# The Impact of Changing Populations on Classifier Performance

Mark G. Kelly, David J. Hand, and Niall M. Adams
Department of Mathematics, Imperial College,
180 Queen's Gate, London,
SW7 2BZ, UK
m.g.kelly, d.j.hand, n.adams[@ic.ac.uk]

## Abstract

An assumption fundamental to almost all work on supervised classification is that the probabilities of class membership, conditional on the feature vectors, are stationary.

However, in many situations this assumption is untenable. We give examples of such *population drift*, examine its nature, show how the impact of population drift depends on the chosen measure of classification performance, and propose a strategy for dynamically updating classification rules.

## 1 Introduction

Much statistical work makes inferences about future sample values from an analysis of a 'design' sample of data. Intrinsic to this paradigm is the assumption that the population being studied does not change over time; that is, that the population from which future samples will be drawn has the same distribution as that from which the design sample was drawn. Unfortunately, this assumption often fails to hold. Indeed, one might argue that it almost always fails to hold. The question then arises as to whether the change in the shape of the distribution is such as to influence the conclusions one draws and the actions one should take. If the changes do influence performance, then we will be interested in developing models which can adapt to such changes. When the population distribution can change over time we say it is subject to *population drift.*

The extent of population drift is clearly a function of how rapidly the population distribution can change and how long a time scale is involved. It is likely to be more of a problem in some application areas than others. In a medical diagnostic problem in which the rate of demographic change is slow relative to advances in medical technology it is likely to be unimportant. In commercial applications, such as the credit granting example we discuss below, it is likely to be very important: the population of applicants will change in response to changing economic conditions and a changing competitive environment over the life of a financial product.

Machine learning work and Bayesian algorithms often emphasise adaptive models, developing methods for sequential estimation, in which the points are added one at a time to produce the final classification rule. Such methods can be readily adapted to the case in which populations evolve over time by extending them so that more recent points are included in the model and the effect of 'earlier' data points is gradually removed from the estimates (either by removing the impact of an individual point completely, or by downweighting it). Apart from these approaches, there has been some more direct work on dynamic updating. Nearest neighbour methods also have particular advantages in permitting ready dynamic updating of classification rules, provided no extensive pre-processing to reduce the number of stored points, choose metrics, or construct a 'multivariate ordering' is used. More recently, Taylor *et al* [7], Nakhaeizadeh *et al* [5] and Nakhaeizadeh *et al* [6] have developed a control chart type of system in which the action taken depends on the performance of the rule.

Population drift is related to *concept drift.* Whereas we regard population drift as referring to changes in the probability distributions of the phenomena under study, the term 'concept drift' has been used both for these and other changes. For example, it can refer to situations where the definitions of the classes in supervised classification problems can change over time (see, for example, Kelly and Hand [3]). Some recent work on concept drift is described in a special issue of *Machine Learning* (1998, Vol. 32, No.2) and in Widmer and Kubat [8] and Lane and Brodley [4].

To explore these ideas we studied a data set consisting of 92,258 unsecured personal loans with a 24 month term given by a major UK bank during the period 1 January 1993 to 30 November 1997. We define an

KDD-99 San Diego CA USA
Copyright ACM 1999 1-58113-143-7/99/08...$5.00

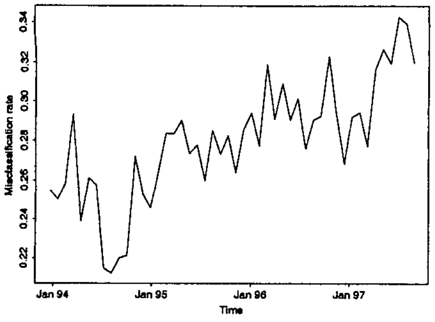

Figure 1: Plot of monthly misclassification rate.

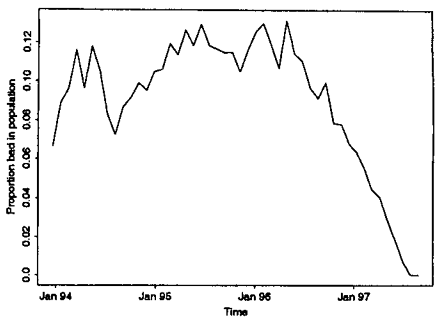

Figure 2: Plot of monthly bad prior.

account as being 'bad' if it has at least three months of arrears, and 'good' otherwise. We are concerned with classification rules for assigning customers to one of these two classes on the basis of seventeen variables describing the application for the loan. Note that the true class of each applicant is not discovered until up to two years after they have been granted the loan. This fact will become important below. Figure 1 shows the proportion of customers misclassified each month by a linear regression classifier built using the 1993 data as design set and applied to customers entering the system between January 1994 and November 1997. In this model the threshold is chosen at each month so that 20% of the applicants at each month are above the threshold (i.e. 20% of applicants will be rejected), so that the bank only accepts the 'best' 80% at each month. There is a clear upward trend over time to this curve; that is, as time progresses, so the proportion of customers who are misclassified increases. It is clear, from this figure, that something is changing over the course of time – that population drift does occur.

## 2 Effect of drift

Evolution of the population may occur in three ways. Firstly, and most simply, the class priors, $p(i)$, $i = 1, 2$, may change over time. Secondly, the distributions of the classes may change; that is, the $p(x|i)$, may alter over time. Thirdly, the posterior distributions of class memberships, the $p(i|x)$ may alter. For classification purposes, it is this third type of change which is relevant. A classification rule will be unaffected if the population to which it is applied evolves in such a way that the $p(i)$ and the $p(x|i)$ change but the $p(i|x)$ remain constant. (This is not entirely true. It is true if the true distributions are used, but inaccuracies in estimating the distributions will mean that classification performance is affected.) Changes of this kind can occur if the population changes are solely functions of $x$, so that they affect the classes in the same way. That is, it depends on whether the drift is merely affecting the distribution of the $x$ values and is not affecting the classes differentially. In general, if the $x$ variables include all the determinants of class membership probability, then changes in $p(x|i)$ will not translate into changes in $p(i|x)$. However, in situations where the class membership probabilities are also functions of other variables, then the $p(i|x)$ may also change. This is the case with our credit data. We examine each of these different potential kinds of drift in subsequent sections. In Section 3 we look at changing priors, in Section 4 we look at changing distributions over $x$, and in Section 5 we outline a strategy for modelling changes to the posterior distribution $p(i|x)$ and illustrate its application to our data. First, however, we demonstrate that each of the three different kinds of population drift does occur with our data. Evidence for the first type of drift, changing priors, is given in Figure 2. This shows the monthly proportion of bad customers plotted against time. The irregularity, month on month, is quite striking. Overall, the rate follows a gradual upward trend, before a sudden dramatic fall. The dramatic fall, though striking, is in fact of little interest. It simply reflects the fact that customers recruited during this period have had less than two years in which to go bad. That is, especially towards the end of this period, many customers are coded as 'good' who will in fact turn out to be bad. Evidence for the second kind of drift, changing $x$ distributions, is given in Figures 3(a) to (d). These show plots, over the five year period, of weekly averages for four of the predictor variables from our data. Figure 3(a) (proportion of applicants aged between 30 and 35) shows no apparent drift. Figure 3(b) (proportion of applicants who have a cheque guarantee card) shows a definite trend. Figure 3(c) (a binary indicator of loan purpose) shows some seasonal variation. Figure 3(d) (a binary 'repayment method' score) results from a policy change.

Figure 4 shows a graph of bad rate amongst accepts. This shows a gradually increasing slope, albeit again surprisingly irregular, and then a sudden fall, paralleling that of the change in bad rate. Thus classification performance is decreasing with time (and with increas-

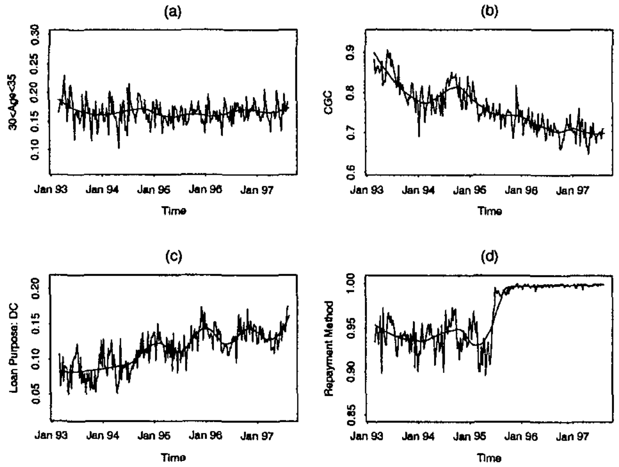

Figure 3: (a) to (d): Weekly averages of four binary predictor variables.

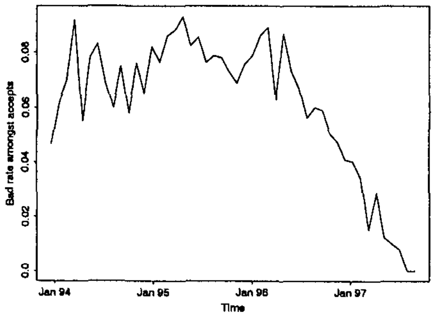

Figure 4: Plot of monthly bad rate amongst accepts.

ing bad rate), at least as measured by this criterion, before the fall. Figure 5 shows a plot of Gini index against time. The Gini index is a measure of how well separated are the distributions of estimated probabilities of belonging to the good class for the true goods and true bads. This measure is independent of the class priors. Even with the influence of class priors removed, there seems to be some evidence of drift over time. The increase in irregularity towards the end of the time period is presumably due to the smaller prior bad rate leading to poorer estimates of the distribution functions used in calculating the Gini index.

## 3 The impact of changing priors

To explore the impact of changing priors when the distributions remain fixed, we took a simple artificial case of two univariate normal distributions, one N(0,1) and one N($\mu$,1) and evaluated misclassification rate and bad rate amongst accepts as the prior proportion of bads varied from 0 to 1, for various values of $\mu$. We used a single point threshold as the decision surface. In this work we kept the proportion classified as bad fixed at 20%, since the strategy of fixing the accept

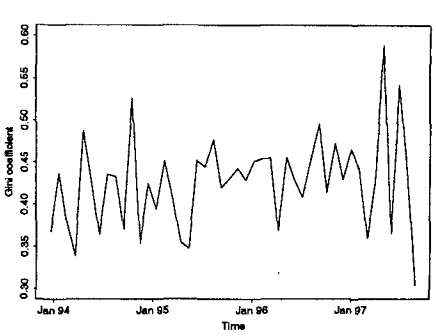

Figure 5: Plot of monthly Gini coefficient.

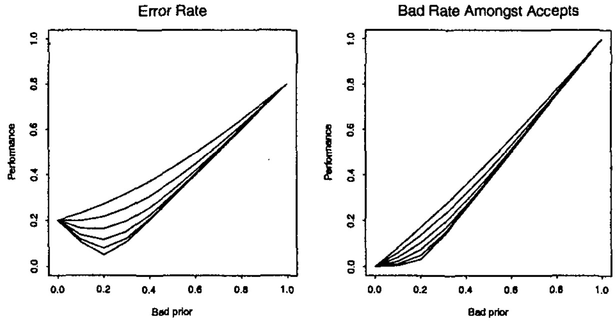

Figure 6: The effect of changing class priors on performance.

rate is one frequently adopted in the context of banking data of the kind used above. (Other rules could equally be adopted. One might simply want to minimise error rate, although this often results in assigning all objects to one class, especially if the other class is small. More generally, one might want to use a cost weighted loss, as described in, for example, Adams and Hand [1]). The results are shown in Figure 6 (with $\mu = 0.5$ being the top curve in each case, and lower curves corresponding to $\mu$ increasing in steps of 0.5). The bad rate amongst accepts increases monotonically, though not linearly, with increasing bad prior. When all customers are bad, the bad rate amongst accepts, with a fixed accept rate of 20%, is, of course, 1.0. The implication is that, for poorly separated classes, as we have in our data above, the bad rate amongst accepts is almost linearly related to the proportion of bads in the population.

The misclassification rate curves are rather more complicated. For well separated classes (high $\mu$) the misclassification rate can decrease initially as the proportion bad in the population increases.

These two families of curves illustrate firstly, the importance of choosing an appropriate performance criterion for the problem in hand, and secondly that the impact of changing class priors can be quite complicated.

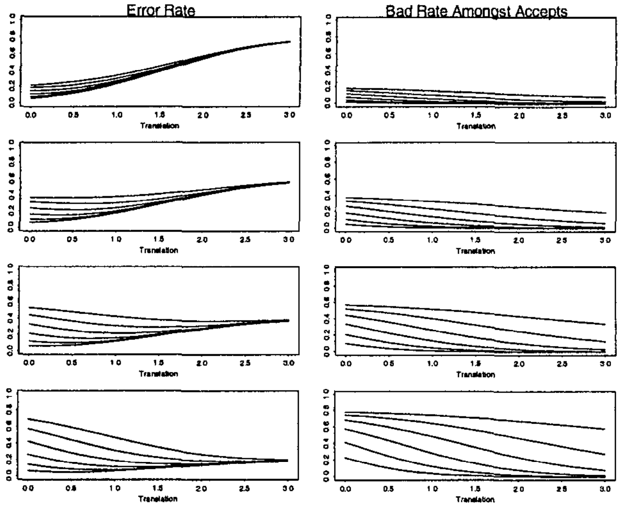

Figure 7: The effect of translating distributions on performance. The top row panels have bad prior 0.2, then 0.4, 0.6 and 0.8. The horizontal axis shows the amount by which the entire population is translated in the direction of the bad class.

## 4 The impact of changing distributions

Unlike priors, class distributions can change in an infinite number of different ways, so that it is difficult to draw conclusions which have general validity. However, so as to produce results which are comparable to those in Section 3, we adopted the same basic situation as in that section (univariate normals, N(0,1) and N($\mu$,1)), but allowed the distributions to be uniformly translated along the axis, in the direction of the bad class, N($\mu$,1). That is, we took the simplest case of population drift, one involving only translation of the combined population. The resulting curves are shown in Figure 7.

As with Figure 6, smaller $\mu$ corresponds to the higher curves in each panel. The horizontal axis corresponds to a translation of the distributions across the decision surface in the direction of the bad class. Thus, although initially the threshold is set so that 20% are classified as bad, for greater degree of translation, a larger proportion will in fact be classified as bad.

Error rate, in particular, shows interesting and changing patterns of behaviour as the $p(x|i)$ drift.

## 5 Adaptive classification

Given that population drift occurs and that it can adversely affect the performance of classification rules, how might one cope with it? The obvious strategy is to embed one's classification rule in a larger model which permits evolution over time. In particular, the estimated values of the parameters determining the model might be permitted to adapt as time progresses and the populations change. A basic dynamic linear model form (which can, of course, be generalised in obvious ways) is

$$\mathbf{y}_t = \mathbf{X}_t\beta_t + \mathbf{e}_t \; ; \; \beta_t = \mathbf{G}\beta_{t-1} + \mathbf{v}_t$$

Here $\mathbf{y}_t$ is a vector of observations made at time $t$, $\beta_t$ is a vector of system parameters at time $t$, $\mathbf{G}$ is a matrix describing the system, $\mathbf{X}_t$ is a matrix of independent variables at time $t$, $\mathbf{e}_t$ and $\mathbf{v}_t$ and are random normal vectors, N($\mathbf{0}$,$\mathbf{E}_t$), N($\mathbf{0}$,$\mathbf{V}_t$) , respectively. As an initial exploration of such methods in the context of our classification problems, we explored the simple special case

$$y_t = \mathbf{X}_t\beta_t + e_t \; ; \; \beta_t = \beta_{t-1} + \mathbf{v}_t$$

yielding a linear prediction. It is easy to extend this to generalised linear models, which may appear to be more appropriate for our context, in which the aim is to produce probability estimates which can be compared with a threshold. However, there is evidence that linear methods, with an appropriate choice of threshold, perform as well as logistic methods for the sorts of problems with which we are concerned here (Henley [2]), even though $y$ can take only two values. Moreover, because we are especially concerned with handling large data sets we required a quick and efficient updating procedure. For this reason, for our initial investigation, we have kept to linear models and used the updating procedure described below.

For a given $n \times d$ matrix, $\mathbf{X}$, of $n$ observations on the $d$ independent predictor variables, and $n$-vector $\mathbf{Y}$ of observations of the class variable, the standard regression estimate is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Incorporating a new observation provides no problem for the $\mathbf{X}'\mathbf{Y}$ factor in this, but the $(\mathbf{X}'\mathbf{X})^{-1}$ requires a fresh matrix inversion, for the matrix incorporating the new point. This can be avoided by updating $(\mathbf{X}'\mathbf{X})^{-1}$using the expression:

$$\left(\mathbf{X}'\mathbf{X} + \mathbf{x}_k\mathbf{x}_k'\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} - \frac{(X'X)^{-1}\mathbf{x}_k\mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}}{1 + \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k}$$

Since $(\mathbf{X}'\mathbf{X})^{-1}$ is a $d \times d$ matrix, storage is trivial. Using this expression, the orientation of the decision surface can easily be updated as soon as a new point, with known class membership, becomes available. A similar expression allows one to remove points without inverting the matrix from scratch. We also need to consider how to update the threshold with which the linear rule is compared. We are especially interested in the bad rate when 80% of applicants are accepted, so we update the threshold so that this is maintained. Figure 8 shows the sorts of results one can obtain. This is one of the simplest of models, in which we have taken a year's worth of data (customers taking out loans in 1994) and

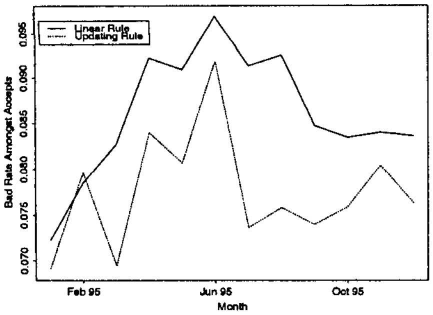

Figure 8: Plot of bad rate amongst accepts for the updating and static classification strategies.

applied the standard model (simple linear regression based on the 1994 data) and a model which begins with the 1994 data and then updates the classification rule as new points become available. Both models are applied throughout 1995, with the classification performance (in terms of bad rate amongst the 80% accepted) plotted each month. The plot starts at month 1 (month 0 is December 1994, at which the two models produce identical results).

## 6 Conclusion

In some classification problems the class structure can change over the lifetime of application of any classification rule. These changes can occur via alterations in the relative class sizes, via alterations in the class conditional distributions, or via alterations in the posterior probabilities of class memberships. Changes in the class conditional distribution which are solely functions of the $x$ variables will not affect the posterior probabilities. These are only affected if the class distributions change differentially. However, in situations where there are determinants of class membership probability beyond the predictor variables, as in our example, such differential effects are likely.

If the class membership probabilities change, a classification rule which does not itself evolve to reflect such changes may exhibit deteriorating performance. This can be tackled by developing more elaborate classification rules which dynamically update themselves to model the changing distributions. These can be based on incrementally adding in the effect of new points, as their true class memberships become known, and on deleting the effect of 'older' points. The precise choice of how much 'older' will determine how rapidly the prediction can increase or decrease. In the latter case, one may want to apply a gradual downweighting rule, rather than a sudden exclusion of points and, again the rate of downweighting will determine the extent of regularisation. Indeed, as Taylor *et al* [7] point out, one might want to let the length of the historical record of points used in the rule vary as time progresses: in times of dramatic change, perhaps the classification rule should be based on a shorter span of previous points.

## Acknowledgements

The work of MGK was supported by a CASE studentship from the UK's Engineering and Physical Sciences Research Council, with additional support from Abbey National Plc. We would like to express our appreciation to Sam Korman and Steve Bull for their interest in and encouragement of this work.

## References

[1] Adams N.M. and Hand D.J. (1999) Comparing classifiers when the misallocation costs are uncertain. To appear in *Pattern Recognition*, **32**.

[2] Henley W.E. (1995) *Statistical aspects of credit scoring*. Unpublished PhD thesis, The Open University, Milton Keynes, UK.

[3] Kelly M.G. and Hand D.J. (1999) Credit scoring with uncertain class definitions. To appear in *IMA Journal of Mathematics Applied in Business and Industry.*

[4] Lane T. and Brodley C.E. (1998) Approaches to online learning and concept drift for user identification in computer security. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, ed. R.Agrawal, P.Stolorz, and G.Piatetsky-Shapiro. AAAI Press, Menlo Park, California. 259–263.

[5] Nakhaeizadeh G., Taylor C.C., Kunisch G. (1997) Dynamic supervised learning: some basic issues and application aspects. *Classification and Knowledge Organization*, ed. R. Klar and O. Optiz, Berlin: Springer-Verlag, 123–135.

[6] Nakhaeizadeh, G., Taylor, C. and Lanquillon, C. (1998). Evaluating usefulness for dynamic classification. *Knowledge Discovery and Data Mining KDD-98*. ed. R. Agrawal, P. Stolorz, G. Piatetsky-Shapiro. AAAI, 87–93.

[7] Taylor, C.C., Nakhaeizadeh, G., and Kunisch, G. (1997) Statistical aspects of classification in drifting populations. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, 521–528.

[8] Widmer, G. and Kubat M. (1996) Learning in the presence of concept drift and hidden contexts. *Machine Learning*, **23**, 69–101.