classify each object as a particular type of star or galaxy. Given the number of features to consider, as well as the huge number of objects, decision-tree learning algorithms have been found accurate and reliable for this task.

3. Medical decision support [4]

Patient records collected for diagnosis and prognosis include symptoms, bodily measurements and laboratory test results. Machine learning methods have been applied to a variety of medical domains to improve decision-making. Examples are the induction of rules for early diagnosis of rheumatic diseases and neural nets to recognise the clustered micro-calcifications in digitised mammograms that can lead to cancer.

The common technique is the use of data instances or cases to generate an empirical algorithm that makes sense to the scientist and that can be put to practical use for recognition or prediction.

7. Example of Predicting Air Quality

To illustrate the data mining approach, both advantages and disadvantages, this section describes its application to a prediction of urban air pollution.

7.1 Motivation

One needs an understanding of the behaviour of air pollution in order to predict it and then to guide any action to ameliorate it. Calculations with dynamical models are based on the relevant physics and chemistry.

An interesting research and development project pursuing this approach is DECAIR [5]. This concerns a generic system for exploiting existing urban air quality models by incorporating land use and cloud cover data from remote sensing satellite images.

To help with the design and validation of such models, a complementary approach is described here. It examines data on air quality empirically. Data mining and, in particular, machine learning techniques are employed with two main objectives:

1. to improve our understanding of the relevant factors and their relationships, including the possible discovery of non-obvious features in the data that may suggest better formulations of the physical models;
2. to induce models solely from the data so that dynamical simulations might be compared to them and that they may also have utility, offering (short term) predictive power.

7.2 Source Data

The investigation uses urban air quality measurements from the City of Cambridge (UK) [6]. These are especially useful since contemporary weather data from the same location are also available. The objectives are, for example, to look for and interpret possible correlations between each pollutant (NO, NO₂, NO_x, CO, O₃ and PM₁₀ particulates) and

- a) the other pollutants;

b) the weather (wind strength and direction, temperature, relative humidity and radiance); looking in particular for lags - that is, one attribute seeming to affect another with a delay of perhaps hours or of days.

7.3 Data Preparation

Before trying to apply machine learning and constructing a model, there are three quite important stages of data preparation. The data need to be *cleaned*, *explored* and *transformed*. In typical applications, this can be most of the overall effort involved.

a) Cleaning

Though not elaborated here, but commonly a major part of the KDD process is data cleaning. In this case, one is concerned with imposing consistent formats for dates and times, allowing for missing data, finding duplicated data, and weeding out bad data - the latter are not always obvious. The treatment of missing or erroneous data needs application dependent judgement.

b) Exploration

Another major preliminary stage is a thorough examination of the data to acquire familiarity and understanding. One starts with basic statistics - means, distributions, ranges, etc - aiming to acquire a feeling for data quality. Other techniques such as sorting, database queries and especially exploratory graphics help one gain confidence with the data.

c) Transformation

The third preparatory step is dataset sampling, summarisation, transformation and simplification. Working with only a sample of the full data, or applying a level of aggregation, may well yield insights and results that are quicker (if not even discernible at all) than with the complete data source. In addition, transforming the data by defining new variables to work with can be a crucial step. Thus one might, for instance, calculate ratios of observations, normalise them, or partition them into bins, bands or classes.

7.4 Initial Analysis

The initial analysis concentrated on the daily averages for the weather measurements and daily maxima of the pollutants. This simplifies the problem, the results providing a guide for a later full analysis. In addition, the peak values were further expressed as bands (e.g. “low”, “medium” and “high”). For example, ozone (O₃) values were encoded as

LOW	<50 ppb
MEDIUM	50-90 ppb
HIGH	>90 ppb

The bands relate to standards or targets set by the UK Expert Panel on Air Quality Standards (EPAQS) that the public can appreciate. (For ozone, the recommended limit is 50 ppb as an 8 hour running average.)

The data exploration and analysis is guided by domain knowledge and enhanced by it. Examination of the Cambridge air pollution data confirmed initial expectations:

- There is a daily cycle with peaks in the afternoon.

- Sundays have low pollution.
- An east wind (from industrial Europe) increases ozone levels.
- Sunlight on nitrogen dioxide (NO₂) produces ozone.
- Particulates (PM₁₀) come from vehicle exhausts.

Cambridge has little industry and within an urban environment traffic is the dominant pollution agent. Its effect depends on the local street topography so mesoscale dynamical models have restricted value.

7.5 Modelling

The two principal machine learning techniques used in this application are neural networks and the induction of decision trees. Expressing their predictions as band values makes the results of such models easier to understand.

a) Decision Trees

Applying the C5.0 algorithm to the data to generate a simple decision tree, one gets for ozone bands:

```
AirTemp <= 28.3 -> LOW
AirTemp > 28.3
  RelHum <= 58.1 -> HIGH
  RelHum > 58.1 -> MEDIUM
```

This suggests how the ozone concentration depends mainly on the air temperature and relative humidity. The same tree, expressed as a rule set is:

```
Rules for HIGH:
  Rule #1 for HIGH:
    if  AirTemp > 28.3
    and RelHum <= 58.1
    then -> HIGH

Rules for LOW:
  Rule #1 for LOW:
    if  AirTemp <= 28.3
    then -> LOW

Rules for MEDIUM:
  Rule #1 for MEDIUM:
    if  AirTemp > 28.3
    and RelHum > 58.1
    then -> MEDIUM

Default : -> LOW
```

In fact, the support for these rules is modest. The handicap is that there are too few instances of HIGH ozone days in the data. Reliable predictions would need something more elaborate, but this illustrates the idea.

b) Neural Networks

Alternatively, the daily data can be fitted with an artificial neural network to model the ozone band value. A first attempt yields:

Neural Network "O3band" architecture

Input Layer : 5 neurons
Hidden Layer #1 : 4 neurons
Output Layer : 4 neurons

Predicted Accuracy : 96%

Relative Importance of Inputs

AirTemp : 0.29
RelHum : 0.06
Rad : 0.04
Wet : 0.02
WindSpeed : 0.004

Again, this shows that air temperature is the dominant predictor. However, given the limited quantity of data summarised to daily values, it is not worth trying to refine the model network.

7.6 Software

The air quality data were analysed using the data mining software package *Clementine* [7] (originally from Integral Solutions Ltd. and now from SPSS Inc.) While this provides standard machine learning algorithms to generate models, its great virtue is the powerful visual environment it offers for data exploration. The figure shows an example of a data processing stream in *Clementine's* graphical interface. This ease of data exploration and modelling is crucial in allowing the domain expert to attack the problem and find applicable results.

7.7 Conclusions

Work so far supports the common experience in data mining that most of the effort is in data preparation and exploration. The data must be cleaned to allow for missing and bad measurements. Detailed examination leads to transforming the data into more effective forms. The modelling process is very iterative, using statistics and visualisation to guide strategy. The temporal dimension with its lagged correlations adds significantly to the search space for the most relevant parameters.

Investigation that is more extensive is needed to establish under what circumstances data mining might be as effective as dynamical modelling. (For instance, urban air quality varies greatly from street to street depending on buildings and traffic.) A feature of data mining is that it can "short circuit" the post-interpretation of the output of numerical simulations by directly predicting the probability of exceeding pollution thresholds. A drawback is the need for large datasets in order to provide enough high pollution episodes for reliable rule induction. More generally, data mining analysis is useful to provide a reference model in the validation of physically based simulation calculations.

8. Summary

Is data mining as useful in science as in commerce? Certainly, data mining in science has much in common with that for business data. One difference, though, is that there is a lot of existing scientific theory and knowledge. Hence, there is less chance of knowledge emerging purely from data. However, empirical results **can** be valuable in science (especially where it borders on engineering) as in suggesting causality relationships or for modelling complex phenomena.

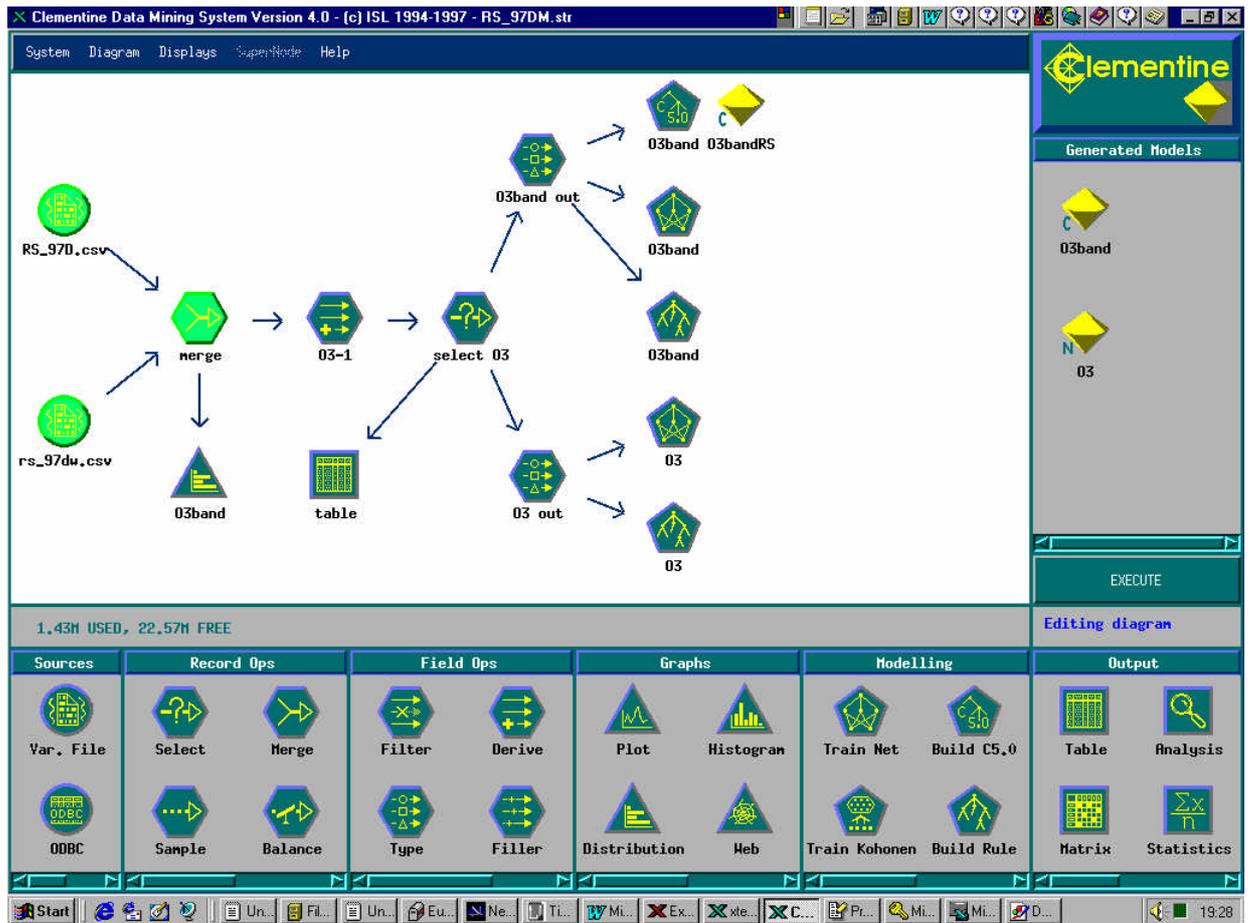
Another difference is that in commerce, rules are soft - sociological or cultural – and assume consistent behaviour. For example, the plausible myth that “30% of people who buy babies’ nappies also buy beer” is hardly fundamental, but one might profitably apply it as a selling tactic (until perhaps the fashion changes from beer to lager).

On the other hand, scientific rules or laws are, in principle, testable objectively. Any results from data mining techniques must sit within the existing domain knowledge. Hence, the involvement of a domain expert is crucial to the data mining process.

Naïve data mining often yields “obvious” results. The challenge is to incorporate rules known *a priori* into the empirical induction, remembering that the whole KDD process is exploratory and iterative.

References

- [1] <http://www.ibm.com/sfasp/locations/milan/index.html>
- [2] P Baldi and S Brunak, *Bioinformatics - The Machine Learning Approach*, MIT Press, 1998.
- [3] U M Fayyad, S G Djorgovski and N Weir, *Automating the Analysis and Cataloging of Sky Surveys*, in U M Fayyad *et al* (eds.), *Advances in Knowledge Discovery and Data Mining*, p471, AAIT Press and MIT Press, 1996.
- [4] N Lavraã, E Keravnou and B Zupan (eds.), *Intelligent Data Analysis in Medicine and Pharmacology*, Kluwer, 1997
- [5] Development of an earth observation data converter with application to air quality forecast (DECAIR): <http://www-air.inria.fr/decair/>
- [6] Cambridge City Council Air Quality Monitor:
<http://www.io-ltd.co.uk/ccc.html>
- [7] <http://www.spss.com/software/clementine/> and
<http://www.isl.co.uk/>



The Clementine user interface: The two source data files of weather and air pollutants measurements are merged, ozone bands selected and models (C5 decision tree and neural net) generated.