Data Mining Methodological Weaknesses and Suggested Fixes

John Maindonald

Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 0200, Australia.
Email: john.maindonald@anu.edu.au

Abstract

Predictive accuracy claims should give explicit descriptions of the steps followed, with access to the code used. This allows referees and readers to check for common traps, and to repeat the same steps on other data. Feature selection and/or model selection and/or tuning must be independent of the test data. For use of cross-validation, such steps must be repeated at each fold. Even then, such accuracy assessments have the limitation that the target population, to which results will be applied, is commonly different from the source population. Commonly, it is shifted forward in time, and it may differ in other respects also.

A consequence of source/target differences is that highly sophisticated modeling may be pointless or even counter-productive. At best, model effects in the target population may be broadly similar. Investigation of the pattern of changes over time is required. Such studies are unusual in the data mining literature, in part because relevant data have not been available.

Several recent investigations are noted that shed interesting light on the comparison between observational and experimental studies, with particular relevance when there is an interest in giving parameter estimates a causal interpretation.

Data mining activity would benefit from wider co-operation in the development and deployment of computing tools, and from better integration of those tools into the publication process.

Keywords: Data mining, statistics, predictive accuracy, target population, observational data, selection bias, reject inference, comparison of algorithms.

1 Introduction

It is now widely though not universally understood that training set accuracy, derived by using the training data for testing also, can be grossly optimistic. Cross-validation or a bootstrap approach is therefore preferred. Where however feature selection and/or model tuning are a component of the model fitting process, care is required to avoid subtler versions of the bias in the training set accuracy measure. For an unbiased assessment, any feature selection and/or model tuning must be repeated at each fold of the cross-validation.

Other important issues relate to the distinction between observational and experimental data, to differences between source and target population, to the stability and interpretability of model parameters, to the comparison of algorithms, to the implications of new technology for the publication process, and to improving cooperation in the development of new tools. The remainder of this section will make preliminary comments on the first two of these issues.

1.1 Observational versus experimental data

Data mining typically uses for prediction or other inferences data that are observational rather than experimental. This introduces hazards that, for data from carefully planned and conducted experiments, are largely absent.

Thus, in a recent study that used a large US car accident database (Meyer and Finney, 2005), the interest is in a model parameter that accounts for the effect of airbag availability on accident mortality. Many factors apart from airbag availability contribute to the outcome. If other factors are ignored, airbags seem to give large benefits. After accounting for the effects of seatbelts and various other factors, benefits appear small or non-existent.

Compare this with a notional experimental study, where cars would be randomly assigned to have airbags fitted, or not, and where other factors (use of a seatbelt, speed of impact, etc) should on average contribute only statistical noise.

Where experimental studies fail, it is typically for one of a small number of reasons, commonly failure of the randomization process. Other possibilities are that experimental subjects (or units) may be untypical of the population to which results will be applied, or that the experiment may answer a question that is different (perhaps subtly different) from the question of interest.

By contrast, it is hardly possible to give a simple and reasonably complete summary of the different ways in which observational studies may fail. See however Rosenbaum (2002). The range and variety of different types of observational study is almost unlimited.

In some business and industrial problems, it may be reasonable to limit attention to a small number of well understood causal factors. The assumption that this is the case should not be made lightly. At the very least, issues such as will be discussed below severely limit the range of problems where relatively automated data mining approaches can be trusted to give useful results. At worst they may make any inferences from available data, however carefully teased out, perilous.

1.2 Accuracy varies with target population

A recurring theme will be that accuracy assessments are specific to a particular target population. A sim-

ple example, using the `Pima.tr` data set that is included with the *MASS* package for R, will illustrate the point. There are seven columns of features that may help explain diabetic status, recorded as `No` or `Yes`. Use of Breiman's random forest algorithm, as implemented in R, gave a classification rule thus:

```
> Pima.rf <- randomForest(type ~ .,
+     data = Pima.tr)
```

The confusion matrix is

```
> Pima.rf$confusion
```

	No	Yes	class.error
No	111	21	0.159
Yes	36	32	0.529

The error rate is estimated as 28%; this is calculated as $(132 \times 0.159 + 68 \times 0.529)/ (132 + 68)$.

If however predictive accuracy is calculated for a population in which the proportions of `No` and `Yes` are equal, the expected error rate changes to $0.5 \times 0.159 + 0.5 \times 0.529 = 34\%$.

Thus use of a balanced sample in cross-validation accuracy assessment may make a large difference to the assessment. Any report of an overall measure should be accompanied by details of the population composition that is assumed. Better still, accompany the report with details of the confusion matrix.

As an aside, note that balanced samples are a poor use of data, unless the relative proportions, perhaps weighted according to misclassification costs, are equal in the target population. Even then, it is better to use prior weights to train a model that is optimal for the relative frequencies and costs in the target population. See Ripley (1996) for the relevant Bayesian decision theory for classification models.

In the case discussed, the difference was in the relative frequencies in two categories. Differences between source and target population are common, and rarely so straightforwardly handled. This discussion will be pursued in the next section, noting also implications for the comparison of algorithms.

2 Honest use of cross-validation

A cross-validation estimate of accuracy, or an estimate obtained from a random split of the data into two parts, is often the best that is available. In default of anything better, it provides an upper bound on the accuracy that can be expected for predictions for the target population. Such estimates are in any case commonly used when algorithms are compared. Unless done correctly, such comparisons are meaningless, and potentially misleading.

Where there is feature selection and/or significant model selection and/or significant model tuning, the following steps are involved:

1. Select features and/or select model and/or tune the model

2. Fit the model that is in due course selected as "best".

Both these steps must be repeated at each fold of the cross-validation process, using what are the training data for that fold.

Consider the following experiment, leading in due course to Figure 1:

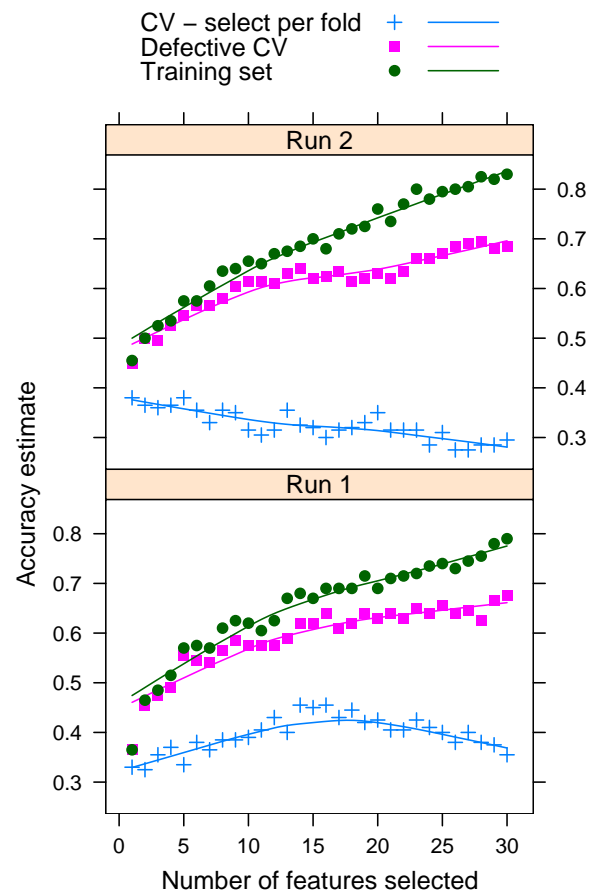- Set up a matrix X whose n rows are observations, and whose $p >> n$ columns are features.

Figure 1: The plots repeat discriminant rule calculations, using different sets of random normal data, to compare different accuracy measures. Each data set had 200 observations, divided approximately equally between three groups, and 1000 features, The training measure (filled circle) is severely biased. Tenfold cross-validation, with features selected using all the data (filled square), has a less severe bias that is nevertheless unacceptable. An acceptable measure (points shown as +) requires re-selection of features at each fold of the cross-validation.
(The code used to create these plots is available from the web page noted in Section 5.)

- Fill the elements with random normal data, independent (though this is not essential to the demonstration) between elements.

- Generate a categorical variable y, with $k = 3$ categories.

Now do the following:

Defective cross-validation – select once
For each of $m = 2, 3 \ldots, 30$, repeat the following steps

1. Using an analysis of variance F-statistic as the criterion, choose the m features that best separate the rows of X into the k categories.

2. Do a cross-validation of a discriminant analysis with the chosen m features, and determine predictive accuracy.

Cross-validation – select at each fold
At each fold of the cross-validation, there is a local set of training data, on which the model is trained, with remaining data providing the local

test set. Modify the above procedure so that, at each fold, the local training set is used for feature selection (and any model tuning), prior to fitting the model and making predictions for the local test set.

Figure 1 shows, for two different X-matrices where the data are "white noise", and with 200 observations that were divided approximately equally between three groups, the resulting assessments of predictive accuracy, plotted against number of features. Notice the different patterns of change in the correct cross-validated error rate (points shown as +), different between the two sets of random data. Similar results to the points shown as + will be obtained if the overall model is assessed on a completely separate set of randomly chosen test data, i.e. on a new matrix Z of the same dimensions as X and filled in the same way with random normal data.

The trap may seem obvious, but a number of authors, including well-known names, have been caught by it. See Ambroise and McLachlan (2002); Zhu et al. (2006). In statistics as in mathematics, plausible notions and methods that have not been validated with proper care are commonly found to be wrong or misguided. Simulation, with random data as in Figure 1, can often give a useful wake-up call.

Suppose however that there is a genuine signal in X, which will now be treated as a sample from the source population. Suppose also that there are systematic differences from the target population, from which we have a sample Z. Assume now that the cross-validation is done correctly. Close tuning to fit the source population will, at some point, lead to the degrading of performance on the target population. Although he does not make the point in this way, this is implicit in the comments in Hand (2006). I will comment further in the next section.

3 Source and target population

Differences in the relative frequencies of different groups in the data are relatively simple to handle. More generally, accuracy assessments that are based on cross-validation, or that are from a random split of the total data into training and test data, are realistic only if the processes that generated the data are the same processes that will apply when results are put to practical use. The source population, from which the data have been sampled, must be closely identical to the target population to which results will be applied. Or, to use different language, the model that describes the processes that have generated the data must also be an adequate description for the processes that will apply when model predictions are used in practice.

A clear and unequivocal near identity of source and target population is, with observational data, unusual. In a business context data that are derived from the past year's activities, or from the past several years, may be the basis for changes in business practice that will affect future years. This point will be taken up below.

3.1 Reject inference

A common further complication is noted in Hand (2006). In assessing credit risk, the sample is distorted as a sample of the potential population of applicants. The true outcome is known only for those applicants who were given credit, yet the inference is required for all applicants, leading to the term *reject inference*. The methodology discussed in Heckman (1979) can in principle address this problem, but requires assumptions that cannot always be checked.

Even if predictions are stable under temporal or other changes that affect the population of interest, it does not follow that model parameters (e.g., regression coefficients) will be stable. There are subtle and complex issues that affect the interpretation of such coefficients when they are derived from observational data.

3.2 Source and Target – A Taxonomy

The following are typical of situations that may occur in practice. This is a slightly modified version of the classification of the range of possibilities that appears in Maindonald & Braun (2006):

1. The data used to develop the model are, to a close approximation, a random sample from the population to which predictions will be applied. If this can be assumed, a simple use of a resampling method will give an accuracy estimate that is unbiased with respect to the population that is the target for predictions.

2. Test data are available that are from the target population, with a sampling mechanism that reflects the intended use of the model. The test data can then be used to derive a realistic estimate of predictive accuracy.

3. The sampling mechanism for the target data differs from the mechanism that yielded the data in 1, or yielded the test data in 2. However, there is a model that predicts how predictive accuracy will change with the change in sampling mechanism. Thus, in the `Pima.tr` and `Pima.te` datasets that were the basis for the calculations in Subsection 1.2, the predictive accuracy is a function of the relative number that are `Yes`.

4. The connection between the population from which the data have been sampled and the target population may be weak or tenuous. It may be so tenuous that a confident prediction of the score function for the target population is impossible. In other words, a realistic test set and associated sampling mechanism may not be available. An informed guess may be the best that is available.

These four possibilities are not completely distinct; they overlap at the boundaries. The distinction between them, such as it is, is however a good starting point for making a judgment on the closeness of the connection between the source and target populations.

Item 3 covers a wide range of possibilities. One simple possibility was discussed earlier, where the remedy is to give groups within the data weights that reflect the relative frequencies and perhaps costs in the target population, rather than those in the source population. The forest cover dataset from the UCI Machine Learning Database (Newman et al, 1998) is interesting in this connection. The relative numbers of the seven different forest cover types change systematically as a window of perhaps 5000 from the 500,000 observations moves through the data. This presumably reflects systematic changes in geographical location – information not included in the data. As the window moves, there are large changes in local predictive accuracy, i.e., using the data within the window as target. This is the case both for a model fitted to the data as a whole, and when the model is fitted to the data locally. While the confusion matrix from the local model changes somewhat between successive windows, the effect on predictive accuracy is of minor consequence relative to that of changes in the proportions of the different cover types.

For reject inference problems, approaches such as in Heckman (1979) are available, but rely strongly on specific modeling assumptions. Validation is accordingly both more necessary and more difficult.

In another common circumstance, there may for example be very extensive data on house prices in two suburbs of a large city. For predicting house prices in another suburb we have what is effectively a sample of two, and must further assume that this can be treated as a random sample. The assumption that errors are independently and identically distributed across the total sample of prices, as in most software that is explicitly aimed at data mining, will lead to optimistic assessments of predictive accuracy. (For other examples see Maindonald, 2003). Similar issues arise with data that are a time series. Again it is necessary, formally or informally, to account for the "error" part of the model.

3.3   Changes with time

Consider again the use of the current year's data to make assessments that will affect next year's business activity. If data from several previous years are available, then it makes sense to run the analysis separately for each of those years, and check for consistency between the different sets of results. If such data are not available, then there may be no good basis for judging the relevance to the subsequent year's business activity. Even where there does seem to be some modest level of consistency over time, this consistency may be placed in jeopardy by changes in external circumstances. Economic shocks – a dramatic increase in oil prices or an economic recession – may depending on the specific context create discontinuities that invalidate or place in doubt assessments that are based on past data.

Where the source and target populations are separated in time, model refinement is readily taken to a point where improved accuracy for the source population leads to reduced accuracy for the target population. What is signal at one point in time may with the passage of time become bias. Under-fitting, relative to estimates of accuracy that are based on a random split between test and training data or on cross-validation, may lead to improved accuracy for the data that matter.

Hand (2006) has two interesting examples that relate to credit scoring. Hand's Figure 4 shows the error rate over a $3\frac{1}{2}$ period, from a classifier built at the start of the period. The error rate drops to almost zero after 8 months, then after a year is back at the initial level, then rises to be $2\frac{1}{2}$ times the initial level by the end of the period. In a second graph (Hand's Figure 5), the performance of a tree-based classifier is compared with that of a linear discriminant function, over customers 1 to 60,000, using odd-numbered customers in the range 1 to 4999 for training. At the beginning of the series, the misclassification cost was around 0.1 less for the tree-based classifier. This difference had reduced to perhaps 0.05 by the end of the series, with the performance of the linear discriminant staying fairly constant at a cost of around 0.225. Other issues concern inevitable changes in the composition of the target population, arbitrariness and drift in the class definition ("concept drift"), and vagueness in the assignment of relative costs.

3.4   Implications for comparing algorithms

In practice then, there is not the identity between source and target populations that the standard comparisons of algorithms assume. Published comparisons of algorithms are at best broad indications of
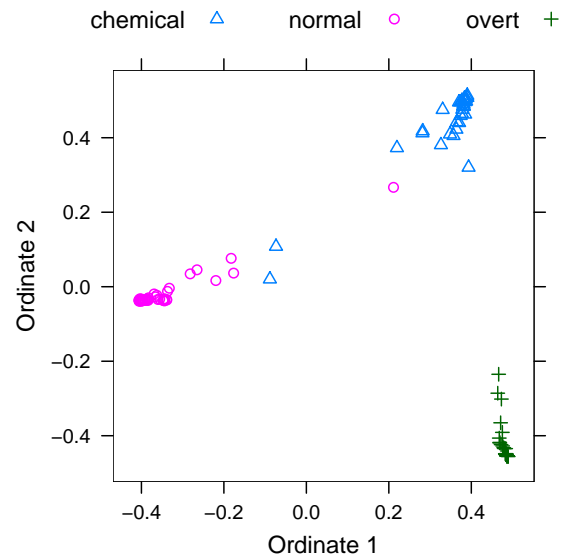

Figure 2: Calculations used the `diabetes` dataset, included in R's *mclust* package. Proximities r_{ij}, calculated for any pair (i, j) of points as the proportion of trees in which they appear at the same terminal node, were derived from use of the `randomForest` function with the `diabetes` dataset. Distances 1-r_{ij} were then used with R's `cmdscale` metric scaling function, yielding a two-dimensional representation.

performance, even once careful attention has been given to advice such as appears on the web site Keogh et al (2006) or in the papers Elkan (2001); Salzberg (1997).

Hand (2006) makes the further point that, in comparisons of different algorithms, users who are more expert with a particular method will have a bias towards obtaining their best results with that method. This, and the inevitability that performance is to an extent data dependent, are yet further reasons for treating published comparisons of algorithms as, at best, broad indications of performance.

A quick check through the UCI Machine Learning repository did not reveal any sizeable data sets that are well suited to studying changes in algorithm performance over time. This is clearly a serious gap in the resources that are currently available for testing and evaluating algorithms. In a number of cases (e.g., the email spam database), it would be highly interesting to have comparable time-stamped data from a period of several years. The newly established UCR Time Series Data Mining Archive (Keogh, 2006) is therefore very welcome. Many practical classification problems have a time-dependent component that should not be ignored.

3.5   Low-dimensional representations

It is helpful to characterize, where possible, conditions under which one or other algorithm is likely to perform well. A low-dimensional representation that shows the separation of groups in supervised classification, or of clusters in a cluster analysis, may give valuable insight. It may indicate gross features of the distribution of data, and give visual clues that highlight differences between one algorithm and another. Where the main effect of tweaking an algorithm is to change which observations are misclassified, the plot will show this. Insight is often more helpful than a 0.1% gain in the cross-validation estimate of predictive accuracy.

Figure 2 was obtained by using the "proximities" from a `randomForest` discriminant rule to derive a low-dimensional representation. The figure legend gives the details. The plot identifies three points where the class labels seem in doubt. Plots of discriminant scores from R's `lda` (*MASS* package) or of the ordination scores from `svm` (*e1071* package) with default parameters, do not show the same clear separation. Why?

4 The Interpretation of Model Parameters

An unequivocal interpretation is usually impossible when there are multiple explanatory features that might be included in the model, perhaps measured with different accuracies. Typically, it is necessary to appeal to other supporting sources of information. Parameter estimates, even if highly significant statistically, cannot necessarily be taken at face value. I will note several instructive case studies. Even if not highly typical of the problems tackled by data miners, they have lessons of which data miners should be aware.

A referee has made the point that whether observational studies are effective in any particular circumstance will depend on the importance, subtlety and nature of the inference. Where the interpretation of parameter(s) is an issue, and there are multiple explanatory features, there is inevitable subtlety.

4.1 Smoking and lung cancer

Notwithstanding the strength of the link between smoking and lung cancer, with papers making the link appearing in the late 1920s, it was not until the 1950s that the connection was placed beyond reasonable doubt. Only when it was clear that multiple independent lines of evidence all pointed in the same direction were the most tenacious critics silenced. See Freedman (1990) for further commentary on the history, on the statistical issues, and for a number of other examples.

Effects that are much smaller than in the connection between smoking and lung cancer may be hard or impossible to tease out, especially if several factors are involved and no one factor strongly predominates.

4.2 Hormone Replacement Therapy

The health effects of hormone replacement therapy (HRT) have been a subject for extensive investigation over a long period of time, with extensive data now available both from observational and from experimental studies. This large collection of studies offers data analysts a unique opportunity to compare results between experimental and observational studies.

Case-control studies, as in Varas-Lorenzo et al (2000), are among the best-regarded of the observational studies. In these "cases", i.e., individuals who have the disease, are first identified. These are then matched with disease-free "controls", chosen to be as similar as possible in all respects except perhaps use of the therapy, in this case HRT. The hope is that over subjects as a whole, disease status will be the same as if the assignment to receive HRT had been done randomly. Almost inevitably the matching is not completely effective, and regression must be used to adjust for remaining differences. If an important explanatory variable is omitted from the adjustment (perhaps, as suggested below, childhood socio-economic status), conclusions may be fatally compromised.

Contrast such studies with experimental studies such as are reported in Rossouw et al (2002). As they enrol, participants are randomly assigned either to HRT or to a placebo, perhaps subject to restrictions that maintain a numeric balance between treatment and control groups. Strict adherence to randomization protocols ensures the identity of the treatment and control populations.

A large meta-analysis of the "best" quality cohort and other observational studies (Stampfer and Colditz, 1991) found a relative reduction in coronary heart disease (CHD) risk of 50% from any use of HRT. Where population based studies gave more or less definitive results, they agreed broadly in their conclusions, to the extent that Stampfer and Colditz could claim

> Overall, the bulk of the evidence strongly supports a protective effect of estrogens that is unlikely to be explained by confounding factors.

Broad agreement across the different studies does not however mean that the estimates are correct. Few would now defend Stampfer and Colditz's conclusion, for reasons that will now be discussed.

The experimental results showed that, far from reducing CHD risk, risk was increased. One large randomized controlled trial (Rossouw et al, 2002) found that HRT use increased CHD hazard by a factor of 1.29 (95% CI 1.02–1.63), after 5 years of follow-up.

This was particularly anomalous because the results of the observational studies have been consistent with the results of randomized trials for other outcomes – breast cancer (increased risk for the combined oestrogen/progesterone HRT; for a 50-year old from 11 in 1000 to maybe 15 in 1000), colon cancer (reduced risk), hip fracture (reduced risk, but diet, exercise and other drugs can achieve the same or better results) and stroke (increased risk; for a 50-year old from 4 in 1000 to 6 in 1000). See again Swan et al (2006) and e.g., Rossouw et al (2002) for further details and references.

Lawlor et al (2004) discuss why there is agreement for most outcomes, but not for CHD. Childhood socio-economic indicators are known to be important as predictors of CHD, independently of adult socio-economic status (SES), behavioural and physiological risk factors. This is not true for the other outcomes considered. Additionally, the use of HRT is "strongly socially patterned"; those with low childhood SES less commonly used HRT. Consider now individuals with low childhood SES, but high adult SES. Their low childhood SES is associated with low use of HRT and consequent lowered risk of CHD. In the analysis, the only adjustment is for their high adult SES. The benefit derived from non-use of HRT is wrongly ascribed, in the analysis and its associated interpretation, to their high adult SES.

If this account is correct, it highlights the importance of accounting properly for socio-economic effects. When studying an outcome of interest from an observational public health study, it is important to ask whether the simpler type of model that can account for breast cancer risk is adequate, or whether the situation that pertains to CHD risk is more likely.

4.3 Other examples and references

Do airbags save lives? The available US data are not encouraging, if analyzed with care. See Meyer and Finney (2005), and articles in a forthcoming issue of *Chance* that will continue the discussion, now with corrected data. The data, although extensive, suffer from a version of the reject inference problem –

they are from accidents that are sufficiently serious that at least one car was towed from the scene. Estimates of the effect of airbags change spectacularly with changes in the other factors that are incorporated into the model.

Leavitt and Dubner (2005) have a number of examples that illustrate the care that must be taken in bringing together multiple sources of evidence, in order to reach conclusions that seem reasonably secure. Their account of the reasons for the reduction in US crime rates in the 1990s, which I find convincing, has attracted huge controversy.

Rosenbaum (2002) teases out practical implications of the use of observational rather than experimental data, using for illustration a number of interesting examples. The insights in this important book have received less attention than they deserve in the statistical community, and scant attention in the data mining community. The brief final chapter, entitled "Some Strategic Issues", makes a number of specific suggestions that merit attention.

5 Re-engineering the publication process

Advances in computer technology allow and demand large changes in the reporting of data, in data analysis, in the total content of publications, and in access to the separate components of the content (Maindonald, 2005). Data mining is among the areas where the potential for change and innovation is greatest. Code and data that are used in papers should be available as a matter of course, preferably as part of a compendium (Gentleman and Lang, 2004) such as will now be discussed, which the reader can readily process through a computer program to create a version of the final paper. The compendium should include or give access to

- the text of the paper

- the data on which it is based, and

- the code used for analysis and for generation of tables and graphs.

The notion of reading a paper is substantially enlarged, to include interaction with the processes involved in moving from data to analysis to published paper.

The `noweb` literate programming syntax (Johnson and Johnson, 1997) is a suitable vehicle for the implementation of these ideas. My experience has been with the implementation in the R system (R Core Development Team, updated regularly). The function `Sweave` (Leisch, 2006) provides a flexible framework for mixing text and R code in an enhanced LaTeX document for automatic report generation. When processed through R's `Sweave` function, markup instructions that surround the R code chunks determine which chunks, and which of the output generated by the code, will be included in the final LaTeX document. Output may include tables and figures.

Gentleman and Lang (2004) argue strongly for the provision of an Sweave type compendium for any paper that presents results of genomic analyses, as a matter of standard practice. Users can then know with certainty the steps that have been followed. Benefits include the opening to scrutiny of any biases in the analysis protocols, and a ready ability to reproduce results and test their sensitivity to analysis choices.

The arguments are surely equally cogent for journals and conferences that publish data mining papers. Provision of `Sweave` type features is a reasonable requirement for any language that is intended for scientific use. A present serious limitation of `Sweave` is

that code that appears in the LaTeX document has comments stripped from it.

An `Sweave` version of this present paper is available from the web page `http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.Rnw`. The R packages *hddplot*, *mclust* (which includes the `diabetes` dataset), *randomForest* and *xtable* must be installed.

The file **ausdm06-jm.Rnw**, when processed through R's `Sweave` function, yields a LaTeX file and associated graphics file from which this present paper can be generated.

6 Final Comments

The issues that I have raised are all in a sense statistical, though not always receiving the attention that they deserve in statistics courses. Here, I will comment on the different traditions of data mining and of statistics, and on the large area of interest that they have in common.

6.1 Different traditions of data analysis

Statistics started as a discipline that had a strong practical orientation. The small number of statistics departments that predated World War II likewise had a strong practical orientation. The three decades that followed World War II saw the widespread establishment of statistics departments, now with a strong theoretical focus. Many of the teachers saw statistics as primarily a mathematical discipline. Over the intervening years, the teaching of statistics has slowly matured to pay more attention to applications, though this change still has some way to go. Over this same period, theory and computing have moved in synergy to bring been huge advances both in theory and in practical computing tools.

The R system (R Core Development Team, updated regularly) is an outstanding product of the new synergy between theory, computing and practice. It demonstrates what is possible when experts co-operate widely across national boundaries. It promises larger achievements yet, more in tune with modern ideas of computer systems.

Where academic statistics took mathematics and a range of practical demands as its points of departure, data mining has taken computing, algorithmics and data bases as its points of departure. It has thrown out a variety of challenges to statistics – challenges which I think valuable for the future development of statistics. A specific challenge is to make statistical methodology available to those who, while bypassing much of the mathematical theory, wish to have access to the fruits of that theory. Simulation is for this as well as for other purposes highly important, especially as it sometimes offers a way ahead in cases where the theory is intractable.

Data miners face, likewise, challenges from the statistical tradition, beyond those raised earlier in this paper. Among these is the challenge to marshal computing skills and tools effectively. Standalone tools are typically deficient in the data manipulation and graphical abilities needed to use them effectively, require the mastering of their own idiosyncratic user interfaces, and do not penetrate widely into the communities where they might find use. Contrast this with the use of R or another such system, as a framework for developing new software, and as an interface into the end product. Many data miners, sensitive to the benefits of such a common interface, are already using and contributing to R. Those who do not wish to go this route have the challenge of finding or developing a system that can equal or better R: in the

expertise that has contributed to its development, in its range of abilities, in the trustworthiness of its output, in its cohesion, in its linkages into other systems, in automated checks that impose minimal standards of consistency across the system as a whole, in the use of the internet to give access to R and to associated resources, in its relative ease of use, and in the wide extent of its user community.

Hard-won insights from both the practical and theoretical streams of statistical development require the attention of data miners. I attach high importance to issues that I have noted in this paper, centering around source and target population, realistic assessment of predictive accuracy, the interpretation of model parameters, and the insights that may be derived by comparing results from observational studies with results from experimental studies.

6.2 The training of data miners

To what extent is understanding of statistical issues, such as I have canvassed, required for effective data mining? Relatively automated use of data mining tools will give better results for some applications than for others. Without however some sense of what issues are important, how will the data analyst know the difference? Anyone who expects to make data mining a substantial part of their work will do well to take time and effort to get on top of the issues that I have canvassed. They can all be understood without recourse to advanced mathematics. To be effective, these points must be reinforced by exposure to, and understanding of, the practical data analysis contexts in which they arise.

6.3 Course materials

Course materials for a course component that includes a statistically focused commentary on data mining are available from my website (Maindonald, 2006). Data that the laboratory exercises explore include two substantial datasets that were mentioned above – the US forest cover data and the US car accident data.

Acknowledgement

References

Ambroise, C and McLachlan, G J, 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, **99**:6262–6266.

Elkan, C 2001. Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000 (postscript) (pdf). In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01), pp. 426-431.
`http://www-cse.ucsd.edu/users/elkan/kddcoil.pdf`

Freedman, D 1979. From association to causation: some remarks on the history of statistics. *Statistical Science* **14**:243-258.

Gentleman, R and Lang, D T 2004. Statistical Analyses and Reproducible Research. Bioconductor Project Working Papers. Working Paper 2.
`http://www.bepress.com/bioconductor/paper2`

Hand, D J 2006. Classifier technology and the illusion of progress. *Statistical Science* **21**:1-14.

Heckman, J J 1979. Sample selection bias as a specification factor. *Econometrics* **47**:153-161.

Johnson, A L and Johnson, B C. 1997. Literate programming using noweb. *Linux Journal*, 64-69, October 1997.
`http://members.shaw.ca/andrew-johnson/noweb_lj.pdf`

Keogh, E 2006. The UCR Time Series Data Mining Archive `http://www.cs.ucr.edu/~eamonn/TSDMA/index.html` Riverside CA. University of California – Computer Science & Engineering Department.

Keogh, E, Xi, X, Wei, L & Ratanamahatana, C A 2006. The UCR Time Series Classification/Clustering Homepage
`www.cs.ucr.edu/~eamonn/time_series_data/`

Lawlor, D A, Davey Smith, D F and Ebrahim, S 2004. Commentary: The hormone replacement – coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology* **33**:464–467.

Lawlor, D A, Davey Smith, D F and Ebrahim, S 2004. Socioeconomic position and Hormone Replacement Therapy use: explaining the discrepancy in evidence from observational and randomized controlled trials. *American Journal of Public Health* **94**:2149–2154.

Leavitt, S D and Dubner, S J, 2005. *Freakonomics. A Rogue Economist Explores the Hidden Side of Everything*. William Morrow.

Leisch F 2006. Sweave User Manual.
`http://www.ci.tuwien.ac.at/~leisch/Sweave`.

Maindonald, J H 2005. Data, science and new computing technology. *New Zealand Journal of Science* **62**:126-128.

Maindonald, J H 2006. Statistical Commentary on Data Mining: Course Materials.
`http://www.maths.anu.edu.au/~johnm/courses/dm/`

Maindonald, J H, 2003. The role of models in predictive validation. Invited Paper. 54th session of the ISI, Berlin, 2003.
`http://www.maths.anu.edu.au/~johnm/dm/isi2003-models.pdf`

Maindonald, J H and Braun, W J, 2nd edn, 2006, in press. *Data Analysis and Graphics Using R – An Example-Based Approach.* Cambridge University Press.
`http://wwwmaths.anu.edu.au/~johnm/r-book.html`

Maindonald, J H and Burden, C J 2005. Selection bias in plots of microarray or other data that have been sampled from a high-dimensional space. In R May and A J Roberts, eds, *Proceedings of 12th Computational Techniques and Applications Conference CTAC-2004*, volume 46, pp. C59–C74.
`http://anziamj.austms.org.au/V46/CTAC2004/Main`.

Meyer, M C and Finney, T 2005. Who wants airbags? *Chance* **18**:3-16.
`http://www.stat.uga.edu/~mmeyer/airbags.htm`
See also `http://wwwmaths.anu.edu.au/~johnm/datasets/airbags/`

Newman, D J, Hettich, S, Blake, C L and Merz, C J 1998. UCI Repository of machine learning databases `http://www.ics.uci.edu/~mlearn/ MLRepository.html` Irvine, CA: University of California, Department of Information and Computer Science.

R Core Development Team. *An Introduction to* R. `http://cran.r-project.org`

Ripley, B D 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rosenbaum, P R 2002. *Observational Studies*, 2nd edn. Springer-Verlag.

Writing Group for the Women's Health Initiative Investigators 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **288**:321-333.

Salzberg, S L 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining. and Knowledge Discovery* **1**:317–327.

Stampfer M J and Colditz G A 1991. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Preventative Medicine* **20**:47–63.

Swan, N, Fry, R, McPherson, A, Trevena, L and Davis, S 2006. Hormone Replacement Therapy - Part Two - Radio National Summer. `http://www.abc.gov.au/rn/healthreport/ stories/2006/1530042.htm#`

Varas-Lorenzo C, Garcia-Rodriguez L A, Perez-Gutthann S, Duque-Oliart A 2000. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation* **101**:2572-2578.

Zhu, X, Ambroise, C, and McLachlan, G J 2006. Selection bias in working with the top genes in supervised classification of tissue samples. *Statistical Methodology* **3**:29–41.