Identification of Credit Risk Based on Cluster Analysis of Account Behaviours

Maha Bakoben\textsuperscript{1,2}, Tony Bellotti\textsuperscript{1}, and Niall Adams\textsuperscript{1,3}

\textsuperscript{1}Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom
\textsuperscript{2}Statistics Department, King Abdulaziz University, P.O.Box 80200, Jeddah 21589, Saudi Arabia
\textsuperscript{3}Heilbronn Institute for Mathematical Research, University of Bristol, Bristol BS8 9AG, United Kingdom

Abstract

Assessment of risk levels for existing credit accounts is important to the implementation of bank policies and offering financial products. This paper uses cluster analysis of behaviour of credit card accounts to help assess credit risk level. Account behaviour is modelled parametrically and we then implement the behavioural cluster analysis using a recently proposed dissimilarity measure of statistical model parameters. The advantage of this new measure is the explicit exploitation of uncertainty associated with parameters estimated from statistical models. Interesting clusters of real credit card behaviours data are obtained, in addition to superior prediction and forecasting of account default based on the clustering outcomes.

Keywords: Behavioural credit scoring; credit behaviour clusters; clustering parameter uncertainty; default prediction.

1. Introduction

Behavioural credit scorecards can be defined as statistical models of customer behaviour, i.e. card usage and repayments, over time (Till and Hand, 2003). The aim of these models is to identify which of the existing customers may experience difficulty paying back the loan (Thomas et al., 2002). Identification of distinct risk levels might support operating decisions with regard to increasing credit limits or offering a financial product (Thomas et al., 2002; Till and Hand, 2003).

In this paper, we present a new methodology for identifying different risk groups based on the available data of customer behaviour. The method aims to assign credit card accounts to clusters such that the behaviours of accounts in the same cluster are similar. This cluster analysis can be used as a tool for building different behavioural scorecards or developing distinct marketing strategies for groups of accounts.

A typical interest in retail banking is predicting the probability of a customer not being able to make the minimum amount of the agreed monthly repayment for three consecutive months, referred to as ‘default’. A default prediction model based on aggregate summaries of account behaviour is traditionally used in behavioural credit scoring (Thomas, 2009). The aggregate summary can be defined as a statistic which describes the time series in a single value such as the mean or median. This approach might result in loss of valuable information inherent in the dynamic behaviour data. We introduce a new approach to the development of the default prediction and forecasting models. This approach utilises the outcomes of the cluster analysis of the credit behaviours. To distinguish between a prediction model and a forecasting model, the former predicts the default status over an observed behaviour period, whereas the forecasting model predicts the default status at a future period after observing the behaviour.

A fundamental aspect in clustering methods is the specification of a dissimilarity measure that is appropriate for the data. As behaviours can be considered as time series, serial dependence needs to be considered in the definition of the dissimilarity measure. Two stages for defining the dissimilarity between pairs of time series objects will be considered in this paper. The first is fitting a multivariate time series model to express the dynamic characteristics of the account. This stage reduces the dimension of the data by providing the model parameters as a summary, in addition it makes the dissimilarity comparison feasible between credit accounts with different numbers of transactions.

The second stage computes the dissimilarity between confidence regions of the model parameters. Since the objects being clustered are parameters of a statistical model, they exhibit statistical uncertainty. Notably, this uncertainty is driven by the amount of data used to estimate the model. This uncertainty-aware dissimilarity measure, introduced in Bakoben et al. (2016), is intended to account for this uncertainty in the estimated model parameters. The consideration of such uncertainty produces more reliable clusters than clusters based only on parameter estimates. Beside the behavioural credit scoring application, this methodology can be applied to other application domains and any model structure with measurable errors. In this paper we use vector autoregression, but the underlying method is more general than this and allows for any distributional form, not just normality in the error term. In addition, it has shown
better performance compared to traditional dissimilarity measures such as Mahalanobis-distance, which has a theoretical basis assuming normality in the distributions (Bakoben et al., 2016). The flowchart in Figure (1) shows the general methodology which can be used in other applications. We are not aware of any literature that has addressed the cluster analysis of credit behaviours using the time series clustering approaches that are described in this paper.

A previous study that considers differentiating credit accounts based on their behaviours is given by Hsieh (2004). The author applies a self-organizing map neural network for the purpose of identifying distinct profitable groups based on transaction variables including repayment behaviour. This was based on aggregate values of the account behaviours. In another study by Wei and Mingshu (2013), credit card accounts were divided into clusters based on an objective cluster analysis (OCA) for application and behavioural variables. The standard Euclidean distance was used for dissimilarity computations. Then a neural network was created for each cluster to predict the ‘good’ and ‘bad’ accounts. Again this study considers aggregate representations of behaviours which may result in loss of valuable information about the dynamic changes in account behaviours over time.

One of the earliest studies concerned with cluster analysis of credit account behaviours is the paper by Edelman (1992) which performs the clustering on delinquency count. In Edelman (1992), the overall total delinquencies of accounts observed at each month over a two-year period were clustered using the \( k \)-medoids clustering method with Euclidean distance, where the main purpose of the analysis is to identify clusters of months or a combination of months and products. Adams et al. (2001) divide credit card accounts into two clusters on the basis of least squares parameter estimates of a linear regression model. The linear model is fitted to the cumulative numbers of missed repayments over a twelve months period. Again with respect to credit card behaviours, Till and Hand (2003) cluster delinquency counts into groups based on Euclidean distance of the linear slope of a polynomial model for the delinquency count over time. Note that those papers were concerned with clustering a univariate behaviour while the clustering approach presented in this paper is applicable to multiple behaviours.

This paper is organised as follows. Section 2 describes the available real data set of credit card account behaviours. Section 3 illustrates the two stages of the cluster analysis. Section 4 introduces the prediction and forecasting models of default. The empirical results of clustering account behaviours are presented in Section 5. Section 6 and 7 show the outcomes of the default prediction and forecasting models, respectively. Finally, Section 8 summarises the work of this study.

2. Data set

Banks are regularly interested in behaviour analysis of small groups of accounts as well as large groups. In this study, the credit card data set includes monthly behaviours for 494 active accounts from an anonymous financial institution in the UK for a maximum period of 37 months from June 2008 to June 2011. The data has full records of 37 months for 7% of the accounts and the mean length of the remaining behaviour records is 29 months. The objective with this data is to assign customers into clusters based on their monthly behaviours and we aim to discriminate between high and low risk customers.

For a single customer \( s \), we denote the corresponding behavioural credit account by \( Y_s \) and its length by \( T_s \). Each account has the following characteristics: \( y_{s, \text{repay}} \) denotes a vector of the monthly repayment amount made by the customer, \( y_{s, \text{bal}} \) denotes a vector of total balance on the account at the end of each month and \( y_{s, \text{cl}} \) denotes a vector of the monthly credit limit which is static for most customers. The latter two behaviours will be considered indirectly through a new behaviour vector, \( y_{s, \text{ut}} \), that is called utilisation rate; the ratio of total balance to credit limit,

\[
\text{utilisation rate} = \frac{\text{total balance}}{\text{credit limit}},
\]

where the value of utilisation rate should be between 0 and 1. However, there are cases when this rate goes below or over the standard range. For example, customers overpay their loans (i.e. \( y_{s, \text{bal}} < 0 \)) or the total balance exceeds the credit limit (i.e. \( y_{s, \text{bal}} > y_{s, \text{cl}} \)). The mean of utilisation rate in the credit data is 0.6355. The minimum and maximum values are \(-7.0990\) and 3.5600, respectively.

Another characteristic in the credit card data set includes delinquency count – a cumulative number of missed number of payment. This ranges between 0 and 12. In addition, a default status, \( x_{s}(t) \in \{0,1\} \) for \( t = 1, \ldots, T_s \), is defined based on the delinquency count. If a customer misses several consecutive payments (usually three) by time \( t \), then the default status \( x_{s}(t) = 1 \) otherwise \( x_{s}(t) = 0 \). The proportion of defaults in the data is 27.33% (135 cases) and 72.67% (359 cases) are non-defaults. This default rate is high because the data provider had already under-sampled the non-default cases prior to delivery of the data. This under-sampling is a common practice in credit scoring to deal with class imbalance.

Figure 1: The general methodology of utilising uncertainty clustering in prediction models.
The repayment amount and utilisation rate are typical variables used in behavioural scorecards, hence they will be used to build the clusters.

For the purpose of evaluation, a proportion of 60% training data from the credit accounts is used for the construction of the model and 40% of the accounts are held out for testing.

3. Clustering method

This section describes the two stages of defining dissimilarity between credit card behaviours. Section 3.1 describes the first stage that is the time series modelling of customer behaviour, Section 3.2 describes the conventional approach for defining dissimilarity between model parameters and Section 3.3 illustrates the inclusion of parameter uncertainty in the dissimilarity measure for the cluster analysis of credit card behaviours.

3.1. Time series modelling

First, we reduce the dimension of the observed behaviours to make the dissimilarity comparison feasible between the credit accounts. An aspect that should be considered in the reduction process is the relationship between several variables over time.

For a single account \( s \), the monthly repayment behaviour, \( y_{s,\text{repay}} = [y_{s,\text{repay}}(t = 1), \ldots, y_{s,\text{repay}}(t = T_s)]^T \), and the utilisation rate behaviour, \( y_{s,\text{ut}} = [y_{s,\text{ut}}(t = 1), \ldots, y_{s,\text{ut}}(t = T_s)]^T \), can be described by a bivariate vector autoregression (VAR) model which is an appropriate choice as it captures the dynamic features of the data. We follow the same approach in this study.

For a single account \( s \), the monthly repayment behaviour, \( y_{s,\text{repay}} = [y_{s,\text{repay}}(t = 1), \ldots, y_{s,\text{repay}}(t = T_s)]^T \), and the utilisation rate behaviour, \( y_{s,\text{ut}} = [y_{s,\text{ut}}(t = 1), \ldots, y_{s,\text{ut}}(t = T_s)]^T \), can be described by a bivariate vector autoregression (VAR) model of order one (Lütkepohl, 2005). Here, the choice of best fitted model is not an issue since the associated error estimates are explicitly considered in the clustering process as will be illustrated later in Section 3.3. The bivariate VAR(1) model is given by:

\[
\begin{bmatrix}
  y_{s,\text{repay}}(t) \\
  y_{s,\text{ut}}(t)
\end{bmatrix} = 
\begin{bmatrix}
  \theta_{s,1} & \theta_{s,2} \\
  \theta_{s,3} & \theta_{s,4}
\end{bmatrix} 
\begin{bmatrix}
  y_{s,\text{repay}}(t-1) \\
  y_{s,\text{ut}}(t-1)
\end{bmatrix} + 
\begin{bmatrix}
  u_1(t) \\
  u_2(t)
\end{bmatrix},
\]

(1)

where \( u = [u_1(t), u_2(t)]^T \) is a vector of weakly stationary white noise process, \( u \sim N(0, \Sigma) \). Each equation in a VAR model is estimated separately by an ordinary likelihood estimator (Lütkepohl, 2005).

By fitting the bivariate VAR model of order one to \( N \) behavioural credit accounts, we obtain \( N \) vectors of VAR coefficients \( \theta_s = [\theta_{s,1}, \ldots, \theta_{s,p}]^T \) where in this case \( p = 4 \).

3.2. Conventional clustering approach

As described in Bakoben et al. (2015), Euclidean distance can be computed directly between a pair of VAR coefficients vectors. For two credit account behaviours \( Y_r = [y_{r,\text{repay}}, y_{r,\text{ut}}] \) and \( Y_s = [y_{s,\text{repay}}, y_{s,\text{ut}}] \), Euclidean distance between their corresponding VAR coefficients \( \theta_r = [\theta_{r,1}, \ldots, \theta_{r,p}]^T \) and \( \theta_s = [\theta_{s,1}, \ldots, \theta_{s,p}]^T \) is computed as follows:

\[
d_{\text{euc}}(Y_r, Y_s) = \sqrt{\sum_{i=1}^{p} (\theta_{r,i} - \theta_{s,i})^2}. \quad (2)
\]

3.3. Uncertainty-aware clustering

These estimated VAR parameters are subject to statistical uncertainty. This type of uncertainty can be characterised by the covariance matrix of the estimated parameter vector, denoted by \( \Psi \). Bakoben et al. (2016) proposed an approach for the explicit inclusion of uncertainty in the computation of dissimilarity between data points.

The idea of the new metric is to measure the overlap between \((1 - \alpha)\) confidence regions of VAR coefficients. Each confidence region is represented geometrically by an ellipsoid defined by:

\[
E_s(\theta_s, \Psi_s) : \{(\mathbf{x} - \theta_s)^T (\hat{\Psi}_s)^{-1} (\mathbf{x} - \theta_s) \leq 1\},
\]

where the scalar \( c = \sqrt{pF_{p/(p-1),\alpha}} \), \( p \) is the number of VAR parameters, \( T_s \) is the length of the corresponding credit account, \( F_{p/(p-1),\alpha} \) is the critical value of the F-distribution with degrees of freedom \((p, T_s - p - 1)\) and \( \alpha \) is the significance level.

The ratio of overlap between each pair of ellipsoids \((E_r, E_s)\) is given by

\[
R_{r,s} = \frac{V_{E_r \cap E_s}}{V_{E_r} + V_{E_s} - V_{E_r \cap E_s}}, \quad r \neq s, \quad V_{E_r}, V_{E_s} > 0,
\]

(3)

where the hyper-volumes of ellipsoids \( V_{E_r} \) and \( V_{E_s} \) are computed by the mathematical formula \( V_{E_r} = \frac{2^n}{(2\pi)^{n/2}} \left(\frac{|\hat{\Psi}_r|}{\sqrt{n}}\right)^{n/2} \) (Friendly et al., 2013). The volume of the overlap region, \( V_{E_r \cap E_s} \), is estimated by Monte Carlo simulations (Robert and Casella, 2010) as there is no-closed formula for the overlap volume. Then, the dissimilarity between confidence regions of VAR coefficients is defined by

\[
d_{\text{ell}}(Y_r, Y_s) = 1 - R_{r,s}, \quad d_{\text{ell}} \in [0, 1].
\]

(4)

The next step in the credit behaviours cluster analysis is the implementation of the \( k \)-medoids partitioning clustering method that is commonly used in industrial practice (Kaufman and Rousseeuw, 1987, 2008). Each account is assigned to the cluster with the closest medoid \( m \). The \( k \)-medoids method with the uncertainty-aware dissimilarity attempts to identify clusters that minimise the sum of this distance to the medoids \( m_1, \ldots, m_k \).

For a credit account \( Y_s \), a vector of cluster allocation \( z_s = (z_{s,1}, \ldots, z_{s,k}) \), is defined where each element in the vector, \( z_{s,l} \) for \( l = 1, \ldots, k \), is given by:

\[
z_{s,l} = \begin{cases} 
  1 & \text{if } l = \text{argmin}_l d_{\text{ell}}(Y_s, Y_{m_l}) \\
  0 & \text{Otherwise.}
\end{cases}
\]

(5)
4. The use of clusters for model predictions and forecasts

We develop a model to predict default. This model will also be used to evaluate the clustering performance. Here, we introduce a binary response variable, \( \tilde{x}_s \), which indicates whether an account has ever defaulted or not. This binary value for an account \( s \) is measured over the available account’s period \( [t = 1, \ldots , t = T_s] \) as follows:

\[
\tilde{x}_s = \max[x_{s}(t = 1), \ldots , x_{s}(t = T_s)].
\]

The default status is predicted based on the cluster assignment that is an explanatory variable in the logistic regression model:

\[
p(\tilde{x}_s = 1|z_s) = \frac{e^{\beta_0 + \sum_{j=1}^{s} \beta_j z_{s,j}}}{1 + e^{\beta_0 + \sum_{j=1}^{s} \beta_j z_{s,j}}}. \tag{6}
\]

Equation 6 is also used for forecasting. Other explanatory variables can be included in the model but in this paper we intend to isolate the effect of the cluster assignment.

In the forecasting model, the cluster analysis is performed on the first 2/3 of each account profile and the forecast default is measured over the last 1/3 period. This is due to variable lengths of the available credit account profiles, hence choosing specific lengths for the observation and forecast periods is not reasonable as the profile length of an account might be less than the specified observation period. Figure 2 illustrates the observation and forecast period in the default forecasting model. Note that some time-window after credit card origination is required to allow for measurement in observation period (e.g. \( t = 13 \) to \( t = 24 \) in Figure 2).

<table>
<thead>
<tr>
<th>( t = 1, \ldots , 12 )</th>
<th>( t = 13, \ldots , 24 )</th>
<th>( t = 25, \ldots , 37 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation period</td>
<td>Forecasting period</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Illustration for observation and forecasting period in default forecasting model.

In order to evaluate the new default model, we compare its performance to the conventional aggregate model. This models the default status with aggregate representations of the time series defined here by a vector \( g_s = (\tilde{y}_{s,\text{repay}}, \tilde{y}_{s,\text{ut}})^T \) which consists of the mean values of the univariate time series \( y_{s,\text{repay}} \) and \( y_{s,\text{ut}} \). The aggregate repayment behaviour is given by

\[
\tilde{y}_{s,\text{repay}} = \frac{\sum_{t=1}^{T_s} y_{s,\text{repay}}(t)}{T_s}. \tag{7}
\]

Similarly, the aggregate utilisation rate behaviour is computed. Then, the aggregate default model is defined by

\[
p(\tilde{x}_s = 1|g_s) = \frac{e^{\beta_0 + \beta g_s}}{1 + e^{\beta_0 + \beta g_s}}, \tag{8}
\]

where \( \beta \) is a 2-dimensional parameter vector for the aggregate representations \( g_s \).

The prediction and forecasting performance of default models are evaluated by the following common assessment criteria: the H-measure (Hand, 2009), Kolmogorov-Smirnov statistic (Duda et al., 2001), Gini-index (Hastie et al., 2009) and area under the receiver-operating characteristic curve (AUC) (Fawcett, 2006).

5. Results: Clusters of credit card behaviours

We apply the uncertainty-aware clustering method described in Section 3. For this particular application, financial institutions would expect the presence of three groups of credit accounts: good, bad and in-between. Hence, assigning the credit accounts into three clusters \((k = 3)\) is reasonable. In addition, in light of the relatively low sample size, creating more than three clusters can result in small clusters and, in particular, some containing only one account. This can be counterproductive and does not serve the main purpose of the cluster analysis. However, developing some means of discovering the optimal number of clusters in a bigger sample would be interesting further work.

The proportions of credit accounts in the three clusters are presented in Table 1. Table 1 also shows the cluster outcomes based on clustering using the standard Euclidean distance. In comparison to the ellipsoid based clusters, Euclidean distance tends to create one cluster that includes a large proportion of accounts whereas the other clusters include a small proportion of the credit account sample. For example, cluster \( C_1 \) comprises 62% of the total accounts. This demonstrates the importance of incorporating uncertainty in clustering.

<table>
<thead>
<tr>
<th>( k )-medoids cluster #</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{\text{ell}} )</td>
<td>244(50%)</td>
<td>115(23%)</td>
<td>135(27%)</td>
</tr>
<tr>
<td>( d_{\text{eucl}} )</td>
<td>307(62%)</td>
<td>144(29%)</td>
<td>43(9%)</td>
</tr>
</tbody>
</table>

**Repayment amount, credit limit and total balance behaviours**

The boxplots in Figure 3 show representations of the account behaviours in data space on the basis of the outcomes of uncertainty-aware clustering which was performed on the parameter space. These plots represent the logarithm of the behaviour sample means for the account profiles separately for clusters \( C_1 \), \( C_2 \) and \( C_3 \). Interesting variations in account behaviours among the three clusters are shown. The credit accounts belonging to cluster \( C_2 \) seem to make low payments compared to accounts in the other two clusters. This amount is slightly larger in cluster \( C_1 \) than \( C_2 \). As shown in the second plot, the highest credit limit seems to be for the accounts assigned to cluster \( C_1 \), whereas almost equal median values of the credit limits are observed for the accounts in clusters \( C_2 \) and \( C_3 \).
Although the median of credit limits for the accounts in cluster \( C_3 \) is lower than the median for accounts in \( C_1 \), both groups seem to have equal amount of outstanding debt (Total balance). That might be because they spend equal amount of money or members of cluster \( C_3 \) are not paying their debt or paying only small amount of the debt. This information can be explored by comparing the boxplots of the total balance and credit limits between the three clusters. Additionally, a few extremely low outstanding amounts in \( C_1 \) and high outstanding amounts in \( C_3 \) are clearly observed.

**Delinquency Counts**

In this section, we explore the delinquency behaviour in the obtained clusters. Recall this variable was not included in the clustering process but the cluster analysis of VAR parameters revealed interesting aspects of the delinquency behaviour as shown in Figure 4. This figure represents the sample means of delinquency for the credit accounts within each cluster at each month from \( t = 1 \) to \( t = 37 \).

From the left plot in Figure 4, the credit accounts in the three clusters might be described as those accounts who never default as the delinquency count is always less than 2. These accounts were assigned to cluster \( C_1 \). Also accounts in \( C_2 \) never defaulted and the mean of the delinquency count for this group is less than the mean of delinquency for accounts in cluster \( C_1 \) particularly when \( t > 20 \). In contrast, the last cluster \( C_3 \) includes those whose delinquency count is gradually increasing over their profile period and consequently default. Thus, cluster \( C_3 \) can be considered as the highest risk group compared to the other two clusters.

Looking at the delinquency plot obtained from the clustering approach that utilises Euclidean distance (right plot in Figure 4), the general structure seems to be similar to the structure obtained using the ellipsoid dissimilarity measure. However, the Euclidean based clustering approach seems to assign some of the high risk accounts to the other clusters. This is clearly apparent by comparing the overall means of the delinquency over time in the high risk cluster in Euclidean clustering and these measured based on the clustering outcomes of the ellipsoid dissimilarity measure.

**6. Default prediction model**

In this section, we present the result of fitting a logistic regression model for default based on the outcomes of clustering VAR model parameters taking into account the associated error estimates. The default status is computed over the available profile period of each account. This prediction model was previously defined in Equation 6, in which the cluster assignment is a predictor variable for the binary default status. In this section, the prediction model is fitted to a training sample of 296 credit accounts.

This model is compared to the default prediction model in which cluster analysis was performed without the consideration of the VAR parameter errors. The performance is measured based on 1000 different splits of training/test samples. Each test sample has 198 credit accounts.

Table 2 displays the frequencies and proportions of default/non-default across clusters for one particular sampling of training data, but are typical of others. The results of ellipsoid dissimilarity measure show the proportion of default in cluster \( C_3 \) is relatively higher than the other two clusters, cluster \( C_1 \) comes next and the lowest proportion is accounted for cluster \( C_2 \).

Table 2: Frequencies of default/non-default status in a training sample. Clusters are obtained using the \( k \)-medoids clustering method with the ellipsoid dissimilarity measure \( d_{ell} \) and Euclidean distance \( d_{euc} \).

<table>
<thead>
<tr>
<th>( k )-medoids clusters</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{ell} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-default(( \bar{x} = 0 ))</td>
<td>127(43%)</td>
<td>62(21%)</td>
<td>29(10%)</td>
</tr>
<tr>
<td>default(( \bar{x} = 1 ))</td>
<td>19(6%)</td>
<td>8(3%)</td>
<td>51(17%)</td>
</tr>
<tr>
<td>( d_{euc} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-default(( \bar{x} = 0 ))</td>
<td>117(40%)</td>
<td>76(26%)</td>
<td>25(9%)</td>
</tr>
<tr>
<td>default(( \bar{x} = 1 ))</td>
<td>67(23%)</td>
<td>11(4%)</td>
<td>0(0%)</td>
</tr>
</tbody>
</table>

Table 3 reports the coefficient estimates of the default prediction models for the \( k \)-medoids clustering outcomes. It is interesting to find that the cluster assignment is statistically significant for predicting the default status in most cases. Different signs for the influence of the cluster assignment on the default status indicates the proportion of the default class in one of the clusters is higher than in other clusters relative to \( C_1 \). In addition, Table 4 shows this effect relative to \( C_2 \). A noticeable high error is observed for cluster \( C_3 \) when clustering using Euclidean distance. This is due to the small number of objects in this cluster of which none are from the default class (see Table 2).

A comparison to the aggregate model (Equation 8) for the prediction performance of account defaults is reported in Table 5. It is interesting to find that including the uncertainty of the statistical model parameters in the cluster analysis improved the prediction performance of the default status with AUC value of 0.7637 (s.e. 0.0397). The prediction model using the cluster assignment based on the ellipsoid dissimilarity measure also performs well in comparison to the aggregate model, AUC 0.5310 (s.e. 0.0450). The low performance of the aggregate model might be an indication of nonlinear relationships. The performance of that model is consistent with results presented in Bakoben (2016) on a larger sample and different forms of aggregate representations.

Additional models are created in an attempt to improve the default prediction performance. Both the cluster assignment and the aggregate behaviours are included as predictors in the logistic regression model. Adding the cluster assignment variable to the aggregate model shows a remarkable improvement in the default prediction performance and the uncertainty clustering made an even bigger improvement to the aggregate model.
Figure 3: Behaviours of accounts at clusters $C_1$, $C_2$ and $C_3$. The assignment of accounts in the clusters is obtained using the $k$-medoids method with the ellipsoid dissimilarity $d_{\text{ell}}$ applied to VAR parameter estimates for credit account behaviours data set. These behaviours are measured on different scales.

Figure 4: The means of delinquency counts in clusters $C_1$, $C_2$ and $C_3$ over the available 37 months of account records. Clusters are obtained using the $k$-medoids method with the ellipsoid dissimilarity $d_{\text{ell}}$ (left plot) and Euclidean distance $d_{\text{euc}}$ (right plot). The cluster analysis was applied to the VAR parameters of 494 accounts.
Table 3: Coefficient estimates of the logistic regression models for predicting account defaults based on cluster assignments (C1 is the base category). Cluster analysis is performed using the k-medoids method with the proposed ellipsoid dissimilarity measure \( d_{\text{ell}} \) and Euclidean distance \( d_{\text{euc}} \). The regression models are built on a training sample.

| \( k \)-medoids method | Estimate | Std. Error | z value | \( p(|z|) \) |
|-------------------------|----------|------------|---------|-----------------|
| (a) Ellipsoid dissimilarity \( d_{\text{ell}} \) | | | | |
| Intercept | \(-1.8307\) | \(0.2351\) | \(-7.7882\) | \(6.80 \times 10^{-15}\) |
| \( C_1 \) | - | - | - | - |
| \( C_2 \) | \(-0.5207\) | \(0.4876\) | \(-1.0678\) | \(0.2856\) |
| \( C_3 \) | \(2.4060\) | \(0.3363\) | \(7.1536\) | \(8.45 \times 10^{-13}\) |
| (b) Euclidean distance \( d_{\text{euc}} \) | | | | |
| Intercept | \(-0.5017\) | \(0.1560\) | \(-3.2167\) | \(0.0013\) |
| \( C_1 \) | - | - | - | - |
| \( C_2 \) | \(-1.6705\) | \(0.3848\) | \(-4.3409\) | \(1.42 \times 10^{-5}\) |
| \( C_3 \) | \(-17.0644\) | \(688.6826\) | \(-0.0248\) | \(0.9802\) |

Table 4: Coefficient estimates of the default prediction model relative to clusters \( C_1 \) and \( C_2 \). The coefficients above the diagonal are for clusters using ellipsoid dissimilarity and the coefficients below diagonal are for clusters using Euclidean. The ** indicates a significant coefficient at \( \alpha = 0.01 \).

<table>
<thead>
<tr>
<th>Clusters</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>-</td>
<td>(-0.5207)</td>
<td>(2.4060)**</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>(-1.6705)**</td>
<td>-</td>
<td>(2.9267)**</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>(-17.0644)</td>
<td>(-15.3938)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5: Performance assessments for default prediction models based on the cluster assignment obtained from the \( k \)-medoids clustering method using the Ellipsoid dissimilarity \( d_{\text{ell}} \) and the standard Euclidean distance \( d_{\text{euc}} \). These models are compared with the default prediction model based on aggregate means of the behaviours. The means of the assessment measures are computed over 1000 different hold-out test samples.

<table>
<thead>
<tr>
<th>d</th>
<th>H-measure</th>
<th>KS</th>
<th>Gini</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{\text{ell}} )</td>
<td>0.3748</td>
<td>0.5443</td>
<td>0.5273</td>
<td>0.7637</td>
</tr>
<tr>
<td>( d_{\text{euc}} )</td>
<td>0.1416</td>
<td>0.3405</td>
<td>0.3563</td>
<td>0.6781</td>
</tr>
<tr>
<td>Aggregate model</td>
<td>0.0573</td>
<td>0.1569</td>
<td>0.0620</td>
<td>0.5310</td>
</tr>
<tr>
<td>Aggregate model + ( d_{\text{ell}} )</td>
<td>0.3962</td>
<td>0.5543</td>
<td>0.5110</td>
<td>0.7555</td>
</tr>
<tr>
<td>Aggregate model + ( d_{\text{euc}} )</td>
<td>0.1818</td>
<td>0.3679</td>
<td>0.3527</td>
<td>0.6764</td>
</tr>
</tbody>
</table>

7. Default forecasting model

This section focuses on forecasting the default status of the credit accounts over unseen future periods of their profiles. As with the prediction models presented in the previous section, the cluster assignment is used as an explanatory variable in the forecasting models. Recall the clusters are obtained from the observation period, whereas the default is computed over the forecasting period. Table 6 presents the frequencies of default/non-default classes in the training data.

Table 7 reports the coefficient estimates of the logistic regression models for forecasting the default status based on the cluster assignments. These coefficient estimates are relative to the base category (\( C_1 \)). Comparison of the effect relative to \( C_2 \) is shown in Table 8. Interestingly, the coefficient estimates are only significant in the model based on the uncertainty-aware dissimilarity measure. Again, as observed in the prediction model in the previous section, some of the forecasting model coefficients have high standard errors as a result of small samples of the default class.

Table 8 compares the forecasting performance between the proposed forecasting models, where the cluster assignment is the explanatory variable. Similar to the prediction model, the proposed forecasting models are compared to the forecast model on aggregate summaries. The most favourable model is based on clustering VAR parameters with the associated uncertainty. The performance values are reasonable for this particular type of application. The AUC for the best model is 0.7251 (s.e. \(6 \times 10^{-4}\)), whereas the AUC for the model based on Euclidean distance is 0.6123 (s.e. \(4 \times 10^{-4}\)). The forecast model based on aggregate summaries shows the lowest performance, AUC 0.5355 (s.e. \(7 \times 10^{-4}\)).

8. Conclusion

This paper introduced a new behavioural clustering approach that can support the construction of behavioural credit scorecards. In the clustering process, the credit accounts were represented by statistical parameter estimates of their behaviours to represent their associated serial dependence. This results in
Both the prediction and forecasting models based on the ellipsoid distinction are used for prediction and forecasting between the high risk group and low risk group. This models default status with cluster assignments. The uncertainty-aware dissimilarity measure.

Taking into account the uncertainty of the model parameters has revealed interesting behavioural clusters. Although the delinquency behaviour of the accounts was not included in the clustering process, the cluster analysis was able to differentiate delinquency behaviour of the accounts was not included in the samples. The assessment measures are the means over 1000 different hold-out test samples.

In this study, we only considered the popular k-mediods clustering method. However, one path of further research would be to consider alternatives such as mixture models.

This research could be extended by performing the uncertainty cluster analysis on time-windows over the account profiles and study the changes in risk levels over profile history. This is an interesting extension of the study but it requires longer behaviour profiles.

An issue that can be addressed in future work is the increase of the computational complexity of the new metric as the sample size increases. Bakoben et al. (2015) discussed this challenge which involves the computation of the overlap volume between confidence regions using Monte Carlo simulations. An intensive study in Bakoben (2016) has considered an iterative MC procedures to reduce the computational complexity, and further optimisation could be undertaken in further work.

Acknowledgments

This work was supported by King Abdulaziz University scholarship fund.

| k-medoids method | Estimate | Std. Error | z value | p(>|z|) |
|------------------|----------|------------|---------|---------|
| (a) Ellipsoid dissimilarity $d_{ell}$ | | | | |
| Intercept | -3.1209 | 0.4171 | -7.4827 | 7.28e−14 |
| $C_1$ | - | - | - | - |
| $C_2$ | 1.8766 | 0.4463 | 4.2044 | 2.62e−5 |
| $C_3$ | 2.8363 | 0.4420 | 6.1755 | 1.39e−10 |
| (b) Euclidean distance $d_{euc}$ | | | | |
| Intercept | -0.7726 | 0.1026 | -7.5342 | 4.91e−14 |
| $C_1$ | - | - | - | - |
| $C_2$ | -17.7934 | 639.5973 | -0.0278 | 0.9778 |
| $C_3$ | -17.7934 | 1581.9722 | -0.0112 | 0.9910 |

Table 8: Coefficient estimates of the default forecasting model relative to clusters $C_1$ and $C_2$. The coefficients above the diagonal are for cluster using ellipsoid dissimilarity and the coefficients below diagonal are for cluster using Euclidean. The ** indicates a significant coefficient at $\alpha = 0.01$.

Table 9: Performance assessments for default forecasting model based on the cluster assignments obtained from the k-mediods clustering method using the ellipsoid dissimilarity measure $d_{ell}$ and the standard Euclidean distance $d_{euc}$. These models are compared with the default forecasting model based on aggregate means of the behaviours. The assessment measures are the means over 1000 different hold-out test samples.

<table>
<thead>
<tr>
<th></th>
<th>H-measure</th>
<th>KS</th>
<th>Gini</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{ell}$</td>
<td>0.1825</td>
<td>0.3382</td>
<td>0.4502</td>
<td>0.7251</td>
</tr>
<tr>
<td>$d_{euc}$</td>
<td>0.0810</td>
<td>0.2237</td>
<td>0.2246</td>
<td>0.6123</td>
</tr>
<tr>
<td>Aggregate model</td>
<td>0.0276</td>
<td>0.1219</td>
<td>0.0709</td>
<td>0.5355</td>
</tr>
<tr>
<td>Aggregate model + $d_{ell}$</td>
<td>0.1744</td>
<td>0.3230</td>
<td>0.4094</td>
<td>0.7047</td>
</tr>
<tr>
<td>Aggregate model + $d_{euc}$</td>
<td>0.1130</td>
<td>0.2799</td>
<td>0.2172</td>
<td>0.6086</td>
</tr>
</tbody>
</table>

References


