Numbercraft: Optimal Customer Account Classification

David Wood, Kate Woodham, Russel Mills and John Perram*

1 Problem Description

The full, detailed, problem description presented to ESGI37 by Numbercraft is contained in Appendix I, but essentially the problem is one of attempting to extract useful information from a large database of customer details. The toy model which was used during the Study Group to formulate our ideas was one of a phone company holding on file, for each customer, details such as total monthly phone usage, monthly off-peak usage, Internet usage, total cost built up etc. In particular note that there will be strong correlation between a lot of the different data: the first task in any analysis would be to reduce the phase-space by way of establishing statistical correlations by, for example Singular Value Decomposition.

The basic idea is now that we wish to classify customers in some way, for our toy example perhaps as 'heavy users', 'medium users' and 'light users', although it is not at all clear that such labels *can* usefully be applied to the data. We then wish to track how customers move between these classes, the natural method being by transition matrices. In particular wish to try and predict when customers will either move into another class or, more importantly, cease being a customer altogether, becoming 'dead' customers.

The two (related) issues with which we are now concerned are:

- How do we sensibly define the classes from raw data so as to maximise the amount of information (partitions of state space)?
- How do we then set up transition matrices between these classes? In particular we would like these matrices to be strongly diagonal and sparse.

1.1 The data provided

Two sets of data were provided during the week for analysis purposes:

*also contributions during modelling week from Liam Clarke, Robert Ketzscher, Neil Stringfellow and Julie Scarrett

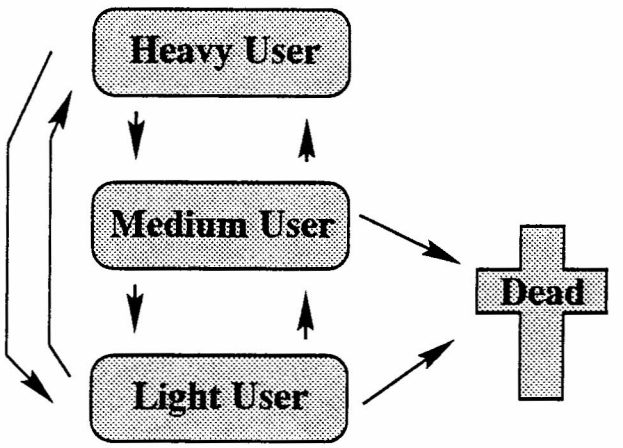

Figure 1: Toy model.

- **Data set 1:** Synthetic Data. Three sets of data, each consisting of 105 time steps, 20 customers, each with 3 state variables. This data was designed so that algorithms which we were developing could be tested.
- **Data set 2:** Real Data (given 3 days later). 14 time steps (monthly), 220 customers, each with 6 state variables. The nature of this data, and what each state variable represented was not known.

1.2 The approaches tried

During the study group several different approaches were considered, which are outlined in more detail later in this report. Briefly however they are:

- The problem is considered in its more abstract setting from a mathematical/AI point of view to highlight the issues involved and to try and make more precise the question with which we are concerned.
- The important question that must be considered is how to define 'classes' of customers, which must be done before we can even begin to consider how a customer moves between such classes. The initial approach outlined here is to start with a large number of arbitrary classes and try to systematically amalgamate classes until we had a manageable (and statistically suitable) number which would give a sparse and strongly diagonal transition matrix.
- We are in particular looking for customers who are about significantly change their behaviour or to stop their use of the service altogether. A statistical analysis of records corresponding to those customers who 'died' versus those who didn't is therefore appropriate in case there is some very simple trait that should be sought in a previously unseen set of data.

- Finally a more novel approach was implemented to try and classify customers by the behaviour of their time-series. A customer who is about to change their customer loyalties would perhaps change their behaviour prior to doing so which would result in a qualitative change in at least some of their time-series. It is not enough just to look for customers who suddenly start using some service less than before, because some customers may do this on a regular basis before increasing usage once more.

2 Some mathematical and AI reflections on the problem

We begin with a description of the problem and its possible background from a partly mathematical, partly AI point of view. That is because this type of problem is also known within the field of multi-agent systems, and is of relevance for a wide range of IT applications. The system consists of a large number of customer-agents who exhibit a range of behaviours or attributes over time in some domain such as banking, telecommunications, financial services, which is recorded in a database. The company owning the database is interested in extracting as much information about customer behaviour from the database as possible, with a view to marketing new products to the customers in a focused way, identifying customers who may be thinking of switching to a competitor or some other purpose in which it is desirable to identify some subset of customers which fulfil some criterion to a greater than average extent.

The issue seems to be to identify the best set of attributes from the data which are provide useful information about the marketability of a range of products. In this sense, the problem is not well defined, which does not make it less interesting. The concepts highlighted my Numbercraft (see Appendix) are discussed at greater length in the sections below

2.1 Customer agents

The records in the customer database contain information about how the individual customer has behaved at regular intervals up to the present. In some sense, this recorded behaviour contains information about how the customers have reacted to activity in the particular domain, such as the emergence of competing products. The activities of the customers in restricted domains are usually modelled as intelligent agents, who exhibit goal-directed responses to external stimuli. One of the main problems in this area is that the notion of agents is not well supported by the legacy databases which are standard in the world of business. If this information was contained in a truly object-oriented or agent-oriented data base, in which the customer record is modelled by an agent or autonomous object, then the agents themselves could handle the sort of clustering we are trying to achieve here, by, for example, imitating humans in conducting networking behaviour.

The fundamental problem here is to identify some small subset of customer agent types from their historical behaviour with a view to satisfying their goals. This will typically be done a postiori using a wide range of probabilistic methods.

We suppose that there are N such customer agents, which we label $C_1, ..., C_N$. Initially all that is known about these agents is the alpha-numeric attribute data about each agent, stored in a set of state vectors $a_1(t), \ldots, a_n(t)$ at a set of time intervals 0,1,...,T, which live in some finite vector state space A. The elements of these vectors can be continuous variables, such as expenditure, age, GPS-coordinates, or discrete variables such as sex, number of children, product ownership, and so on.

2.2 State and phase space descriptions

Complete information about customers is contained in the curves or trajectories followed by their attribute vectors in the state space A. These trajectories need not be smooth and will generally be very noisy. System properties which can be expressed in terms of population attributes by functions such as $F(a_1(t), \ldots, a_n(t))$ can be computed as averages over the trajectories

$$< F >= \frac{1}{T} \sum_{t=0}^{T} F(a_1(t), \ldots, a_n(t)) \tag{1}$$

An alternative representation of the system is in phase space. Imagine that the state space A has been decomposed into K subspaces $A_1, \ldots, A_K$. Each state vector can thus be assigned to a subspace, using either numeric or fuzzy criteria. Let $n_k(t)$ represent the number of state vectors in the subspace A_k at time t. Dividing these histograms by N gives probability densities

$$P_k(t) = \frac{n_k(t)}{N} \tag{2}$$

The evolution of the system can be followed by defining a transition matrix of conditional probabilities, defined by

$$P_{k,k'}(t+1) = \frac{n_{k,k'}(t+1)}{n_k(t)} \tag{3}$$

where $n_{k,k'}(t+1)$ is the number of customers which move from subspace A_k at time t to subspace $A_{k'}$ at time t+1. That this is a conditional probability measure can be shown from equation (4) by multiplying both sides by $n_k(t)$ and summing over k, using the result

$$n_{k'}(t+1) = \sum_{k=1}^{K} n_{k,k'}(t+1). \tag{4}$$

Within the limits of accuracy imposed by the granularity of the subspaces, approximations system averages can be obtained by averaging quantities over the phase space.

2.3 Choosing the subspaces

The key issue is the choice of the subspaces, which should in some way reflect behaviours typical of the agents which belong to it. Too many subspaces will not be very useful and too few will not contain much information. If the world was ideal, it should be possible to deduce the subspaces corresponding to customer group behaviour by examining the customer data.

One issue to settle is the one of trying to increase the number of subspaces or prune that number from a larger number. We eventually decided on a pruning approach, since coarser grained information can always be obtained from finer grained. For example, suppose that we store the histograms $n_{k,k'}(t+1)$, and for some reason want to merge two subspaces A_k and A_l into a single subspace, which we will denote by A_{kl} before renumbering them. Then the relevant members of the histogram are updated using

$$\begin{aligned} n_{kl,kl}(t+1) &= n_{k,k}(t+1) + n_{k,l}(t+1) + n_{l,k}(t+1) + n_{l,l}(t+1) \\ n_{kl,k'}(t+1) &= n_{k,k'}(t+1) + n_{l,k'}(t+1),\ k' \neq k,l \\ n_{k',kl}(t+1) &= n_{k',k}(t+1) + n_{k',l}(t+1),\ k' \neq k,l \end{aligned} \tag{5}$$

Thus a first exercise is examining criteria for choosing which subspaces to merge. One idea is that since the tridiagonal elements will normally be dominant (an agent will tend to stay in its own subspace or move to a neighbouring one for uninteresting reasons), we should look for similarities in the remaining elements, which could be signalling significant events. These events will normally be hidden by the dominant tridiagonal ones in any norm-based distance.

The time average of the conditional probabilities should converge if the underlying process is stationary and we have found the right classes. It would be interesting to find out to what extent looking at the convergence of the time averages of the transition matrix elements computed for a given partition of synthetic data generated from a transition matrix for a given (unknown) number of subspaces, can give information about the goodness of the partition.

3 Identifying Customer Classes

As mentioned above a bottom-up approach was used to try and group customers into a manageable number of classes, ideally, but not necessarily, in physically meaningful ways. For example, our Toy Model had customers grouped into light, medium and heavy users - a classification that would be preferential to explain any results to a layman, but the most efficient groupings in practice may not correspond to such a physical interpretation (convex verses concave sets).

Essentially, we assume that customers can be modelled by a Markov process whose transition matrix we wish to try and recreate by looking at the data and in each data-set following points from one time step to the next. There is an unresolved issue that such matrices may in practice be time-dependent, but ignore we this for the purposes of this investigation.

Various approaches were suggested, and in some cases tentatively tested, such as applied neural networks, self organising maps and radial basis functions, but in the end a more 'hands on' method was settled on:

- The interval in which each data set was contained was partitioned into 20 arbitrary 'bins'.
- From this partitioning a routine was written to create the transition matrices for successive time-steps between these bins. Each row i of such a matrix corresponds to the probabilities of leaving class i and going to the other classes.
- Using the hypothesis that bins belonging to the same class should share the same transition probabilities, i.e. if two rows of the transition matrices are in fact from the same class, then the probability of leaving these two classes should be similar, we look for bins that can be amalgamated.
- Amalgamation is done by looking at the difference of squares between classes. For example take two states i and j and consider transition probabilities from state i to all states other than state j and visa-versa and if the sum of squares of differences is smaller than some tolerance level then amalgamate states i and j into one class.
- We can either compare only adjacent rows (convex sets) which would at least give more physically describable states, or compare all rows with all rows (non-convex sets). Both approaches were attempted during the Study Group.
- Due to time/computing constraints only the synthetic data was analysed in this way during the Study Group.

Results from synthetic data

Early analysis of the first data set we were confronted with (the 'synthetic data') showed promising success, finding partitioning that fitted in with the method used to derive the data, and almost immediately writing off one of the state variables as purely random, which in fact it was!

More precisely, from the procedure detailed above, all three example data sets appeared to be structured in pretty much the same way. State variables x_1 and x_3 were split into the groupings $0, (0,1], (1,2]$, and x_2 was random numbers. Thus the data could be split into five classes:

- class 1: $x_1 = x_3 = 0$ [omega, or dead class]
- class 2: $x_1 \in (0,1],\ x_3 \in (0,1]$
- class 3: $x_1 \in (0,1],\ x_3 \in (1,2]$
- class 4: $x_1 \in (1,2],\ x_3 \in (0,1]$

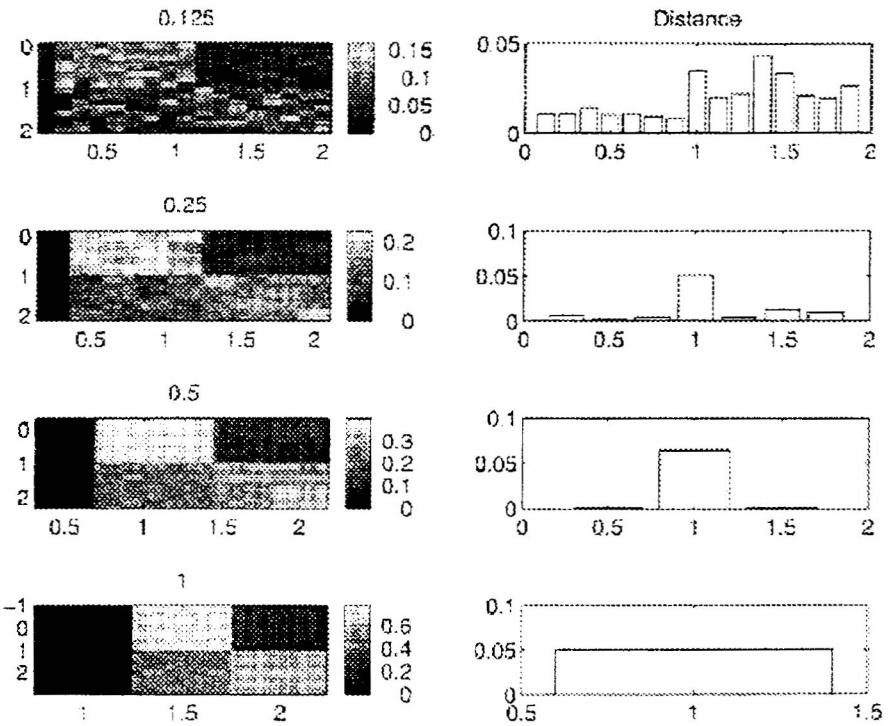

Figure 2: Analysis and amalgamation of one of the state variables from one of the first data sets.

- class 5: $x_1 \in (1,2]$, $x_3 \in (1,2]$

MATLAB then gave the probability transition matrices between these classes as follows.

For example data set 1,

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.014 & 0.561 & 0 & 0.237 & 0.189 \\ 0.002 & 0.477 & 0.521 & 0 & 0 \\ 0 & 0 & 0.398 & 0.312 & 0.290 \\ 0 & 0 & 0.422 & 0.281 & 0.297 \end{bmatrix}$$

For example data set 2,

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.02 & 0.58 & 0 & 0.39 & 0 \\ 0.02 & 0.44 & 0.54 & 0 & 0 \\ 0 & 0 & 0.37 & 0.63 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

For example data set 3,

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.005 & 0.528 & 0 & 0.261 & 0.206 \\ 0.007 & 0.489 & 0.504 & 0 & 0 \\ 0 & 0 & 0.408 & 0.297 & 0.294 \\ 0 & 0 & 0.411 & 0.285 & 0.304 \end{bmatrix}$$

In Example 2, the probability transition matrix had no way of moving into class five. This was borne out by the data, which had no entries in this class. The classes with $x_1 > 1$ had no way of moving to the dead class, reflecting the way that the examples were set up by Numbercraft. From these a simple diagram for the transition probabilities could be constructed (Figure 4).

4 Statistical analysis of the dead

Introduction

It was evident from the type of data to be analysed that some customers, the dead, terminated their transactions during the data period. Since an objective of any analysis or clustering is to highlight such customers prior to their demise, and target them for extra publicity, it was deemed useful to investigate the life time of customers who died during the period. The object of such an investigation is to try and see a behaviour pattern common to such customers, and use this to detect the subset of live customers who may be "considering dying". The investigation was undertaken in two parts, the first of which looked at the first synthetic data set, and the 8 dead customers contained therein. The second part of the investigation was concerned with the real data set, and the first 8 dead customers contained within it. In the case of the real data there was a fundamental difference in behaviour, as some customers apparently died and then returned to life again. This necessitated differentiating between the temporarily dead, whose overall behaviour was similar to those who never die, and the permanently dead.

Synthetic Data

The first investigations on the synthetic data were concerned with the first synthetic data set, and several indicators of "death-like" behaviour were considered. The first synthetic data set contained data for 3 states for a time period of 2

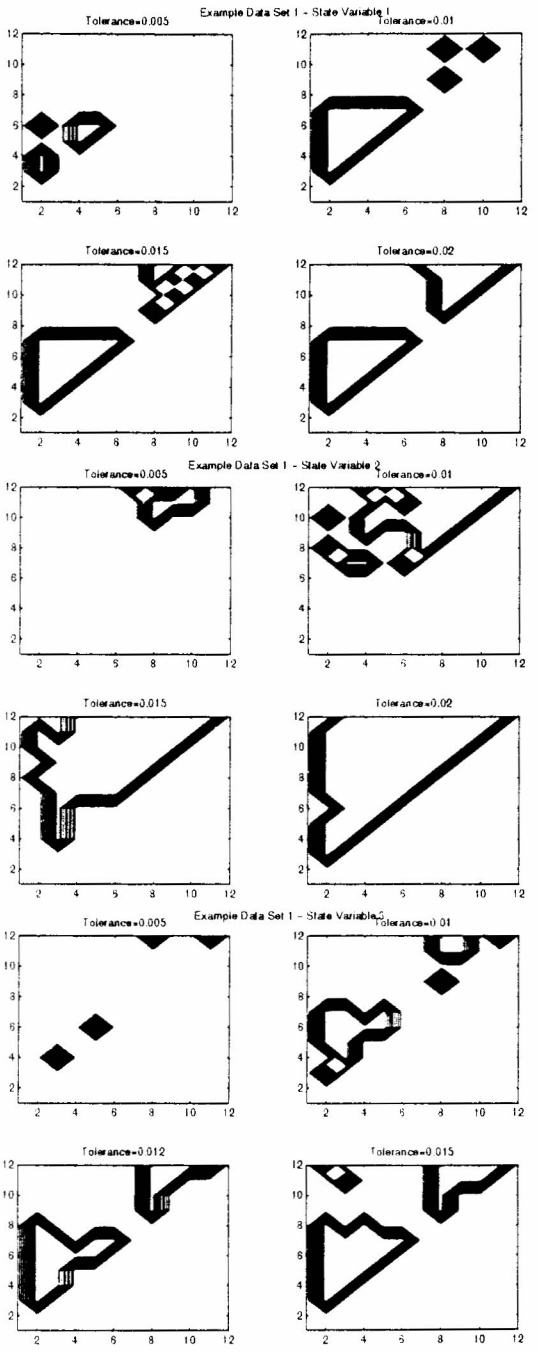

Figure 3: Example data set one, state variable one: amalgamating rows with various tolerance levels

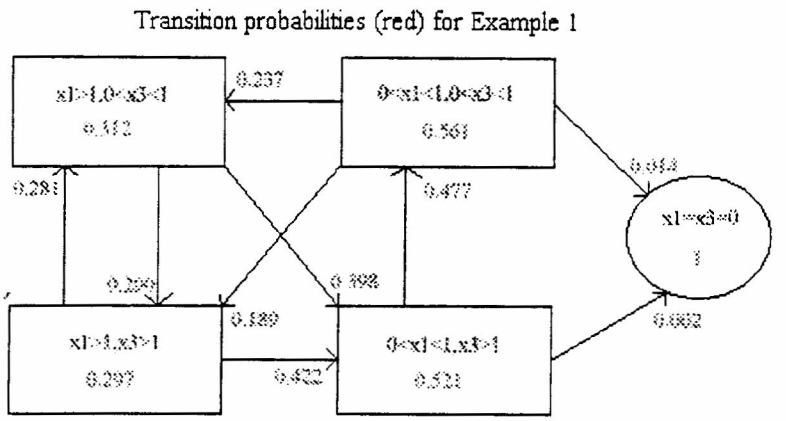

Figure 4: Transition probabilities between final class classification for example data set 1.

years, with a time step of 2 weeks. The first investigation of the dead customers was concerned with the maximum and minimum values of each state for each dead person, over the period of their life. It was postulated that if the maximum value of a state occurred near the beginning of the lifetime, and the minimum occurred towards the end of the lifetime, this could indicate a dying customer. A Matlab program was written to extract the minimum and maximum value of each state for each customer, and the time step at which it occurred. This enabled the range of each state to be calculated for each customer, and also a normalised value for the position of the maximum and minimum, which was calculated as follows:

$$p_{\max} = \frac{t_{\max}}{t_{tot}} \qquad p_{\min} = \frac{t_{\min}}{t_{tot}}$$

where: $t_{\max}$ is the time step when the maximum occurs, $t_{\min}$ is the time step when the minimum occurs and t_{tot} is the total lifetime of the customer.

The results of this investigation on the dead customers show that the maxima, minima and ranges for the 3 states are very similar for each customer, with states 1 and 3 showing slightly more variation than state 2. The data contained in the mean max and mean min columns is the $p_{\max}$ and $p_{\min}$ data, and this has a little more variation. If the value is close to 1, then this indicates that the maximum or minimum occurred towards the end of the customer's lifetime, and if the value is close to 0 this indicates that the maximum or minimum occurred towards the beginning of the customer's lifetime. Support of the initial postulation would therefore be indicated by having the $p_{\max}$ data close to 0 and the $p_{\min}$ data close to 1. The results only support the initial postulation for

some customers, and a range of behaviour is illustrated from having the maximum towards the end and the minimum towards the beginning of the lifetime through to having the maximum and the minimum occurring close together, as for customer 5 state 3 for example.

For comparison the same analysis was run on the first 4 live customers, having the full 2 year range of data. For live customers, one might suppose that the maximum would not occur near the beginning and the minimum not near the end of the data period, so as to indicate not dying. Thus if the $p_{\max}$ data is close to 1 and the $p_{\min}$ data close to 0, this might be supposed to be the live postulation. Once again, this postulation is not born out by the results, as although some customers exhibit this pattern for some states it is not true for all customers and all states.

The results from the first investigation were quite disappointing, since they didn't result in a method for differentiating between live and dying customers. However, since no information on the states was available, it might be incorrect to suppose the minimum and maximum value gave a good indication of behaviour. For the live customers, it may have been more realistic to look for a minimum and maximum per year, for example. Also, since the data is synthetic, it may include behaviour or anomalies which would not be present in real data.

The second investigation of dead customers was concerned with statistical calculations. The mean, standard deviation, average of the absolute deviation from the mean, and correlation coefficient were calculated for all the data, for the 8 dead customers and also for the first 4 live customers. It was postulated that the range of the average of the absolute deviation from the mean would be broadly constant for live customers, and would show more fluctuation for dying customers. The correlation coefficient was expected to be close to zero for live customers, so that behaviour between states was different, and to move closer to ± 1 for dying customers, indicating a trend towards similar behaviour across all states.

The results of these calculations on the entire data set are as follows:

	Mean	Std	Av. Dev	Correl coeff	Av. Dev. %
State 1	0.612627731	0.614327463	0.523099242	0.437321791 (1 - 2)	85.39
State 2	0.714272022	0.669188592	0.5922021	0.420554519 (1 - 3)	82.91
State 3	0.699627759	0.662305893	0.584578175	0.452933904 (2 - 3)	83.56

The first interesting point to note is that the values for the mean and standard deviation are very similar to each other, and in the case of State 1, the standard deviation is actually greater than the mean. Thus, it is possible to be only one standard deviation from the mean and be in a region containing very small, or negative, values. The average of the absolute deviation from the mean is however smaller than the mean in each case, and interestingly, is always around 84%. This may well be due to the method used to generate the synthetic data. The correlation coefficient takes a value of 0 if there is no correlation and a value of ± 1 if there is perfect correlation. The correlation between each combination of pairs of states gives a value of around 0.4, and is therefore indeterminate.

Looking first at the values for mean and standard deviation, it can be seen that there are no cases for which the standard deviation is greater than the mean. It might be deduced that including the zero data for each person who dies is skewing these measures when the whole data set is considered. For all cases, the average of the absolute deviation from the mean is fairly similar, and lies between 0.36 and 0.6. Looking at the average deviation as a percentage of the mean value it can be seen that customers 2, 8 and 10 have a broader range than the other customers, however this may be because these customers have a shorter life span. The correlation coefficients all lie between about -0.4 and 0.6, which is rather disappointing since this is the region from no correlation through to indeterminate. It was hoped that the behaviour patterns of the states for dead customers might tend to one pattern, thus giving a correlation coefficient close to ± 1 and a measure of likelihood of death.

For comparison the same analysis was run on the first 4 live customers, having the full 2 year range of data. For these live customers, the mean values are similar for each state, and the standard deviation is also similar, and smaller than the mean in each case. This supports the supposition that including the zero data for each person who dies is skewing these measures when the whole data set is considered. The average of the absolute deviation from the mean is also similar for each person and for each state, and when expressed as a percentage of the mean it lies between 47% and 57%, which is a much smaller range than for the dead customers. The correlation coefficients all have absolute value less than about 0.2, indicating no correlation, which is as anticipated.

Overall, the results from the second investigation are more promising. The live customers agree with the anticipated behaviour by having a small range for the average of the absolute deviation from the mean (expressed as a percentage) and by having correlation coefficients close to zero. The results from the dead customers don't show the pattern expected, but it is possible that better results may be obtained from the real data set.

Real Data

The real data set contained data for 220 customers over a period of 14 time steps, for 6 states. For ease of comparison, the first 4 live customers and the first 8 dead customers were investigated, and their behaviour was analysed as for the synthetic data. For interest, the maxima and minima analysis was carried out on all 6 states, but the statistical analysis only considered the first 3 states. Of the 8 dead customers, only 3 remain dead (customers 7, 11 and 14) and the remainder only die temporarily, though perhaps more than once. This is a significant difference from the synthetic data, and will have an effect on the results of the analysis. It would be helpful if an indicator for each of the 3 types of customer behaviour could be developed, so that living, temporarily dead and permanently dead customers could be differentiated from each other.

The first investigation was identical to that for the synthetic data, with the maximum and minimum value of each of the six states, $p_{\max}$, $p_{\min}$ and the range being calculated for the first 8 dead customers. Due to the difficulties

of calculating the actual lifespan for the temporarily dead customers, and since the time period for the data is short, t_{tot} was taken to be 14 for all customers. The maxima, and ranges for the 6 states vary widely both between states and between customers, although state 1 contains the largest range value for each customer. The minima are zero for every state and every customer, which was not true or the synthetic data, and hence the range and the maximum in each case is identical. The data contained in the mean max and mean min columns is again the $p_{\max}$ and $p_{\min}$ data, and this appears to show a pattern. Customers 7, 11 and 14 who die and remain dead have $p_{\max}$ and $p_{\min}$ values less than 0.5, indicating that their minima and maxima occur towards the beginning of the time period. The remaining customers, who die temporarily, have one or more states with a $p_{\max}$ value close to 1, indicating a maxima towards the end of the time period, and one or more states with a $p_{\min}$ value close to 0, indicating a minima towards the beginning of the time period, as postulated initially. Customer 8 has four states for which the $p_{\max}$ values are close to 1, perhaps indicating a very lively dead person!

For comparison, the same analysis was run on the first 4 live customers: the maxima, minima and range values vary widely between states and customers, though once again state 1 contains the largest maximum, minimum and range value for each customer. Customers 2, 4 and 6 have at least one state with a small $p_{\min}$ value, and one with a large $p_{\max}$ value, which agrees with the initial postulation However, customer 3 has only small $p_{\max}$ values, and both small and large $p_{\min}$ values, which indicates a possible death-like trend in behaviour. Clearly, to confirm these initial results the analysis must be run on all 220 customers.

Again, the second investigation was concerned with statistical calculations, with the mean, standard deviation, average of the absolute deviation from the mean and correlation coefficients being calculated for all the data, the first 8 dead customers and the first 4 live customers. It was anticipated, as before, that the range of the average of the absolute deviation from the mean would be broadly constant for live customers and show more fluctuation for dying customers, and that the correlation coefficient between each combination of states would be close to zero for live customers and close to ± 1 for dying customers. The results of these calculations on the entire data set of 220 customers are as follows:

	Mean	Std	Av. Dev	Correl coeff	Av. Dev. %
State 1	97.98367857	120.8339754	89.45534752	0.685057717 (1 - 2)	91.30
State 2	3.385714286	3.874776295	2.780751391	0.224356566 (1 - 3)	82.13
State 3	14.31309091	22.53093884	15.1082242	- 0.107393043 (2 - 3)	105.56

Firstly, as for the synthetic data, the values for the mean and standard deviation are similar for all three states, and the standard deviation is greater than the mean. This is again attributed to the inclusion of zero data for each person who dies. The average of the absolute deviation from the mean is again smaller than the mean, but the values as a percentage of the mean are no longer similar. This supports the supposition that the similarities for the synthetic

data were due to the method of data generation. The correlation coefficient values are small for states 1 to 3 and 2 to 3, suggesting little or no correlation, and larger for states 1 to 2, which indicates possible correlation.

As has already been noted, some customers died on several occasions during the time period (1,5,8,16,17), whilst others died and then remained dead (7,11,14). Looking at the values for the mean and standard deviation it can be seen that they are still similar, but for customers 8 (states 1 and 3) , 11, 14 (states 1 and 3) and 16 the standard deviation is still greater than the mean, as is the value for average of the absolute deviation from the mean, leading to large percentage values, often ¿ 100%. This set of customers only contains 2 out of the three who die, but contains all of the customers with 3 or more deaths during the data period. It is clear therefore that the zero values are skewing the statistical results, and that there is no clear indication of the true dead customers. The correlation coefficients between the various combinations of the first 3 states appear to indicate the two types of dying customer behaviour. Customers 7, 11 and 14, who actually die, have the absolute value of all correlation coefficients greater than 0.67. Customer 11, who is hardly alive at all, has all three correlation coefficients larger than 0.93, which indicates a high degree of correlation. Conversely, customers 1, 5, 8 and 16, who only die temporarily have a correlation coefficient greater than 0.5 between states 1 and 2 but values of less than 0.5 for the other two combinations. Person 17, who has a death at the final time step, has correlation coefficients in the same pattern as customers 1, 5, 8 and 16. This seems to suggest that the correlation coefficient may be used to differentiate between the dead for whom all correlation coefficients have absolute value greater than 0.6, and the temporarily dead who only have a value greater than 0.65 between states 1 and 2. It is possible, however, that the results indicated by the correlation coefficient values are being affected unduly by the number of death occurrences.

With the exception of customer 3 state 3, all the mean values are larger than the standard deviation values by a factor of approximately 2. The average of the absolute deviation from the mean has its largest magnitude for state 1 in each case, but when expressed as a percentage the largest value is that for stage 3. These results differ from those for the temporarily and permanently dead, as they had the standard deviation greater than the mean, and the percentage values of the average of the absolute deviation from the mean all similar. This is no doubt due to the appearance of zero data for the dead customers, which is not present in the live customers. The correlation figure for states 1 to 2 is greater than 0.67 for all the live customers. All the remaining values are less than 0.5, with the exception of the values for states 2 to 3 for customers 4 and 6 which are close to -0.7. These results appear to support use of the correlation coefficient for all possible combinations of states to detect dying customers.

Comment

Due to the small range of data available for analysis, and the relatively short time period of the initial investigations it is perhaps a little dangerous to draw

any definite conclusions, particularly since only a small part of the real data set was investigated. However, it would certainly appear that there are some results worth further investigation, in particular the use of the correlation coefficient for distinguishing between the two types of dead people.

With regard to the statistical measures, it would appear that if the standard deviation is larger than the mean for all states, and the averages of the absolute deviation from the mean expressed as a percentage of the mean are similar, then the customer has been dying either permanently or temporarily. Conversely, if the standard deviation is smaller than the mean for all states and the averages of the absolute deviation from the mean are varied then the customer has always been alive. Correlation coefficients close to ± 1 for all possible combinations of states appear to indicate a customer who is dying, and thus it may be possible to differentiate such a customer from those who only die occasionally. As stated earlier, it is possible that all the results are unduly affected by the zeros from the death occurrences.

5 A method to quantify 'Customer Behaviour'

The final approach that was undertook was an analysis of the data as time-series using a 'dynamical systems' mentality. In particular the methods used were those used in the 'FRACMAT' project to analyse the suitability of wire to be coiled into springs [1].

The basic idea is that we are looking for customers that are changing their behaviour, and so we must be able to differentiate customers who have similar records week in week out to those whose weekly records vary quite wildly. So for example a customer who usually has a steady time-series and then suddenly decreases usage of some service and a customer who every few weeks dramatically decreases usage but then increases will appear to be behaving identically in the short term when both decrease usage.

After a suitable projection onto the most informative subspace (for example using Singular Value Decomposition [2]) the time-series from real data of ten customers can be represented as in Figure 5.

From looking at this representative sample is should be clear that the bottom two time-series are very similar, as are the two graphs on the second row. The question is how can we try to quantify just how similar such time-series are?

Let us re-cap the method used for the analysis of the suitability of wire to make springs.

- A length of wire is coiled into one long spring, with perhaps $50 - 100$ distinct coils. The successive distances between each coil (the 'pitch') is then measured to give a time series $(p_1, p_2, \ldots, p_N)$.
- This data is then normalised by dividing each measurement through by the average $\mu = \sum p_i/N$. Call the new time-series $(q_1, \ldots, q_N)$.

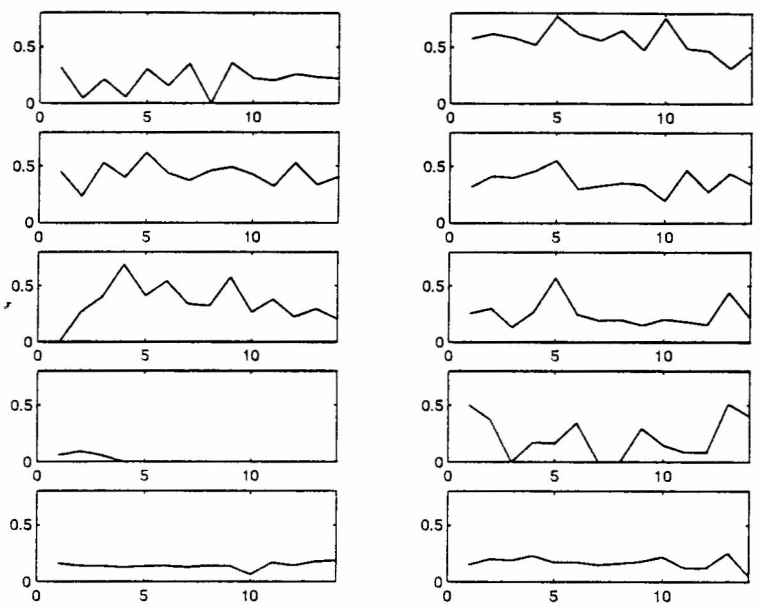

Figure 5: Projection of the behaviour of 10 real customers.

- The matrix

$$\begin{bmatrix} a_0 & a_1 \\ a_1 & a_0 \end{bmatrix}$$

  is then calculated where $a_0 = 1/N \sum q_i^2$ and $a_1 = 1/(N-1) \sum q_i q_{i+1}$.

- The eigenvalues $a_0 + a_1$ and $a_0 - a_1$ of this matrix are then calculated with the larger of the two being assigned the variable σ_1 and the smaller σ_2.

- On the same graph, for each such data set plot σ_2/σ_1 and $\sqrt{\sigma_1^2 + \sigma_2^2}$.

- On this plot points to the upper right corner correspond to wire that is poor (inconsistent with high scatter), to the lower left good (low scatter and consistent), top left high scatter but consistently so, lower right moderate scatter but inconsistent (see Figure 6).

What is really being measured is a measurement of how elliptical the 'Packard-Takens' plot of each time-series is. This is the plot of p_i against p_{i+1} which will produce a scatter of points. If each p_i is similar to p_{i+1} then each point will lie close to the line $p_i = p_{i+1}$, at the other extreme they will lie a long way from it and so it is instructive to quantify how 'elliptical' this cloud of points is. The quantifiers detailed above are essentially standard deviation of the time series and the ratio of the lengths of the principal axis of the ellipse found by principal component analysis of the cloud.

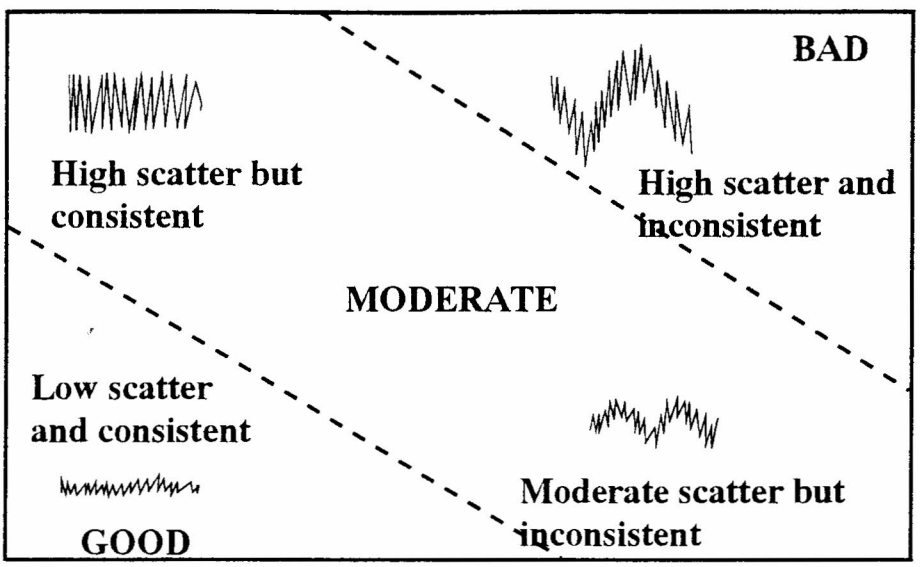

Figure 6: Classification of springs due to method outlined in body of text. X-axis is σ_1/σ_2 and Y-axis is $\sqrt{\sigma_1^2+\sigma_2^2}$, and typical time-series of pitch measurements are shown.

In the industrial setting this method has proved extremely accurate in discriminating between different wires, but how well does it perform on the data we are considering here? An initial problem is that we have been presented with data sets that only consist of 14 time steps, but some preliminary analysis has shown that this does not seem to matter. Preliminary investigation also showed that it was more informative not to 'divide through by the average' and so keep more of the quantitative information.

When this was taken into account and the above algorithm was run on the same 10 customers shown in Figure 5, the plot in Figure 7 was obtained. The numbers indicate the customers in Figure 5 where the graphs are numbered from left to right, top to bottom. It should be clear that there is some obvious clustering involving customers 1,6,8, customers 9 and 10, customers 3,4,5 with customers 2 and 7 both being distinct. A comparison with the original graphs do indeed suggest that the algorithm has been quite successful in differentiating between behaviour of time-series, even with the small number of data points, and such a classification could be used to describe our 'classes' of customer.

The issue of whether such a method could be used to detect a sudden change in behaviour was not addressed, or rather how many data points would be needed in order to detect a change in the long-term behaviour. The method as a whole though does show considerable promise.

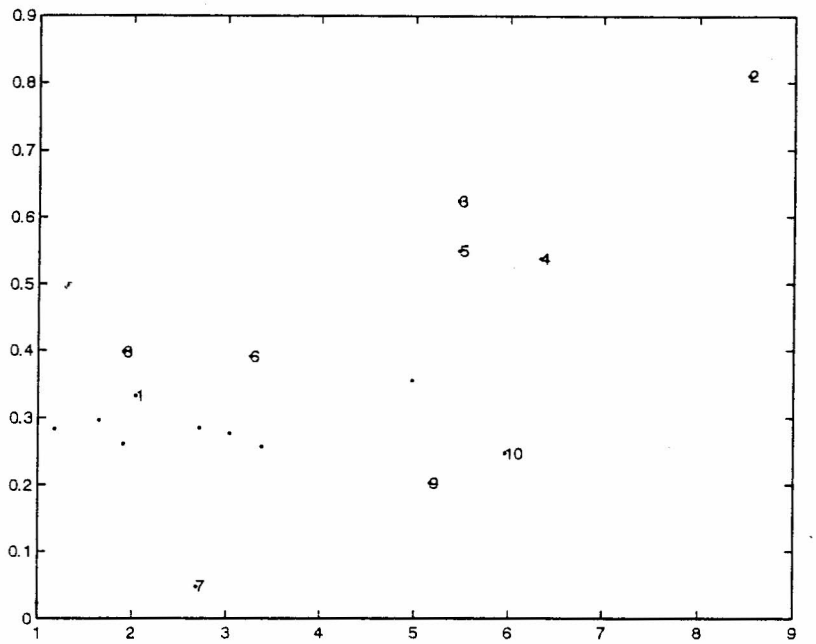

Figure 7: FRACMAT classification of 10 real customers behaviour.

6 Conclusions

Despite time limitations and some serious computing issues, we have followed several lines of attack on the problem which show a great deal of promise on both the synthetic and real data sets. It is believed that further advances can be made, but within a 'study group' setting this became infeasible due to the sheer amount of data with which we were presented.

In particular analysis could be extended to include more than one state-variable, or higher-dimensional projective spaces or even the inclusion of 'Fuzzy set' theory in our description of classes.

Appendix I - Problem Description

"Consider the situation where a company (e.g. a digital TV, telephone, or Internet service provider) monitors the behaviour of its (probably many thousands of) customer accounts. It does this by partitioning them into a small number of classes (of the order of ten), in such a way that customer accounts which are deemed to belong to the same class exhibit similar behaviour. This has applications in marketing strategy: for example, one might discover that a particular

pattern of customer account behaviour is a precursor to a customer either terminating his/her account or trading his/her customer up. If one had an automated procedure for identifying all customer accounts displaying this pattern of behaviour, then one could offer incentives so as to try and retain their business.

"We will consider a situation where information on each account is given at discrete intervals of time (possibly every week), and this information will take the form of (up to ten) numeric performance indicators for each customer account. Some of the indicators may be irrelevant, others may be highly correlated so that not all the indicators need be considered in the allocation of accounts to classes.

"The indicators will be bounded above and below in each case. Hence, wlog, we assume that each indicator is given on an interval. The cross product of these intervals is the state space in which all 'live' accounts must exist. There will also be one additional special class, Ω, the class of closed or 'dead', accounts where all indicators are blank.

"While the accounts have indicators which can be plotted in state space and the accounts can be broken into classes, it should not be assumed that there will be any detectable clustering of the points when plotted in state space. That is to say, methods which seek to find natural clusters in the points of state space which represent the accounts are probably not suitable.

"At the most basic level, the choice of the number of classes is also important: one wants enough classes to distinguish usefully between different kinds of customer account, but not so many that designing action for each class becomes a burden, or so many that stochastic noise dominates. Moreover, a large number of classes may result in sampling uncertainty.

"Given a set of classes - a complete partition of the possible state space - one will see that some customer accounts change class from one time period to the next. From these changes one naturally derives a transition matrix whose entries P_{ij} give the probability of a customer residing in class j given that he/she was in class i the previous week. We want to find a method for designing 'clean' classifications where the transition matrix is fairly sparse, in the sense that although most elements will be non-zero, only a small number (perhaps the diagonal elements and a few others) will be significant.

"In cases where the classification leads to a sparse transition matrix as described above there is a possibility of using clustering techniques in phase space even though the may be of little or no use in state space. For each indicator we can plot its value at time t against its value at time $t+1$ for each account. If the indicator is related to transitions between classes then there should be a heavy distribution of points around the line representing $indicator(t) = indicator(t+1)$ and a light distribution far from this line. In this case some clustering may be apparent and the boundaries between clusters will be related to critical values in changing between classes. This could be a useful tool but has the disadvantage of doubling the dimension of the problem.

"We may require that the estimates of transition probabilities are accurate - the distributions of 'residence times' should be exponentially decreasing.

"Once such a 'clean' classification has been designed and the corresponding transition probabilities calculated,it will be possible to draw a directed graph

showing the main flows of customers from class to class in the system. Using the directed graph one should be able to calculate the expected profit from (and hence the value of) a customer whose account lies in a particular class over the lifetime of his/her contract. This has applications, for example, in the valuation of Internet companies, many of which presently make a loss, yet have large stock market quotations based on the size of their customer databases.

"The number of accounts in a data set may affect the choice of algorithm (special methods may either become available, or become necessary, if the dataset is very large).

"A further issue is the resolution in time at which account data is gathered. For example, in the case of a telephone company, an individual customer might make a very different number of calls from week to week, even though his/her long term behaviour is effectively constant. In this situation we don't want a method which spuriously indicates the customer changing classes from week to week. The solution might be to average performance indicators over two or more weeks so as to try and eliminate stochastic noise for an individual customer. The disadvantage to this approach is that the reaction time, e.g. to signs that a customer is about to terminate his/her contract, is increased.

"Numbercraft will supply some data for testing and evaluation purposes, the project requires:

- *the generation and manipulation of alternative partitions of state space,*
- *definition of suitable performance measures for the transition matrices (trading the desired sparseness, accuracy of transition probabilities etc. with complexity),*
- *application/testing under both simple and noisy circumstances.*

References

[1] L. Reynolds, D. Wood, M. Muldoon, M. Bayliss, I. Stewart, *Chaos Theory, Fractal Materials and Spring Wire*, Wire Industry Feb 1997, pp 144-148.

[2] Press, Flannery, Teukolsky, Veterling, *Numerical Recipes in .....*, Cambridge University Press, 1989.