Statistics and Data Mining: Intersecting Disciplines

David J. Hand
Department of Mathematics
Imperial College
London, UK
+44-171-594-8521
d.j.hand@ic.ac.uk

ABSTRACT

Statistics and data mining have much in common, but they also have differences. The nature of the two disciplines is examined, with emphasis on their similarities and differences.

Keywords

Statistics, knowledge discovery.

1. INTRODUCTION

The two disciplines of *statistics* and *data mining* have common aims in that both are concerned with discovering structure in data. Indeed, so much do their aims overlap, that some people (perhaps, in the main, some statisticians) regard data mining as a subset of statistics. This is not a realistic assessment. Data mining also makes use of ideas, tools, and methods from other areas - especially computational areas such as database technology and machine learning - and is not heavily concerned with some areas in which statisticians are interested.

The commonality of aims between statistics and data mining has naturally caused some confusion. Indeed, it has even sometimes caused antipathy. Statistics has formal roots stretching back at least throughout this century, and the appearance of a new discipline, with new players, who purported to be solving problems that statisticians had previously considered part of their dominion, inevitably caused concern. The more so since the new discipline had an attractive name, almost calculated to arouse interest and curiosity. Contrast the promise latent in the term 'data mining' with the historical burden conveyed by the word 'statistics', a word originally coined to refer to 'matters of state' and which carries with it the emotional connotations of sifting through columns of tedious numbers. Of course, the fact that this historical image is far from the modern truth is neither here nor there. Furthermore, the new subject had particular relevance to commercial concerns (though it also had scientific and other applications).

The aim of this article is to put the two disciplines side by side, to note the places where they overlap and the places where they differ, as well as to draw attention to some of the difficulties associated with data mining. We might begin by observing that the term 'data mining' is not a new one to statisticians. Everitt [5], for example, defines it as 'a term generally used in a pejorative sense for the process of considering a large number of models including many which are "data-driven" in order to obtain a good fit'. Statisticians are thus careful about the ad hoc analysis of a set of data implied by the term data mining because they are aware that too intensive a search is likely to throw up apparent structures purely on the basis of chance. While this is true, it is also true that large bodies of data may well contain unsuspected but valuable (or interesting or useful) structures within them. It is this which has attracted the attention of others and it is this which is generally meant by data mining nowadays.

2. THE NATURE OF STATISTICS

It is pointless attempting a definition of a discipline as broad as statistics. All that such an attempt would achieve would be to attract disagreement. Instead I want to focus on a few of the properties of the discipline which stand in contrast to those of data mining.

One such difference is related to the remarks in the last paragraph of the previous section. This is that statistics as a discipline has a certain conservativeness. There is a tendency to avoid the ad hoc, and prefer the rigorous. Of course, this is not of itself bad: only through rigour can mistakes be avoided and truth be unearthed. However, it can be detrimental to discovery if it promotes an overcautious attitude. This conservativeness may derive from the perspective that statistics is a part of mathematics - a perspective with which I do not agree (see, for example, [15], [9], [14], [2], and the discussion in [3]). Although statistics clearly has mathematics at its base (as do physics and engineering, for example, and likewise neither is regarded as a 'part' of mathematics), it also has very strong links with each of the disciplines which generate the data to which statistical ideas are applied.

The mathematical background and the emphasis on rigour has encouraged a tendency to require proof that a proposed method will work prior to the use of that method, in contrast to the more experimental attitude which is at home in computer science and machine learning work. This has meant that sometimes researchers in those other disciplines, looking at the same problems as statisticians, have produced methods which apparently work, even if they cannot be (or have not yet been) proven to work. The statistical journals, in general, tend to avoid publishing ad hoc methods in favour of those which have been established, by relatively rigorous mathematics, to work. Data mining, being an offspring of several parents, has inherited the adventurous attitude of its machine learning progenitor. This does not mean that data mining practitioners do not value rigour, but merely implies that they are prepared to forgo it if this can be seen to give results.

It is the statistical literature which reveals the (perhaps exaggerated) emphasis of statistics on mathematical rigour. This literature also reveals its heavy emphasis on *inference*. Although there are subdisciplines within statistics whose concern is description, a glance in any general statistics text will demonstrate that a central concern is how to make statements about a population when one has observed only a sample. This is often also a concern of data mining. As we note below, a defining attribute of a data mining problem is that the data set is large. This means that often one will want, for reasons of practicability, to work only with a sample and yet make statements about the

larger data set from which the sample was drawn. However, data mining problems also often have available the entire population of data: details of the entire workforce of a corporation, of all customers in the database, of all transactions made last year. In such cases notions of significance testing lose their point: the observed value of the statistic (the mean value of all the year's transactions, for example) is the value of the parameter as well. This means that whereas statistical model building may make use of a sequence of probability statements in building a model (for example, that some proposed feature of a model is not significantly different from zero, and so may be dropped from the model) such statements are meaningless in a data mining context if the entire population is involved. In their place, one may simply use score functions: measures of the adequacy of description a model provides for the data. This fact, that one is often simply concerned with model fit rather than its generalisation ability, in many ways makes model search easier. For example, one can often make use of monotonicity properties of goodness-of-fit measures in model search algorithms (for example, in branch and bound methods), while these properties may be lost when one uses probabilistic statements about generalisability based on them.

A third feature of statistics which overlaps only partly with data mining problems is the central role played by the *model* in modern statistical work. It is perhaps unfortunate that the term 'model' means various rather different things. On the one hand statistical models may be based on some theory about the relationships between the variables analysed, while on the other they may be an atheoretical summary description of the data. A regression model for size of credit card transactions may include income as an independent variable because one believes it is likely to lead to larger transactions. This would be a theoretical model (albeit based on a very weak theory). In contrast, one might simply carry out a stepwise search over a set of potential explanatory variables, to obtain a model which yielded good predictive power, even if one had no idea why. (When data mining is aimed at producing a model, then it is normally concerned with this second kind of meaning.)

There are other ways to distinguish between statistical models, but I will not go into this here. See [10] for details. The point to which I want to draw attention here is simply that in modern statistics the model is king. The computation, the model selection criteria, and so on are all secondary, mere details in the task of building a good model. This is not so generally true in data mining. In data mining the algorithm plays a much more central role. (There are isolated exceptions in statistics where the algorithm is central. The Gifi school of nonlinear multivariate analysis is one such. For example, Gifi ([7], p34) says: 'In this book we adopt the point of view that, given some of the most common MVA [multivariate analysis] questions, it is possible to start either from the model or from the technique. As we have seen in Section 1.1 classical multivariate statistical analysis starts from the model, …. In many cases, however, the choice of the model is not at all obvious, choice of a conventional model is impossible, and computing optimum procedures is not feasible. In order to do something reasonable in these cases we start from the other end, with a class of techniques designed to answer the MVA questions, and postpone the choice of model and of optimality criterion.')

It is not surprising that algorithms are more central in data mining than in statistics, given that the mixed parentage of the former includes computer science and related disciplines. In any case the size of the data sets often means that traditional statistical algorithms are too slow for data mining problems and alternatives have to be devised. In particular, adaptive or sequential algorithms are often necessary, in which the data points are used one at a time to update estimates. Although some such algorithms have been developed within the statistical community, their more natural home is in machine learning (as the very word 'learning' suggests).

To many, the essence of data mining is the possibility of serendipitous discovery of unsuspected but valuable information. This means the process is essentially exploratory. This is in contrast to the rather optimistically styled 'confirmatory' analysis. (Optimistic because one can never actually confirm a theory, only provide supporting evidence or lack of disconfirming evidence.) Confirmatory analysis is concerned with model fitting - establishing that a proposed model does or does not provide a good explanation for the observed data. Thus much, perhaps most, statistical analysis addresses confirmatory analysis. However, exploratory data analysis is not new to statisticians, and this is perhaps another basis on which statisticians might consider that they already undertake data mining. All of this is true, but for the fact, again, that the data sets often encountered in data mining are huge by statistical standards. The well-established statistical tools may fail under such circumstances: a scatterplot of a million points may well simply yield a solid black display. ([11] contains examples.)

Given that the central aim of data mining is discovery, it is not concerned with those areas of statistics involving how best to collect the data in the first place so as to answer a specific question, such as experimental design and survey design. Data mining essentially assumes that the data have already been collected, and is concerned with how to discover its secrets.

3. THE NATURE OF DATA MINING

Since the roots of statistics predate the invention and development of the computer, it is hardly surprising that common statistical tools include many which may be applied by hand. On this basis, to many statisticians a data set of 1000 points will be regarded as large. But this 'large' is a far cry from the 350 million annual transactions handled by the UK's largest credit card company, or the 200 million daily long distance phone calls handled by AT & T. It is clear, in the face of such numbers, that techniques different from those that 'in principle may be applied by hand' are required. What it means is that computers (which are ultimately responsible for the possibility of the existence of such vast data sets) are essential for the manipulation and analysis of the data. It is no longer possible for analysts to interact directly with the data. Rather, the computer provides a necessary filter between the data and the analyst. This necessity is another reason for the extra emphasis on algorithms in data mining. Although necessary, the separation of analyst from data obviously carries associated risks. There is the real danger that unsuspected patterns may be distorting the analysis - a point to which we return below.

I do not want to give the impression that computers are not also an essential tool in modern statistics. They are, but for reasons other than the sheer size of the data sets. Data intensive analysis schemes such as bootstrap methods, randomisation tests, complex iterative estimation methods, and the large and complex models which can now be fitted are all only possible because of the

computer. The computer has allowed the scope of traditional statistical models to be extended dramatically, and also permitted the development of radically new tools.

I drew attention to the possibility of unsuspected patterns distorting the data. This is associated with the general issue of data quality. All conclusions of data analysis are subject to qualifications about the quality of the data. The computer acronym GIGO, standing for Garbage In, Garbage Out, applies just as much here as elsewhere, and data analysts, of whatever flavour, cannot perform miracles and extract gems from rubbish. With very large data sets, and in particular when one is seeking small and subtle patterns or departures from regularity, the problems are particularly acute. Deviations in the second decimal place can begin to matter when one is looking for one part in a million. With experience one can learn to be wary of the sorts of problems which are most common, but unfortunately there is an infinite number of ways in which things can go wrong.

Such problems may arise at two levels. The first is the micro level, that of the individual record. For example, particular attributes may be missing or misrecorded. I know of a case in which, unbeknown to those mining the data, missing values were recorded as 99 and were included in the analysis as genuine values. The second is the macro level: whether or not the overall data set being analysed has been distorted by some selection mechanism. Road accident statistics provide a nice example of this. The more serious accidents, those resulting in fatalities, are recorded with great accuracy, but the less serious ones, those resulting in minor or no injury, are not recorded so rigorously. Indeed a high proportion are not recorded at all. This gives a distorted impression - and could lead to mistaken conclusions.

Relatively little of statistics is concerned with real time analysis, though data mining problems often require this. For example, banking transactions happen every day, and one does not want to wait three months for an analysis alerting one to possible fraud. Associated issues arise from the fact that populations evolve over time. My research group has clear examples showing how the characteristics of applicants for bank loans change as time progresses and the competitive environment and economic climate fluctuate ([13]).

Up to this point we have described data analytic issues, showing how there are differences in emphasis between data mining and statistics, despite the considerable overlap. However, data miners must also contend with entirely non-statistical issues. One example is the problem of obtaining the data in the first place. Statisticians tend to view data as a convenient flat table, with cases cross-classified by variables, stored on the computer and simply waiting for analysis. This is fine if the data set is small enough to fit in the computer's memory, but in many data mining problems this is not possible. Worse, very large data sets are often dispersed across several machines. Perhaps the extreme of this is arises when analysing data from the World Wide Web, which may exist on many computers around the world. Problems of this kind make the very possibility of extracting a random sample questionable (let alone the possibility of analysing the 'complete data set', a concept which may not exist if the data are constantly evolving, as with telephone calls, for example).

When describing data mining techniques, I find it convenient to distinguish between two general classes of tools, according to whether they are aimed at *model building* or *pattern detection*. I have already noted the central role of the concept of a model in statistics. In model building one is trying to produce an overall summary of a set of data, to identify and describe the main features of the shape of the distribution. Examples of such 'global' models include a cluster analysis partition of a set of data, a regression model for prediction, and a tree-based classification rule. In contrast, in pattern detection, one is seeking to identify small (but nonetheless possibly important) departures from the norm, to detect unusual patterns of behaviour. Examples include sporadic waveforms in EEG traces, unusual spending patterns in credit card usage (for fraud detection), and objects with patterns of characteristics unlike any others. To many, it is this second exercise which is the essence of 'data mining' - an attempt to locate 'nuggets' of value amongst the dross. However, the first kind of exercise is just as important. Note that working with a sample is acceptable when one is concerned with global model building (one will be able to characterise the important features with a sample of a hundred thousand just as effectively as with a sample of ten million, although clearly this depends in part on the size of the features one wants to model). However, the same is not true of pattern detection. Here, selecting only a sample may discard just those few cases one had hoped to detect.

Although statistics is mainly concerned with analysing numerical data, the mixed parentage of data mining means that it also has to contend with other forms of data. In particular, logical data sometimes arise - for example, in searching for patterns composed of conjunctive and disjunctive combinations of elements. Likewise, higher order structures sometimes arise. That is, the elements of the analysis may be images, text chunks, speech signals, or even (as, for example, in meta-analysis) entire scientific studies.

4. DISCUSSION

Data mining is sometimes presented as a one-off exercise. This is a misconception. Rather, it should be perceived as an on-going process (even if the data set is fixed). One examines the data one way, interprets the results, looks more closely at the data from a related perspective, looks at them another way, and so on. The point is that, except in those very rare situations when one knows what sort of pattern is of interest, the essence of data mining is an attempt to discover the unexpected - and the unexpected, by its very nature, can arise in unexpected ways.

Related to the view of data mining as a process is the recognition of the novelty of the results. Many data mining results are only what one would expect - *in retrospect*. However, the fact that one can explain them does not detract from the value of the data mining exercise in unearthing them. Without this exercise, it is entirely possible that one would never have thought of them. Indeed, it is likely that only those structures for which one can retrospectively formulate a plausible explanation will be valuable. Those which still seem improbable, no matter how one twists and turns the likely causal mechanisms, may well not be real phenomena at all, but simply chance artifacts of the particular data at hand.

There is clear potential, opportunity, and indeed even excitement in data mining. The possibilities for making discoveries in large data sets certainly exist, and the number of very large data sets is growing daily. However, this promise should not conceal the risk from us. All real data sets (even those collected by entirely automatic processes) have the potential for error. Data sets

concerned with human beings (such as transaction and behaviour data) especially have such potential. This may well mean that most 'unexpected structures' discovered in the data are intrinsically uninteresting, being solely due to departures from the ideal process. (Of course, such structures may be interesting for other reasons: if the data have problems which might interfere with the purpose for which they were collected it is as well to know about them.) Associated with this is the deep issue of how to ensure (or at least provide support for the fact) that any observed patterns are 'real' in the sense that they reflect some underlying structure or relationship rather than merely how a particular data set, with a random component (for example, if it is a sample) happens to have fallen. Scoring methods may be relevant here, but more research, by statisticians and data miners is needed.

The data mining literature is now burgeoning. An important basic work is the edited text by Fayyad *et al* [6] and the breadth of current work is demonstrated by the range of topics and areas dealt with in the proceedings of the *International Conference on Knowledge Discovery and Data Mining* series (the two most recent proceedings being [12] and [1]) and the journal *Data Mining and Knowledge Discovery*. Papers discussing the relationship between statistics and data analysis include [8], [4], and [10].

5. REFERENCES

[1] Agrawal R., Stolorz P., and Piatetsky-Shapiro G. (eds.) (1998) *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press.

[2] Bailey R.A. (1998) *Journal of the Royal Statistical Society, Series D, The Statistician*, **47**, 261-271.

[3] Discussion (1998) Discussion on the papers on 'Statistics and mathematics'. *Journal of the Royal Statistical Society, Series D, The Statistician*, **47**, 273-290.

[4] Elder J, IV, and Pregibon D. (1996) A statistical perspective on knowledge discovery in databases. In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press. 83-113

[5] Everitt B.S. (1998) *The Cambridge Dictionary of Statistics*. Cambridge: Cambridge University Press.

[6] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) (1996) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press.

[7] Gifi A. (1990) *Nonlinear Multivariate Analysis*. Chichester: Wiley.

[8] Glymour C., Madigan D., Pregibon D., and Smyth P. (1996) Statistical inference and data mining. *Communications of the ACM*, **39**, 35-41.

[9] Hand D.J. (1998a) Breaking misconceptions - statistics and its relationship to mathematics. *Journal of the Royal Statistical Society, Series D, The Statistician*, **47**, 245-250.

[10] Hand D.J., (1998b) Data mining: statistics and more? *The American Statistician*, **52**, 112-118.

[11] Hand D.J., Mannila H., and Smyth P. (forthcoming) *Principles of Data Mining*, MIT Press.

[12] Heckerman D., Mannila H., Pregibon D., and Uthurusamy R. (eds.) (1997) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press.

[13] Kelly M.G., Hand D.J. and Adams N.M. (1999) The impact of changing populations on classifier performance. Open University Department of Statistics Technical Report.

[14] Senn S. (1998) Mathematics: governess or handmaiden. *Journal of the Royal Statistical Society, Series D, The Statistician*, **47**, 251-259.

[15] Sprent P. (1998) Statistics and mathematics - trouble at the interface? *Journal of the Royal Statistical Society, Series D, The Statistician*, **47**, 239-244.

About the Author:

REFERENCES David J. Hand is the Professor of Statistics in the Department of Mathematics at Imperial College in London. Previously, he was the Professor of Statistics and Head of the Department of Statistics at the Open University. He is the founding editor and continuing editor-in-chief of the journal *Statistics and Computing*, and is former editor of the *Journal of the Royal Statistical Society, Series C*. Professor Hand has published over 100 research papers and fifteen books, including *Construction and Assessment of Classification Rules*, *Practical Longitudinal Data Analysis*, and *Statistics in Finance*. His research interests include data mining, classification methods, and the interface between statistics and computing. He has consulted broadly, including in the areas of medicine, psychology, and finance