Mining in the presence of selectivity bias and its application to reject inference

A.J. Feelders
Tilburg University
Faculty of Economics
CentER for Economic Research
PO Box 90153, 5000 LE Tilburg
The Netherlands
e-mail: A.J.Feelders@kub.nl

Soong Chang and **G.J. McLachlan**
University of Queensland
Department of Mathematics
St. Lucia, Brisbane
Australia
gjm@maths.uq.edu.au

Abstract

Standard predictive data mining techniques operate on the implicit assumption of random sampling, but data bases seldomly contain random samples from the population of interest. This is not surprising, considering company data bases are primarily maintained to support vital business processes, rather than for the purpose of analysis. The bias present in many data bases poses a major threat to the validity of data mining results. We focus on a form of selectivity bias that occurs frequently in applications of data mining to scoring. Our approach is illustrated on a credit data base of a large Dutch bank, containing financial data of companies that applied for a loan, as well as a class label indicating the repayment behavior of accepted applicants. With respect to the missing class labels of the rejected applicants, we argue that the missing at rondom (MAR) case becomes increasingly important, because many banks nowadays use formal selection models. The classification problem is modeled with mixture distributions, using likelihood-based inference via the EM-algorithm. Since the distribution of financial ratios is notably non-normal, class-conditional densities are modeled by mixtures of normal components as well. The analysis shows that mixtures of two normal components usually give a satisfactory fit to the within-class empirical distribution of the ratios. The results of our comparative study show the selectivity bias caused by ignoring the rejects in the learning process.

Introduction

Standard predictive data mining techniques operate on the implicit assumption of random sampling, but data bases seldomly contain random samples from the population of interest. This is not surprising, considering the fact that company data bases are maintained primarily to support vital business processes, rather than to perform data mining analyses. The bias present in many data bases poses a major threat to the validity of data mining results.

Consider the following example to illustrate the problem. A direct marketing bank receives both written and telephonic loan applications. The data of written applications are always entered into the data base, regardless of whether the loan is accepted or rejected. In case of a telephonic application however, the data of the applicant are not always entered into the data base. If the bank employee quickly finds out that the applicant cannot be accepted, the conversation is usually ended without the data being entered. This allows the bank employee to help more clients, but clearly data quality suffers.

There is no generally applicable method to correct for the multitude of biases that may occur. Instead, we focus on a particular form of bias that frequently occurs, particularly in applications of data mining to scoring.

In credit scoring, loan applicants are either rejected or accepted depending on characteristics of the applicant such as age, income and marital status. Repayment behaviour of the accepted applicants is observed by the creditor, usually leading to a classification as either a good or bad (defaulted) loan. As repayment behaviour of rejects is for obvious reasons not observed, complete data is available only for accepted applicants. Since the creditor does not accept applicants at random, this constitutes a non-random sample from the population of interest. Construction of a classification rule based on accepted applicants only, may therefore lead to invalid results. This is, in a nutshell, what is called the reject inference problem in the credit scoring literature.

In the next section we formulate reject inference as a problem of learning with missing data. Subsequently, we discuss the analysis of a credit data base of a large Dutch bank, using mixture modeling. We compare the results obtained by using reject inference, to those using standard supervised learning techniques. Finally we draw a number of conclusions, and indicate some topics for further research.

Credit scoring and missing data

In order to structure the following discussion, we distinguish between two stages in the credit scoring proces. The first stage is the *selection mechanism* that determines whether an applicant is rejected or accepted by the bank. The second stage is the *outcome mechanism* that determines the response of the selected cases. We also refer to the first stage as the *missing-data mechanism*, since it determines for which cases the response is observed in the next stage. Our primary interest is to model the outcome mechanism.

We assume some set of variables X is completely observed, and the class label Y is missing for the rejected applicants. Following the classification used in (Little & Rubin 1987), we distinguish between the following situations.

Missing completely at random

The probability that Y is missing is independent of X and Y. We don't consider this case any further here, since it would only apply if the bank were to accept applications at random.

Missing at random

The probability that Y is missing depends on X but not on Y. This situation frequently occurs in practice, since many credit institutions nowadays use a formal selection model. In that case, Y is observed only if some function of variables occurring in X exceeds a threshold value, say $f(X_s) \geq c$, where $X_s \subseteq X$. The missing-data mechanism is ignorable for likelihood based inference. The use of standard supervised learning techniques, ignoring the rejected cases altogether, may lead to invalid results however (Hand & Henley 1993).

Non-ignorably missing

The missing-data mechanism is non-ignorable when it depends on Z, which includes variables not contained in X. If the variables in Z supply any 'extra' information about class membership, then the probability that Y is missing (given X), depends on Y as well. This typically occurs when selection is partly based on characteristics that are not recorded in X, for example the 'general impression' that the loan officer has of the applicant. It may also occur when a formal selection model is used, but is frequently 'overruled' by a loan officer on the basis of characteristics not recorded in X.

It is required to model the selection mechanism as well as the outcome mechanism. For methods applicable to scoring when the selection mechanism is non-ignorable, we refer the reader to (Boyes, Hoffman, & Low 1989; Copas & Li 1997).

Analysis and Modeling of credit data

In this section we discuss the analysis of a credit data set from a large Dutch bank. Since the use of real rejected applications would make model evaluation impossible, we used accepted applications only, and created a realistic selection criterion. As remarked by Hand (Hand 1997), unclassified observations can only help in estimating the outcome mechanism if *global parametric forms* are assumed for the class-conditional distributions. We use mixture distributions (McLachlan & Basford 1988) to model the outcome mechanism.

Data description and study design

The data base consists of about 11,000 records containing financial data over the year 1992 of companies from several industry branches. All these companies succesfully applied for a loan of 5 million dutch guilders maximum somewhere before 1992. Companies in the loan portfolio are periodically evaluated by the bank, and receive a rating from 1 to 7, the lower rating indicating the better credit risk. For the purpose of this study we used data from the retail branch, consisting of about 2,400 records.

In addition to the financial data over 1992, the data base contains two such ratings for each company: one for the year '92 and one for '94. An artificial group of 'rejected applications' was created using the 1992 rating: companies having ratings ≥ 5 were considered to be rejected. Consequently, their credit rating in 1994 was pretended to be unknown, and was actually only supplied to us by the bank *after* the modeling stage. The class label assigned to the accepted loans was derived from the credit rating in '94, with ratings < 5 receiving the label 'Good', and ratings ≥ 5 receiving the label 'Bad'. The rating in 1992 represents the selection mechanism, and the 1994 rating represents the outcome. The problem then is to predict the outcome, on the basis of financial data of two years before.

We are particularly interested in the improvement in predictive performance that may be obtained by including the rejects in the analysis. To this end we compare the performance of a model estimated on the accepted loans only, to a model estimated on the accepted and rejected loans together.

Modeling the outcome

For our analysis, we focused on two financial ratios, debt ratio and working capital/total assets, denoted by X_1, and X_2. Initially, we worked with the data of known classification; that is, with the loans which had a known class label of good or bad. An inspection of the histograms of these two variables for the good and bad group of loans considered separately suggests that the group-conditional (marginal) distributions of these variables are non-normal. Consequently, we adopted a mixture of g normal component densities with unequal variances and covariances to model the joint distribution of these variables. It was concluded that a mixture of $g = 2$ normal components suffices for this purpose. To illustrate the good fit provided by the two-component normal mixture models to the initially classified groups of good and bad loans, we have plotted for the first variable the fitted mixture cumulative distribution function (CDF) versus the empirical CDF in Figure 1 and 2, corresponding to the good and bad loans, respectively.

In the sequel therefore, the distribution of the vector $X = (X_1, X_2)^T$ for the class of good loans is modeled

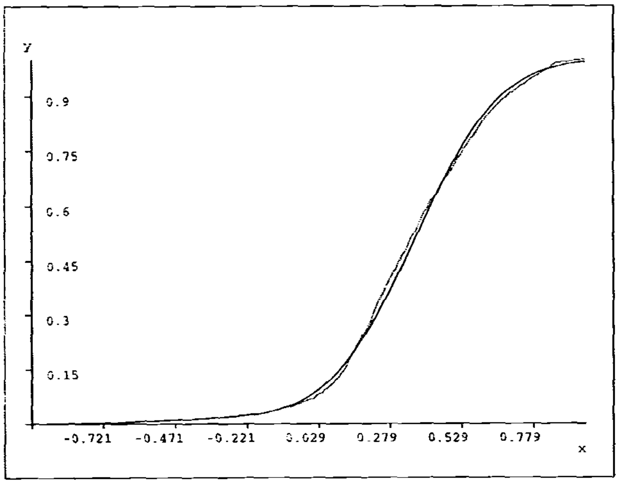

Figure 1: Fitted (thick line) and empirical (dotted line) cdf of debt ratio for good loans

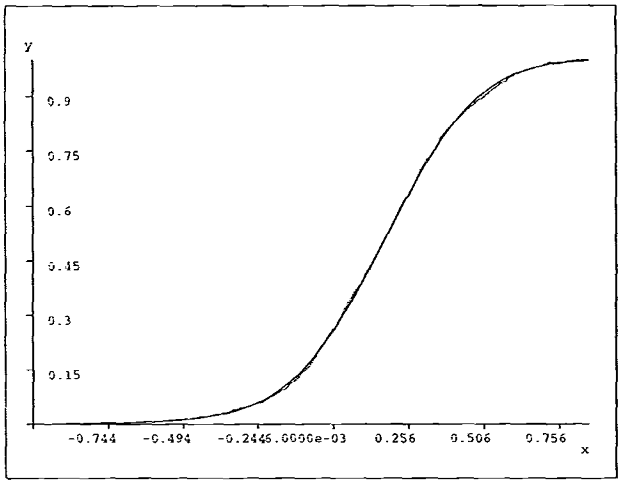

Figure 2: Fitted and empirical (dotted line) cdf of debt ratio for bad loans

component	mixing proportions	mean		variance	
		X_1	X_2	X_1	X_2
11 (good)	0.804	0.358	0.170	0.058	0.062
12 (good)	0.046	-0.320	-0.300	0.092	0.143
21 (bad)	0.140	0.165	0.015	0.067	0.083
22 (bad)	0.010	-0.440	-0.474	0.079	0.104

Table 1: Component estimates including rejects

as

$$f_1(x) = \sum_{h=1}^{2} \pi_{1h}\phi(x;\ \mu_{1h},\ \Sigma_{1h}), \qquad (1)$$

where $\phi(x;\ \mu,\ \Sigma)$ denotes the (bivariate) normal density with mean μ and covariance matrix Σ, and where π_{11} and π_{12} denote the mixing proportions, which are nonnegative and sum to one. The corresponding mixture distribution for the class of bad loans is

$$f_2(x) = \sum_{h=1}^{2} \pi_{2h}\phi(x;\ \mu_{2h},\ \Sigma_{2h}). \qquad (2)$$

In order to include the rejects in our analysis, we fitted the four-component normal mixture model

$$\begin{aligned} f(x) &= \sum_{i=1}^{2} \pi_i f_i(x) & (3)\\ &= \sum_{i=1}^{2} \pi_i \{\sum_{h=1}^{2} \pi_{ih}\phi(x;\ \mu_{ih},\ \Sigma_{ih})\}, & (4)\end{aligned}$$

where π_1 and $\pi_2 = 1 - \pi_1$ denote the proportion of good and bad loans. The mixture model (4) was fitted by maximum likelihood via the EM algorithm of (Dempster, Laird, & Rubin 1977) to all the data, including the rejects; see also (McLachlan & Krishnan 1997) for a recent account of the EM algorithm.

Results of the analysis

In Table 1, we have listed the estimated means, variances, and mixing proportions of the four components in the normal mixture model (4), as obtained on the basis of accepted (classified) and rejected (unclassified) loans. In Table 2 we listed the estimates obtained on the accepted loans only, and Table 3 contains the estimates obtained on the complete data, including the true class label of the rejects. To illustrate the bias in the estimates obtained by working with just the data on the initially classified loans, we have plotted for the first variable in Figure 3, the estimate of the two-component normal mixture density (1) for the class of good loans, along with the estimated density obtained by working with just the initially classified group of good loans.

In Figure 4, we give the corresponding plot for the bad loans. In practice, the classification of the rejected loans is unknown. But for the data under analysis here, we do know their true classification. Hence to illustrate how effectively the unclassified data on the rejects has been used in forming the estimates of the class-conditional densities, we have plotted in Figure 5 the density estimates based on all the good loans (that is, both initial and those in the rejects) along with the estimate based on the initially classified loans and the rejects. The plots of the corresponding estimates for the density of a bad loan is given in Figure 6. The good agreement between these latter two estimates loans is

component	mixing proportions	mean		variance	
		X_1	X_2	X_1	X_2
11 (good)	0.965	0.373	0.180	0.054	0.061
12 (good)	0.035	-0.303	-0.274	0.106	0.142
21 (bad)	0.954	0.205	0.039	0.058	0.081
22 (bad)	0.046	-0.432	-0.396	0.090	0.125

Table 2: Component estimates on accepted loans

component	mixing proportions	mean		variance	
		X_1	X_2	X_1	X_2
11 (good)	0.956	0.365	0.174	0.056	0.061
12 (good)	0.044	-0.284	-0.285	0.095	0.131
21 (bad)	0.963	0.115	-0.013	0.079	0.094
22 (bad)	0.037	-0.717	-0.597	0.020	0.100

Table 3: Component estimates on complete data

encouraging.

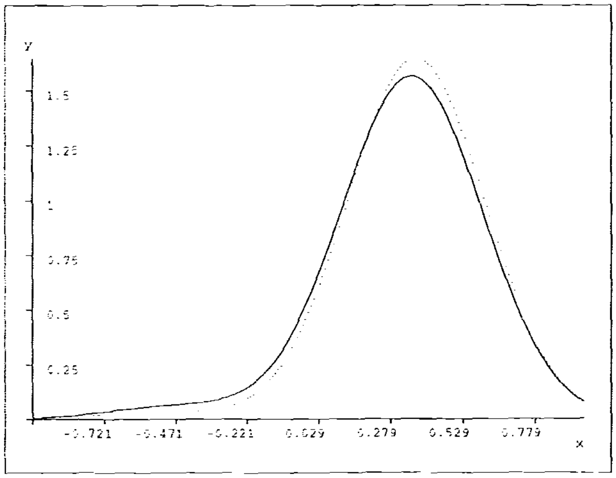

Figure 3: Estimated density of debt ratio of good loans with (thick) and without (dotted) rejects

We let r_1 be the allocation rule based on just the initially classified loans, and r_2 the rule based on all the data (initially classified loans plus rejects). These rules are formed by replacing the unknown parameters with their estimates in the Bayes allocation rule, where an observation is assigned to the class to which it has the highest posterior probability of belonging. On applying these two rules to the rejects, it was found that both rules allocated all the rejected loans to the good class. Hence the bad loans among the rejected loans were all misallocated. As the prior probability for the class of good loans is so much greater than that for the class of bad loans, the observed data on a loan has to have a very small density in the good class for it to be assigned to the bad class.

Although the two rules r_1 and r_2 give the same outright assignment of the rejected loans, the estimates of the posterior probabilities of class membership tend to be better for the second method that uses all the data rather than for the first method that uses the data on just the initially classified loans. A general measure for the fit of the estimated posterior probabilities of the rejects is cross-entropy, defined as follows

$$E = -\sum_n t^n \ln(y^n) + (1 - t^n)\ln(1 - y^n)$$

where $t^n = 1$ for good loans and $t^n = 0$ for bad loans, and y^n is the estimated posterior probability of a good loan. The results are depicted in Table 4, with lower values indicating a better fit.

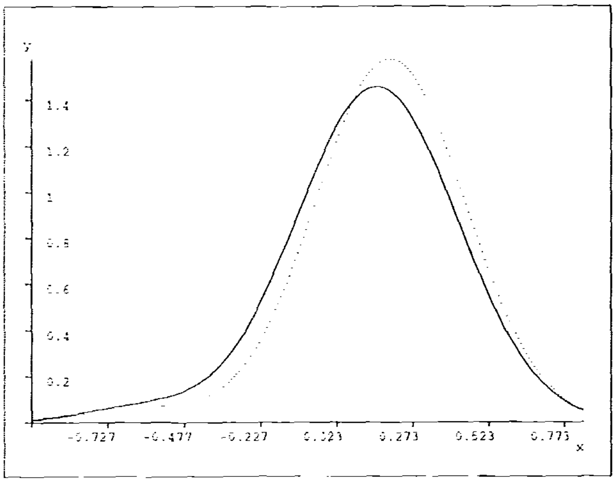

Figure 4: Estimated density of debt ratio of bad loans with (thick) and without (dotted) rejects

Method	Rejects	All loans
Accepts only	185.96	958.84
Accepts and rejects	184.23	956.50
Complete data	158.10	939.92

Table 4: Cross-entropy on rejects and all loans

The leftmost column of Table 4 indicates which data have been used to fit the model. Column 2 and 3 indicate the cross-entropy of the fitted model on the rejects and all loans respectively. The overall conclusion is that including the unclassified rejects gives a slightly better model fit than the fit obtained when ignoring them altogether. Both models are however clearly worse than the model fitted on the complete data.

Conclusions and further research

We presented a mixture modeling approach to learning from data that suffer from a frequently occurring form of selectivity bias. Analysis of a credit data set demonstrated the bias resulting from ignoring the unclassified cases. Inclusion of the rejects gave slightly better results, but there clearly is scope for further improvement. We showed that the distribution of financial ratios can often be modeled adequately by a mixture of two normal components. This result may be of interest to other applications of data mining in finance, for example bankruptcy prediction models. It should be noted however that in credit scoring attributes are often discrete, and in that case the proper choice of parametric form is more problematic (Hand 1997). The approach suggested by (Lawrence & Krzanowski 1996) appears to be an interesting line of research here.

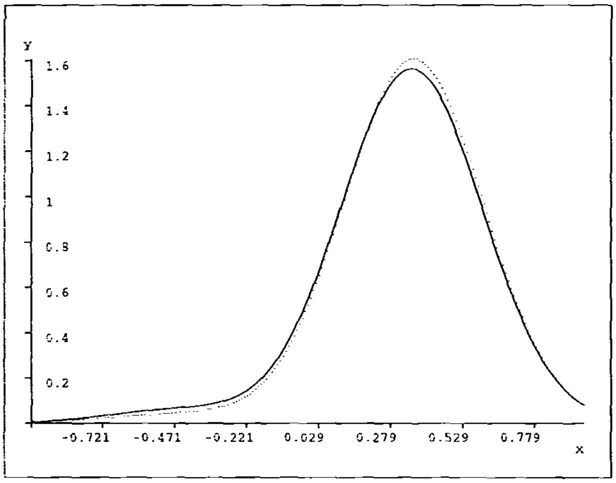

Figure 5: Estimated density of debt ratio of good loans with complete (dotted) and incomplete (thick) data

References

Boyes, W.; Hoffman, D.; and Low, S. 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* 40:3–14.

Copas, J., and Li, H. 1997. Inference for non-random samples. *Journal of the Royal Statistical Society B* 59(1):55–95.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B* 39:1–38.

Hand, D., and Henley, W. 1993. Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry* 5(4):45–55.

Hand, D. J. 1997. *Construction and Assessment of Classification Rules.* Chichester: John Wiley.

Lawrence, C., and Krzanowski, W. 1996. Mixture separation for mixed-mode data. *Statistics and Computing* 6:85–92.

Little, R. J., and Rubin, D. B. 1987. *Statistical analysis with missing data.* New York: John Wiley & Sons.

McLachlan, G. J., and Basford, K. E. 1988. *Mixture models, inference and applications to clustering.* New York: Marcel Dekker.

McLachlan, G., and Krishnan, T. 1997. *The EM Algorithm and Extensions.* New York: John Wiley.

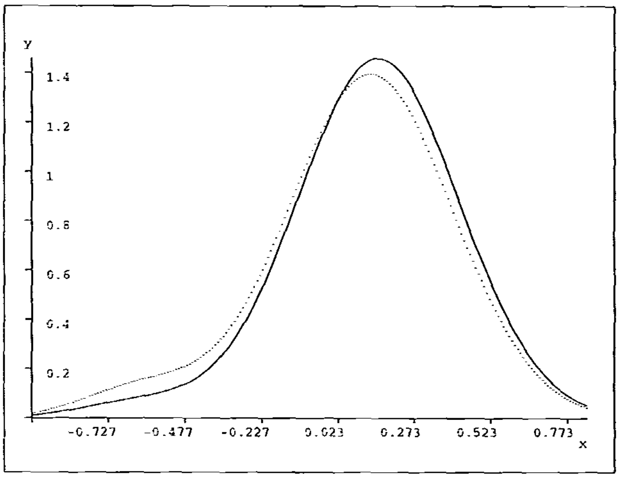

Figure 6: Estimated density of debt ratio of bad loans with complete (dotted) and incomplete (thick) data