Learning from biased data using mixture models

A.J. Feelders
Data Distilleries Ltd.
Kruislaan 419
1098 VA Amsterdam
The Netherlands
email: ad@ddi.nl

Abstract

Data bases sometimes contain a non-random sample from the population of interest. This complicates the use of extracted knowledge for predictive purposes. We consider a specific type of biased data that is of considerable practical interest, namely non-random partially classified data. This type of data typically results when some screening mechanism determines whether the correct class of a particular case is known. In credit scoring the problem of learning from such a biased sample is called "reject inference", since the class label (e.g. good or bad loan) of rejected loan applications is unknown. We show that maximum likelihood estimation of so called mixture models is appropriate for this type of data, and discuss an experiment performed on simulated data using mixtures of normal components. The benefits of this approach are shown by making a comparison with the results of sample-based discriminant analysis. Some directions are given how to extend the analysis to allow for non-normal components and missing attribute values in order to make it suitable for "real-life" biased data.

Introduction

It frequently occurs that company data bases used for data mining contain a non-random sample from the population of interest. This situation complicates generalization of the patterns found in the data base, especially when the knowledge extracted is intended for predictive purposes.

We discuss a special case of such a "biased" data base, that is of considerable practical interest. We consider data bases of *partially* classified data, where the "mechanism" that determines whether the class label of a particular case is known, is non-random. Furthermore, we assume that the objective of the data mining excercise is to learn a classification rule that is able to predict the class label of an unseen case that is taken at random from the population of interest.

The situation described above typically occurs when some kind of "screening mechanism" determines whether the correct class of a particular case is known. For example, a bank decides on the basis of a combination of attributes, whether a loan application is accepted or rejected. Only when the loan is accepted will the bank eventually be able to label the loan as "good" or "bad", depending on the payment behaviour of the client. For the loan applications that are rejected, the correct class label cannot be determined with certainty. When the bank wants to use the data base to learn to tell the difference between good and bad applications, a problem arises. If only the classified cases are used, bias is introduced, since the classified cases are a non-random sample from the population of interest. A rule thus learned will only generalize reliably to the group of applicants that pass the screening mechanism (Feelders, le Loux, & Zand 1995). Furthermore, one is not able to determine in this way whether one has been incorrectly rejecting a particular group of applicants. In the credit-scoring literature, this problem is known as the *reject inference* problem. For convenience we continue to use this term, although the problem is obviously more general than credit-scoring alone. Similar cases of screening may for example occur in application areas such as marketing, insurance, and medicine.

In the next section, we formulate and classify the reject inference problem as a problem of learning with missing data. It turns out that likelihood-based approaches are appropriate for this class of missing-data problems. After that we derive the appropriate likelihood based on a mixture model formulation of the problem. We present a case study, using simulated data, to show the potential benefits of the approach. Finally, we draw a number of conclusions and indicate how mixture modeling may be applied to reject inference in more realistic situations.

Reject inference as a missing data problem

Reject inference may be formulated as a missing data problem, where the attributes X are completely observed, and the class label Y is missing for some of the observations. Following the classification used in (Little & Rubin 1987), we distinguish between the following situations, according to whether the probability of Y being missing

1. is independent of X and Y,
2. depends on X but not on Y,
3. depends on Y, and possibly X as well.

If case 1 applies, both sampling-based and likelihood-based inference may be used without the results being biased. For sampling-based, one could just perform the analysis on the classified observations, and ignore the unclassified ones. In the credit-scoring example, case 1 would apply when loan applications are accepted at random, e.g. by simply accepting all applications up to a certain number. This way of "buying experience" has been used to a certain extent by credit institutions, although there are obvious economic factors that constrain the use of this method (Hsia 1978). One may of course also consider using the standard selection mechanism, but accepting rejects with a predetermined probability. This bias could then be corrected easily by weighting the observations according to their probability of ending up in the sample.

Case 2 applies, when the observed values of Y are a random sample of the sampled values within subclasses defined by the values of X (Little & Rubin 1987). This is the case for the reject inference problem, since applications with particular predefined combinations of attributes are accepted and the other applications are rejected. Under these conditions, the missing-data mechanism is ignorable for likelihood-based inference, but not for sampling-based inference.

Finally, if case three applies the missing-data mechanism is nonignorable.

From the foregoing we conclude that likelihood-based inference is a viable approach to the reject inference problem, and start the analysis by formulating the appropriate likelihood, using mixture models.

Mixture distributions

Mixture distributions (Everitt & Hand 1981; Titterington, Smith, & Makov 1985; McLachlan & Basford 1988) are distributions which can be expressed as superpositions of a number of component distributions. Henceforth we assume that the number of components equals the relevant number of classes, so each component models a class-conditional distribution.

As an example, one might express the income density function in the form

$$f(income) = f_1(income; b)\pi_b + f_2(income; g)\pi_g$$

where π_b and π_g are, respectively, the probabilities that a loan application is bad or good (called the mixing proportions), and f_1 and f_2 are the income density functions for bad and good applications. Thus the density function of income has been expressed as a superposition of two conditional density functions. This idea is easily generalized to the multivariate case, e.g. when the joint density of income and age is expressed as a mixture of the joint densities of these features for bad and good loans respectively. In general, a finite mixture can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{x}; \boldsymbol{\theta}_i)$$

where c is the number of components, π_i the mixing proportions and $\boldsymbol{\theta}_i$ the component parameter vectors. Usually, one assumes that it is unknown from which component observations are drawn, and one wants to estimate the *mixing proportions* and the parameters θ_i of the component distributions.

Suppose there are available attribute vectors $\mathbf{x}_j$ observed on m entities of unknown class, and sampled from a mixture C of $C_1, \ldots, C_c$ in unknown proportions $\pi_1, \ldots, \pi_c$. Then the relevant parameters can be estimated using maximum likelihood, taking the following likelihood function

$$L_1(\Psi) = \prod_{j=1}^{m} \left\{ \sum_{i=1}^{c} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right\}$$

where $\Psi = (\boldsymbol{\pi}', \boldsymbol{\theta}')'$ denotes the vector of all unknown parameters.

Likelihood L_1 is appropriate when all observations are of unknown class. When we also include n classified observations, the likelihood has to be adjusted accordingly. With respect to the classified entities one distinguishes between two sampling schemes, separate sampling and mixture sampling. In case of separate sampling, random samples of size n_i are drawn from each class separately. Consequently, the relative frequencies n_i/n of the classes do not give any information about the mixing proportions. In case of mixture sampling, the classified entities are obtained by sampling from mixture C, and the resulting relative frequencies of the different classes do provide information on the mixing proportions. For reasons that become clear shortly, we proceed with formulating the likelihood under the assumption that the classified entities have been obtained by mixture sampling.

$$L_2(\Psi) = L_1(\Psi) \prod_{j=m+1}^{m+n} \left\{ \sum_{i=1}^{c} z_{ij} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right\}$$

where z_{ij} equals 1 if observation j has class-label i, and zero otherwise.

For computational convenience one often considers the loglikelihood $L_3 = \log L_2$

$$\begin{aligned} L_3(\Psi) \quad = \quad & \sum_{j=1}^{m} \log \left\{ \sum_{i=1}^{c} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right\} + \\ & \sum_{j=m+1}^{m+n} \sum_{i=1}^{c} z_{ij} \log(\pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i)) \end{aligned}$$

Let us recall that likelihood L_2 was formulated under the assumption that both the classified and unclassified cases are random samples from the relevant mixture C.

This does unfortunately not apply to the situation considered here, since the selection of classified and unclassified observations has been performed in a systematic way; an observation is not classified if it is located in some subregion of the attribute space. If one assumes however that the total sample of size $m+n$ is a random sample from mixture C, then it can be shown (see (McLachlan 1992), section 2.8) that the relevant likelihood, apart from a combinatorial additive term, reduces to L_2. This means that one can estimate the parameters of the class-conditional (component) distributions using likelihood L_2, even when the separation between classified and unclassified observations is non-random.

A maximum likelihood estimate of Ψ can be obtained using the EM algorithm. The general strategy is based on optimizing the complete-data loglikelihood

$$L_C = \sum_{j=1}^{m+n} \sum_{i=1}^{c} z_{ij} \log(\pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i))$$

In the first E-step, one uses some initial estimate $\Psi^{(0)}$, to calculate the expectation of the complete-data loglikelihood. This is done by calculating the posterior probabilities

$$\tau_{ij} = \frac{\pi_i f_i(\mathbf{x}_j)}{\sum_{i=1}^{c} \pi_i f_i(\mathbf{x}_j)} \quad (1)$$

of group membership for the unclassified cases, and entering these as values of z_{ij} in the complete-data loglikelihood. In the M-step, the algorithm chooses $\Psi^{(k)}$ that maximizes the complete-data loglikelihood that was formed in the last E-step. The E and M steps are alternated repeatedly until convergence. It has been shown that, under very weak conditions, this algorithm will yield a local maximum of likelihood L_2 of the incomplete-data specification. For a more detailed and rigorous account of the application of EM to this problem, the reader is referred to (McLachlan 1992), pages 39–43.

Example of reject inference using mixture models

In this section we give an example of the possibility of performing reject inference using mixture models. To this end we generate a synthetic data set of loan applications, and a decision rule to determine whether a loan application is accepted or rejected. For the sake of simplicity we assume that only two normally distributed attributes are recorded for each loan application. The 1000 bad loans are drawn from the following distribution

$$\mu_b = \begin{pmatrix} 96.8 \\ 15.6 \end{pmatrix} \qquad \Sigma_b = \begin{pmatrix} 584.6 & -39.7 \\ -39.7 & 3.6 \end{pmatrix}$$

The 1000 good loans are drawn from

$$\mu_g = \begin{pmatrix} 137.8 \\ 9.2 \end{pmatrix} \qquad \Sigma_g = \begin{pmatrix} 720.8 & 44.5 \\ 44.5 & 9.2 \end{pmatrix}$$

	Predicted		
True	Bad	Good	Total
Bad	959	41	1000
Good	49	951	1000
Total	1008	992	2000

Table 1: True vs. Predicted class of loans: quadratic discriminant

	Reject	Accept	Total
Bad	966	34	1000
Good	240	760	1000
Total	1206	794	2000

Table 2: True class vs. Accept/Reject

The parameter estimates resulting from the particular sample drawn are $\hat{\pi}_b = \hat{\pi}_g = 0.5$,

$$\hat{\mu}_b = \begin{pmatrix} 98.0 \\ 15.5 \end{pmatrix} \qquad \hat{\mu}_g = \begin{pmatrix} 135.8 \\ 9.1 \end{pmatrix}$$

and for the covariance matrices

$$\hat{\Sigma}_b = \begin{pmatrix} 553.0 & -38.1 \\ -38.1 & 3.6 \end{pmatrix} \qquad \hat{\Sigma}_g = \begin{pmatrix} 734.0 & 45.0 \\ 45.0 & 9.3 \end{pmatrix}$$

Since within each class, the attributes are normally distributed, with unequal covariance matrices, a quadratic discriminant function is optimal. Quadratic discriminant analysis was performed on the complete sample. The in-sample prediction performance of the resulting function is summarized in table 1. Overall, 95.5% of the observations is classified correctly. This classification result can be obtained if the correct class of all loan applications is known, which is not the case since part of the loan applications is rejected.

For each application the following score is calculated

$$S = 0.16x_1 - x_2$$

If $S > 10$, the loan is accepted, otherwise it is rejected. This score function represents the acceptance policy of the bank, which may have been determined by loan officers or by analysis of historical data. On the particular sample drawn, this yields the results as shown in table 2. The fraction of accepted loans that turns out to be bad is quite low, $34/794 \approx 4.3\%$. On the other hand, quite a number of the rejected loans are in fact good loans, $240/1206 \approx 20\%$. The predictive accuracy of the quadratic discriminant function (estimated on the complete sample) on the rejected loans is summarized in table 3. The overall accuracy of the quadratic discriminant function on the rejected loans is approximately 95.4%.

Next, we removed the class label of the rejected loans for the analysis that follows. This corresponds to the

	Predicted		
True	Bad	Good	Total
Bad	957	9	966
Good	46	194	240
Total	1003	203	1206

Table 3: True vs. Predicted class of rejected loans: quadratic discriminant

situation that the bank faces in practice. It is interesting to obtain an estimate of how many loans are incorrectly rejected, and perhaps more importantly, which of the rejected loans are in fact very likely to be good risks. We try to answer these questions in the subsequent analysis.

Application of EM algorithm

In order to estimate the class-conditional densities of good and bad loans, using the partially classified data, we use a program for fitting a mixture of normal distributions with arbitrary covariance matrices. The number of classified cases from each group must be larger than the number of attributes p, in order to avoid the occurrence of singularities in the likelihood.

The program used has been taken from (McLachlan & Basford 1988), pages 218–224. The program uses EM to find maximum likelihood estimates for the component parameters and the mixing proportions. Under the normality assumption, the likelihood estimates of π_i, μ_i, and Σ_i satisfy

$$\hat{\pi}_i = \frac{\sum_{j=1}^{m} \hat{\tau}_{ij} + \sum_{j=m+1}^{m+n} z_{ij}}{m+n}$$

and

$$\hat{\mu}_i = \frac{\sum_{j=1}^{m} \hat{\tau}_{ij}\mathbf{x}_j + \sum_{j=m+1}^{m+n} z_{ij}\mathbf{x}_j}{\sum_{j=1}^{m} \hat{\tau}_{ij} + \sum_{j=m+1}^{m+n} z_{ij}}$$

and finally

$$\hat{\Sigma}_i = \frac{\sum_{j=1}^{m} \hat{\tau}_{ij}(\mathbf{x}_j - \hat{\mu}_i)(\mathbf{x}_j - \hat{\mu}_i)'}{\sum_{j=1}^{m} \hat{\tau}_{ij} + \sum_{j=m+1}^{m+n} z_{ij}} + \frac{\sum_{j=m+1}^{m+n} z_{ij}(\mathbf{x}_j - \hat{\mu}_i)(\mathbf{x}_j - \hat{\mu}_i)'}{\sum_{j=1}^{m} \hat{\tau}_{ij} + \sum_{j=m+1}^{m+n} z_{ij}}$$

Where, because of the normality assumption the posterior probability τ_{ij} that $\mathbf{x}_j$ belongs to C_i is obtained by substituting

$$(2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp\{-1/2(\mathbf{x}_j - \mu_i)'\Sigma_i^{-1}(\mathbf{x}_j - \mu_i)\}$$

for $f_i(\mathbf{x}_j)$ into equation 1. These equations are solved by substituting the initial estimates into the right-hand sides of the equations to obtain new estimates, which are substituted into the right-hand sides, and so on, until convergence.

In case of random classification, the classified observations can be used to choose reasonable initial estimates for the mixing proportions, mean vectors and covariance matrices. This procedure is not the most sensible here since the classified observations are not a random sample from mixture C. Therefore, the intial estimates were determined as follows. The size of the total sample of loan applications equals $m+n$. There are n accepted applications and m rejected applications. The class label of the rejected applications is missing.

Furthermore, the accepted loans can be subdivided in bad (b) and good (g) loans ($n = b+g$). Then the initial estimates for the mixing proportions are chosen as follows

$$\hat{\pi}_b^{(0)} = (b+m)/(n+m), \qquad \hat{\pi}_g^{(0)} = g/(n+m)$$

The initial estimates of $\hat{\mu}_b, \hat{\mu}_g, \hat{\Sigma}_b$ and $\hat{\Sigma}_g$ are also calculated from $b+m$ and g respectively. The rationale is that one simply assumes that all rejects are in fact bad loans (which is the reason they were rejected in the first place).

Thus we get the following initial estimates for the mixing proportions

$$\hat{\pi}_b^{(0)} = \frac{1206+34}{2000} = 0.62, \qquad \hat{\pi}_g^{(0)} = \frac{760}{2000} = 0.38$$

For the covariance matrices we have

$$\hat{\Sigma}_b^{(0)} = \begin{pmatrix} 562.9 & -30.1 \\ -30.1 & 9.7 \end{pmatrix} \qquad \hat{\Sigma}_g^{(0)} = \begin{pmatrix} 552.7 & 48.4 \\ 48.4 & 8.7 \end{pmatrix}$$

Finally, for the group means we get

$$\hat{\mu}_b^{(0)} = \begin{pmatrix} 100.4 \\ 14.4 \end{pmatrix} \qquad \hat{\mu}_g^{(0)} = \begin{pmatrix} 143.8 \\ 8.9 \end{pmatrix}$$

After 15 iterations the algorithm converged, yielding the following parameter estimates. For the mixing proportions

$$\hat{\pi}_b^{*} = 0.507, \qquad \hat{\pi}_g^{*} = 0.493$$

For the covariance matrices

$$\hat{\Sigma}_b^{*} = \begin{pmatrix} 552.8 & -38.7 \\ -38.7 & 3.7 \end{pmatrix} \qquad \hat{\Sigma}_g^{*} = \begin{pmatrix} 739.7 & 46.2 \\ 46.2 & 9.3 \end{pmatrix}$$

The estimates for the group means are

$$\hat{\mu}_b^{*} = \begin{pmatrix} 98.3 \\ 15.4 \end{pmatrix} \qquad \hat{\mu}_g^{*} = \begin{pmatrix} 136.0 \\ 9.1 \end{pmatrix}$$

To test the sensitivity of the solution to the initial estimates, we performed the same analysis with initial estimates determined on the classified observations only. In that case, the initial estimates for the mixing proportions are way of,

$$\hat{\pi}_b^{(0)} = 34/794 = 0.043, \qquad \hat{\pi}_g^{(0)} = 760/794 = 0.957$$

The initial estimates for the means and covariance of good loans are near the true value, but for the bad

	Predicted		
True	Bad	Good	Total
Bad	959	7	966
Good	51	189	240
Total	1010	196	1206

Table 4: True vs. Predicted class of rejected loans: mixture model

	Predicted		
True	Bad	Good	Total
Bad	966	0	966
Good	169	71	240
Total	1135	71	1206

Table 5: True vs. Predicted class of rejected loans: linear case

loans they are strongly biased because of the selection effect. For the covariance matrices we have

$$\hat{\Sigma}_b^{(0)} = \begin{pmatrix} 135.5 & -2.6 \\ -2.6 & 0.9 \end{pmatrix} \quad \hat{\Sigma}_g^{(0)} = \begin{pmatrix} 552.7 & 48.4 \\ 48.4 & 8.7 \end{pmatrix}$$

Finally, for the group means we get

$$\hat{\mu}_b^{(0)} = \begin{pmatrix} 149.0 \\ 11.6 \end{pmatrix} \qquad \hat{\mu}_g^{(0)} = \begin{pmatrix} 143.8 \\ 8.9 \end{pmatrix}$$

After 21 iterations the algorithm converged to the same solution as obtained in the previous analysis.

Most relevant is how well the resulting discriminant rule classifies the rejects. This is summarized in table 4. The proportion of correct classifications of rejects is about 95.2%, which is only slightly worse than the performance of the quadratic discriminant function trained on the complete sample. Perhaps more importantly, 189 of the 196 cases predicted to be good loans are in fact good loans, which is approximately 96.4%.

Comparison to sample-based approaches

To illustrate the severe bias that sample based approaches may suffer from, we have performed linear and quadratic discriminant analysis on the accepted loans, as if it were a random sample from the population of loan applicants. The results for the linear case are summarized in table 5, for the quadratic case in table 6. In both cases, the class priors (mixing proportions) were taken to be the same as the initial estimates in the EM algorithm.

For the linear case, the overall percentage of correct classifications is about 86%. Of the 240 good loans, only 71 are predicted to be good. For the quadratic case, the overall result is about 69.8% correct classifications. Out of 966 bad rejects, 285 (± 30%) are predicted to be good loans.

	Predicted		
True	Bad	Good	Total
Bad	681	285	966
Good	79	161	240
Total	760	446	1206

Table 6: True vs. Predicted class of rejected loans: quadratic case

Discussion

The example discussed has admittedly been constructed to show the possible benefits of the mixture modeling approach to biased data, or more specifically non-random partially classified data. More realistic case studies should be performed to test the practical usefulness of this approach.

One may for example consider situations where the attributes are not real valued, but categorical. In that case mixtures of bernouilli or multinomial components may be used. Problems with mixed real, binary and categorical attributes can be analysed using joint densities with mixed components of the three types (Ghahramani & Jordan 1994; Lawrence & Krzanowski 1996).

One may also consider situations where the class-conditional densities are themselves mixtures. Preliminary data analysis may reveal that a class-conditional density should be modeled as a mixture of component densities rather than a single density. This is for example done in (McLachlan & Gordon 1989) for an application in medicine. A semi-parametric approach is taken in (Hastie & Tibshirani 1996), where a method and algorithm for discriminant analysis by normal mixtures is described. The algorithm can be adjusted quite easily to allow for missing class labels.

Finally, one may also consider situations where missing values occur in the attributes, and not just in the class label. This situation is analysed in (Little & Rubin 1987) and (Ghahramani & Jordan 1994).

Although each of these extensions will obviously complicate the analysis to a certain extent, they can all be handled within the mixture modeling framework using EM to obtain maximum likelihood estimates of the relevant parameters.

Conclusion

The mixture modeling approach is well suited to analyse non-random partially classified data, and avoids the bias that sample-based approaches have. This approach may be applied to any partially classified data set, where some kind of "screening" mechanism determines which observation is classified and which not. Furthermore the mixture modeling framework using EM is flexible enough to allow for non-normal data, class-conditional densities that are mixtures, and missing attribute values. This flexibility indicates that it

is also applicable to real-world messy data sets. Performance of realistic case studies, and development of suitable software to perform these studies, must substantiate this claim in the future.

References

Everitt, B., and Hand, D. 1981. *Finite mixture distributions.* London: Chapman and Hall.

Feelders, A.; le Loux, A.; and Zand, J. v. t. 1995. Data mining for loan evaluation at ABN AMRO: a case study. In Fayyad, U., and Uthurusamy, R., eds., *Proceedings of KDD-95*, 106–111. AAAI Press.

Ghahramani, Z., and Jordan, M. I. 1994. Supervised learning from incomplete data via an EM approach. In Cowan, J.; Tesauro, G.; and Alspector, J., eds., *Advances in Neural Information Processing Systems 6.* San Fransisco, CA: Morgan Kaufmann.

Hastie, T., and Tibshirani, R. 1996. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society B* 58(1):155–176.

Hsia, D. 1978. Credit scoring and the equal credit opportunity act. *The Hastings law journal* 30:371–448.

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 Gordon, R. 1989. Mixture models for partially unclassified data: a case study of renal venous renin in hypertension. *Statistics in Medicine* 8:1291–1300.

McLachlan, G. J. 1992. *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Titterington, D.; Smith, A.; and Makov, U. 1985. *Statistical analysis of finite mixture distributions.* Chichester: John Wiley & Sons.