Defining the goals to optimise data mining performance

Mark G. Kelly, David J. Hand*, and Niall M. Adams

Department of Statistics
The Open University
Milton Keynes
MK7 6AA, UK
{m.g.kelly, d.j.hand, n.adams}@open.ac.uk

Abstract

In many data mining problems the definition of what structures in the database are to be regarded as interesting or valuable is given only loosely. Typically this is regarded as a source of ambiguity and imprecision. However, we propose taking advantage of the looseness of the definition by choosing a particular definition which optimises some additional criterion. We illustrate using a consumer credit data set, where the definition of what constitutes a bad risk customer is somewhat arbitrary. Instead of adopting the common strategy of freely choosing some definition, we choose that which optimises predictability. That is, we choose to define our classes on the grounds that they are the ones amongst those which can be most accurately predicted.

Introduction

In many supervised classification problems the classes are defined in a rather arbitrary way. For example, although student grades may be based on performance in an examination, the choice of the thresholds defining the grades is seldom based on a clear-cut and objective criterion. Similarly, in many medical domains (such as psychiatry) the diseases are defined in terms of symptom patterns rather than in organic terms and there is often some ambiguity in deciding if a symptom is present. A third example - the one we use below - is that, in a bank's decision about whether or not to grant credit, the definition of a 'good' or 'bad' customer will often be based on some rather arbitrary indicators of behaviour. The arbitrariness in situations like these can arise from several causes. One is change over time: the very subject matter that students learn evolves over time (at least, one hopes so), so that the grades need to evolve to reflect this, and as economic circumstances alter so what is to be regarded as a bad risk may change as different loans become economical or uneconomical. Another is that the variables used to define the classes may be proxies for the variables which are really of interest, with the correlation being imperfect. A third is simple ambiguity: perhaps there is no sound reason for preferring one definition to another.

Typically the problem of potential ambiguity in the class definitions in supervised classification problems has been ignored in the past. When it has been recognised it has been regarded as a problem to be tackled through the use of new models or extensions of existing ones. For example, Kelly and Hand (1997) and Hand and Kelly (1998) describe models in which the definition of the classes need only be given at the time at which the classification is required, and not when the model is constructed and Hand, Li, and Adams (1998) discuss models in which the classes are defined in terms of multiple underlying variables. Here, however, we take a complementary approach, adopting the viewpoint that, if there is ambiguity in the definition then perhaps one can capitalise on it to produce more effective models. This is very much in the spirit of data mining: instead of trying to build models which will predict a given type of structure, we let the data determine the structure which is to be predicted, and then build models for that. Although we restrict ourselves to supervised classification, the general principles apply much more widely.

Our basic premise is that in (at least some) situations where there is ambiguity in the definitions of the classes, it is often difficult to argue that the chosen definition is in any sense superior to alternative definitions slightly perturbed from the chosen one. To illustrate we consider a set of data taken from a portfolio of current account holders with a major bank. A 'bad' account in a particular month is defined as one for which, in the month in question, (a) the excess amount overdrawn above the nominal limit is greater than £500; or (b) this excess is greater than £100 and the maximum balance over the course of the month in question is less than £0; or (c) total credit turnover in the month is less than 10% of the month's end balance. A 'good' account is defined as the complement of this. This definition is similar to that used in practice, but slightly modified so as to preserve commercial confidentiality (but we shall, for simplicity, refer to this definition as 'the bank's original definition'). Now it is difficult to believe that the choice of £500 in (a) has been based on any formal reasoning or data analysis. That is, it would seem difficult to defend the view that a threshold of £500 in defining the classes is any more appropriate than a definition of £490, or

*Author to whom correspondence should be addressed.

one of £510, for example. Indeed, the roundness of 500 raises one's suspicions that this has been chosen on grounds other than mathematical or financial. On the other hand, clearly the thresholds used in the definition have some substantive import - somehow they convey the sort of thing that the bank wants to mean by 'bad'. The same applies to the other thresholds used in the definition.

Discussions with bankers confirm that there is considerable arbitrariness in the choice of thresholds in the above. To a large extent, they are chosen on grounds of practical convenience. However, in the days of computer technology, it is arguable that a threshold of £500 is any more convenient than a threshold of (say) £492. Clearly the definition of good and bad in this example is meant to serve as a crude measurement (merely binary) scale for the underlying concept of 'goodness/badness'. But it does more than merely measure the concept. It also defines it. That is, the definition above is both an operational definition of what is meant by 'bad' and a way of determining if a customer is bad.

Given that the definitions of the classes are not precise, and that alternative definitions slightly perturbed from those described above may equally legitimately be adopted, we can talk of a 'region of legitimate definitions'. This is the region defined by varying the above thresholds such that any definition within it would be equally acceptable for the purpose to hand.

In general, the existence of such a region of legitimate definitions in a problem is a source of imprecision and inaccuracy. The standard approach to restricting the region - and the approach implicitly illustrated above - is to *impose* some definition, thus removing the ambiguity by fiat. The bank simply adopts the thresholds in the example above. This is all very well, and it certainly achieves the purpose. However, the existence of the region of legitimate definitions also invites one to restrict it by forcing the problem formulation to satisfy some additional criterion which is useful. That is, perhaps we can take advantage of the ambiguity, so that the ultimate solution is satisfying on other grounds as well.

The choice of the additional criterion must depend on the problem. In the example we are using to illustrate the ideas, we adopt as the criterion a measure commonly used in credit scoring applications (Hand and Henley, 1997), namely the *Gini* coefficient. This is a measure of performance of supervised classification rules, taking values between 0 and 1 (at least, it takes the value 0 for a classifier which is no better than chance), larger values indicating better performance. It is defined as twice the area between the curve and the diagonal in a Lorentz diagram, and is formally equivalent to the Mann-Whitney-Wilcoxon two sample test statistic of the hypothesis that two distributions are identical (Hand, 1997). The Gini coefficient is not an ideal measure for assessing classification rules (see, for example, Adams and Hand, 1998), but it is a very widely used one, especially in credit scoring. We use this measure to determine how good our classification rules are, and use the effectiveness of the classification rules to choose a definition from the region of legitimate definitions. In this way we adopt a good/bad definition which is not only acceptable to the bank, but which can be accurately predicted (or, at least, we do the best that can be done, and certainly better than using the bank's current definition).

Section 2 describes the data and the region of legitimate definitions for our banking example. Section 3 shows the results, demonstrating how the external criterion of classification performance can be used to help define the classes. Section 4 presents some conclusions.

The data

For our example we use records of 7956 bank accounts. A 'bad' account in any particular month is defined as above, with an account being 'good' otherwise. Our aim was to use the values of five variables in two consecutive months to predict the likely good/bad class six months in the future. These five predictor variables were debit turnover during the month (x_1), number of cheques (x_2), number of direct debits (x_3), value of debits (x_4), and value of charges (x_5). In the definition of the classes six months later we will use t_1 to denote the excess amount overdrawn above the nominal limit in (a), t_2 to denote excess amount overdrawn above the nominal limit in (b), t_3 to denote the maximum balance over the course of the month in (b), and t_4 to denote the ratio of total credit turnover in the month to the month's end balance in (c). Thus, in (our slightly modified version of) the bank's definition of the classes t_1 = 500, t_2 = 100, t_3 = 0, and t_4 = 0.1.

Since the aim of this paper is to compare different potential definitions of the classes, rather than the more common problem of studying the effectiveness of classification rules, we have adopted a simple and familiar method for constructing supervised classification rules, namely logistic discriminant analysis.

The models we will be looking at have only ten predictor variables, are based on 7956 data points, and are restricted to have a simple form, so that overfitting is unlikely to be a serious problem. However, a more subtle issue arises. We intend to search over the space of possible definitions (and, in fact, examine almost 6000 definitions, as explained below) and choose those which yield high values of our Gini criterion. The extra flexibility that this introduces (by requiring the estimation of another four parameters) may increase the scope for overfitting. In particular, although we can choose our definition according to the predictive performance on the data used for analysis, if overfitting has occurred this will be an unreliable guide to future performance.

To remove this risk, we split the data into ten parts, and took nine of these as the design set. Since these 9/10ths of the complete data are almost all of it, the resulting classification rules will be similar to those which would have been constructed using all of the data. (It would, of course, be better to use the leaving-one-out method, but computation time requirements made this impossible.) Each definition then yields ten resubstitution Gini coefficients, each based on design sets of size 9/10ths of the entire data set. The average of these is taken as the overall measure and is used to choose between alternative definitions. Once a definition has been chosen, the likely future performance of the classification rule with that definition is taken to be the average of the Gini coefficients for the ten test sets.

This procedure results in a choice of model in which the estimate of future performance is not subject to overfitting bias (being based in independent test sets), in which the estimate of performance is almost unbiased as an estimate of that based on the entire data set (since almost all of the available data is used in constructing the rules), and in which the measure of likely future performance is conditional on the design set (and not averaged over possible design sets), which is typically the focus of interest in real problems. Hand (1997) discusses such issues in detail.

Results

We take the region of legitimate definitions as being spanned by $t_1 \in$ [200, 800], $t_2 \in$ [50, 600], $t_3 \in$ [-150, 150], and $t_4 \in$ [0.05, 0.5].

A slight complication which needs to be considered is that, when the definition of the classes is altered, the sizes of the classes also changes. This, in turn has an effect on some measures of performance of classification rules. For example, it influences the very common measure misclassification or error rate. However, the Gini coefficient is invariant to changes in the sizes of the classes, provided the class-conditional distributions are unchanged. These points are further discussed in Hand and Kelly (1998).

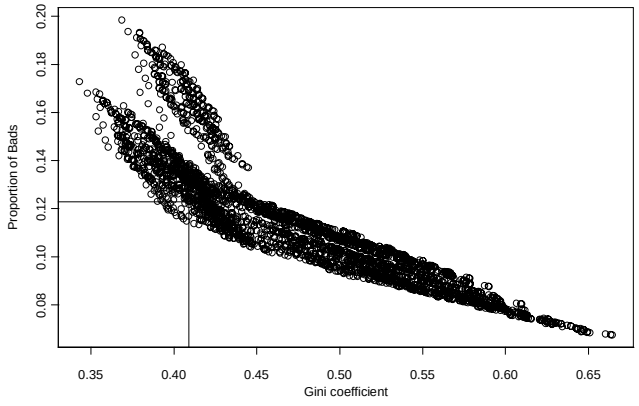
Of course, changing class sizes are still of interest: if an alternative definition of 'bad' means that hardly any accounts are defined as bad then this definition will probably be of limited interest to the bank. However, this sort of situation should seldom arise because the region of legitimate definitions will have been chosen by the bank to contain only definitions which were regarded as equally legitimate. One which defined virtually everyone as good would probably not have been proposed in the first place.

Figure 1 shows a scatterplot of proportion defined as bads against Gini coefficient (estimated by the 10-fold cross-validation) for definitions in the above region of legitimate definitions. (The increments in this plot were 100, 50, 50, and 0.05, for t_1 to t_4 respectively, leading to 5880 possible definitions.) From this one can see immediately that the Gini coefficient and bad rate are indeed negatively correlated, so that the definitions which lead to more accurate predictions are likely to be those which define a lower proportion of the population as bad. However, even if this correlation were perfect, one could still take advantage of the region of legitimate definitions. In this case one would simply adopt that definition leading to the greatest Gini coefficient - which would coincidentally be that leading to the smallest bad rate, but this would be acceptable by definition of the region of legitimate definitions, as noted above.

In any case, the correlation is not perfect. This means that one could, if one were so inclined, restrict one's choice to definitions which had the same bad rate as the original one, and still choose one with a higher Gini coefficient.

A word or two about random variation is appropriate here. All of the definitions under investigation are applied to the same data sets, so that any differences are genuine, for that data set: the differences in Gini coefficient displayed in the diagram are real for our data. Of course, since all data sets are finite, one would like to know the extent to which the obtained Gini coefficients are likely to change if a different sample was drawn. This is an important question, but one which does not affect the general thrust of our argument, so we do not discuss it here. If such random variability is thought to be a significant factor, it can be reduced by increasing the sample size.

Figure 1: Scatterplot of the bad rates for the 5880 definitions against the Gini coefficients of the logistic discriminant analyses.



The point in Figure 1 corresponding to the bank's original definition had 12% bads and a Gini coefficient of 0.41. However, one can also see immediately that the Gini coefficient of 0.41 achieved using the bank's current

definition of bad is rather poor compared with the predictability which can be achieved using alternative definitions. Figure 1 is rather deceptive in this regard, because of the overprinting - remember that there are 5880 points in the diagram. It is apparent from the histogram of Gini coefficient values for the 5880 different models given in Figure 2 that the majority of them do better than the bank's choice, some substantially better.

Figure 2: Histogram of the Gini coefficients for the logistic discriminant analyses applied to the 5880 definitions.

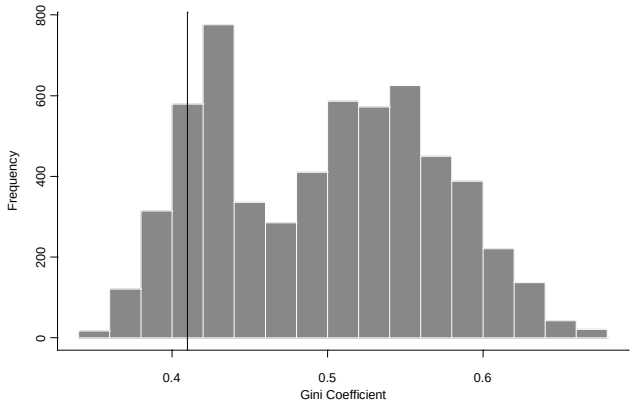


Table 1: Thresholds yielding four alternative definitions of 'bad account'. *Definition 1* is that given in Section 1.

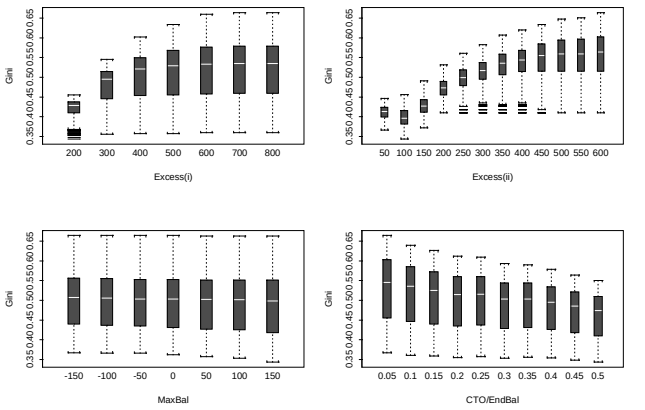
Definition	t_1	t_2	t_3	t_4	*Bad rate(%)*	*Gini*
1	500	100	0	0.10	12	0.41
2	400	150	-50	0.05	10	0.46
3	200	100	150	0.10	15	0.36
4	600	400	0	0.05	7	0.61

To illustrate the differences between definitions, we examine three alternative definitions in detail. We take the bank's original definition, given in Section 1, as *Definition 1*. The details are given in Table 1.

Of the three alternative definitions we have chosen to look at, *Definition 4* leads to the largest Gini coefficient (note that this is not the largest in the region of legitimate definitions defined above, which is 0.66, associated with a 7% bad rate). In the context of this problem, a change in the value of the Gini coefficient from 0.41 to 0.61 is vast. On the other hand, one might feel that this definition is on the bounds of legitimacy: the values of t_1, t_2, and t_3 are quite different from the corresponding values in *Definition 1* (even if they are not quite on the edge of what was specified as the region of legitimate definitions). *Definition 2* might be regarded as more acceptable. The difference of 5% in the values of the Gini coefficients between *Definitions 1* and *2* is, if not vast, certainly very large in this context, and definitely enough to be worthy of attention.

Insight into how the different thresholds t_i influence behaviour over the region of legitimate definitions can be obtained from studying the impact of different values of these variables individually and jointly. Figure 3 shows box plots for the alternative definitions at given values of each variable. It is clear from this that t_1 and t_2 have substantial effects on the Gini coefficient, that t_3 has virtually no effect, and that t_4 has a slight effect. It is also clear that most of the increase in the value of the Gini coefficient is achieved for values of t_1 and t_2 towards the centre of the region of legitimate definitions that we have defined. It is not necessary to go to the less acceptable extremes in order to realise improvement.

Figure 3: Box plots for the 5880 definitions by values of each defining variable separately. (t_1 is *Excess(i)*, t_2 is *Excess(ii)*, t_3 is *MaxBal*, and t_4 is *CTO/EndBal*.)



Although the various alternative definitions all lie within the region of legitimate definitions which we have specified, it is still conceivable (if extremely unlikely) that the definitions may lead to very different accounts being classified as bad. In general it is possible that two equal bad rates correspond to entirely different subjects - although in our example there must be significant overlap by virtue of the relationships between the different definitions. In any case, it would be reassuring to the bank to know that they did have a substantial number in common. A suitable measure of commonality is given by the proportion of the data on which two definitions yield the same classification. The commonalities between *Definition 1* and *Definitions 2* to *4* are 0.978, 0.979, and

0.952 respectively. If we were to recommend *Definition 2*, with its 5% improvement in the value of the Gini coefficient, it would disagree with *Definition 1* on only 2.2% of the definitions. In fact, of the 6951 goods according to *Definition 1*, 6946 would also be defined as good by *Definition 2*, and of the 1005 bads according to *Definition 1*, 832 would also be defined as bad according to *Definition 2*. *Definition 3*, with its hugely improved Gini coefficient, is perhaps even more interesting here. It agrees with *Definition 1* on all of this definition's 6951 goods. However, of the 1005 accounts which *Definition 1* defines as bad, *Definition 4* defines only 625 as bad. The huge improvement in Gini coefficient has been achieved by relaxing the definition of what is bad (as is clear from the definition above). In general, on these data, as we have already noted from Figure 1, bad rate and Gini coefficient are negatively correlated.

Conclusion

Many problems have elements of ambiguity about the precise meaning of some of the variables used. This may be because the variables are proxies for others, or it may be because there is simply no reason to prefer one measure to another closely related one, or it may be for other reasons. Often such ambiguity is removed by an overt operational definition of the variables in question. This is all very well, and it certainly removes the imprecision arising from the lack of a clear specification, but it is arbitrary and hence unsatisfying. We suggest, as an alternative, that advantage can be taken of the ambiguity by choosing that definition of the variables which optimises some other attractive criterion. We illustrated with a supervised classification example where there is a set of equally acceptable possible definitions for the classes.

Although our example spanned the region of legitimate class definitions by changing the thresholds on underlying continuous variables, this is by no means the only way of defining such a region. For example, one could consider including variables additional to those already used: different sets of variables can be used to span a region of legitimate definitions. Likewise, although we explored definitions within the region of legitimate definitions in some detail, this is not necessary. Having determined this region, one could simply numerically find that definition which maximised the external criterion.

Much data mining work seeks to identify interesting patterns in databases. So as to provide as much opportunity as possible, one would like to avoid specifying too narrowly what might be meant by 'interesting'. Our proposal takes a general definition of 'interesting' (in our example, any set of classes within the region of legitimate definitions is acceptable) and then chooses a specific definition on the grounds that it optimises some other criterion (in our example, predictability). Other 'interesting' definitions, since they could be less well predicted, though interesting, would be of less practical value.

Acknowledgements

The second author was supported by a CASE studentship from the Engineering and Physical Sciences Research Council in the UK, with additional support from Abbey National Plc. During the course of his work on this project, the third author was supported by grant number R022250001 from the Economic and Social Research Council. We would also like to express our appreciation to Sam Korman for his encouragement of this work.

References

Adams N. and Hand D.J. (1998) Comparing classifiers when the misallocation costs are uncertain. Submitted to *Pattern Recognition.*

Hand D.J. (1997) *Construction and Assessment of Classification Rules*. Chichester: Wiley.

Hand D.J. and Henley W.E. (1997) Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, **160**, 523-541.

Hand D.J. and Kelly M.G. (1998) Supervised classification when the class definitions are initially unknown. Submitted to *Data Mining and Knowledge Discovery*.

Hand D.J., Li H.G., and Adams N.M. (1998) Supervised classification with structured class definitions. In preparation.

Kelly M.G. and Hand D.J.(1997) Credit scoring with uncertain class definitions. Presented at *Credit Scoring and Credit Control V*, Edinburgh, September.