PAM
Analytics
Predictive Asset Management

PAM Analytics Projects: Database Marketing

Introduction

PAM Analytics has carried out many database marketing projects, including missing value imputation, segmentation using cluster analysis and decision trees, customer survival time, propensity (churn) modelling and profiling. This document describes missing value imputation and segmentation in general, and then describes projects that PAM Analytics has carried out in these areas.

Missing Value Imputation

Missing data (sometimes called gaps or holes) is a very serious but unfortunately often neglected problem in most commercial databases. It is a particular problem for marketing companies, lifestyle and demographic data owners and sellers, and companies working in customer relationship management and credit (risk) analysis. Missing data make the *effective* size of databases smaller than their *actual* size because they are not as rich or useful as they appear to be as judged by their actual sizes, and so they reduce the quality and commercial value to the business of the databases.

PAM Analytics has developed a new and unique software product for imputing missing data in large databases. It uses an advanced and specially enhanced form of the method of *k* nearest neighbours to impute the missing data and so convert partially populated databases, i.e. databases with incomplete records, to fully populated databases, i.e. databases where all the records are complete because all the missing data have been replaced by imputed values. The software is of benefit to organisations that own, maintain or analyse databases with missing data, and which require the missing data to be replaced by well-based imputed values.

PAM Analytics has used the software for a high street jeweller, a mobile phone operator and a database marketing company. It can be used to impute data that help define and classify individuals and businesses. Personal data that can be imputed include age, income, occupation, newspaper readership, propensity to buy particular products, and business data that can be imputed include SIC sector (a classification system for businesses) and business turnover.

k Nearest Neighbours

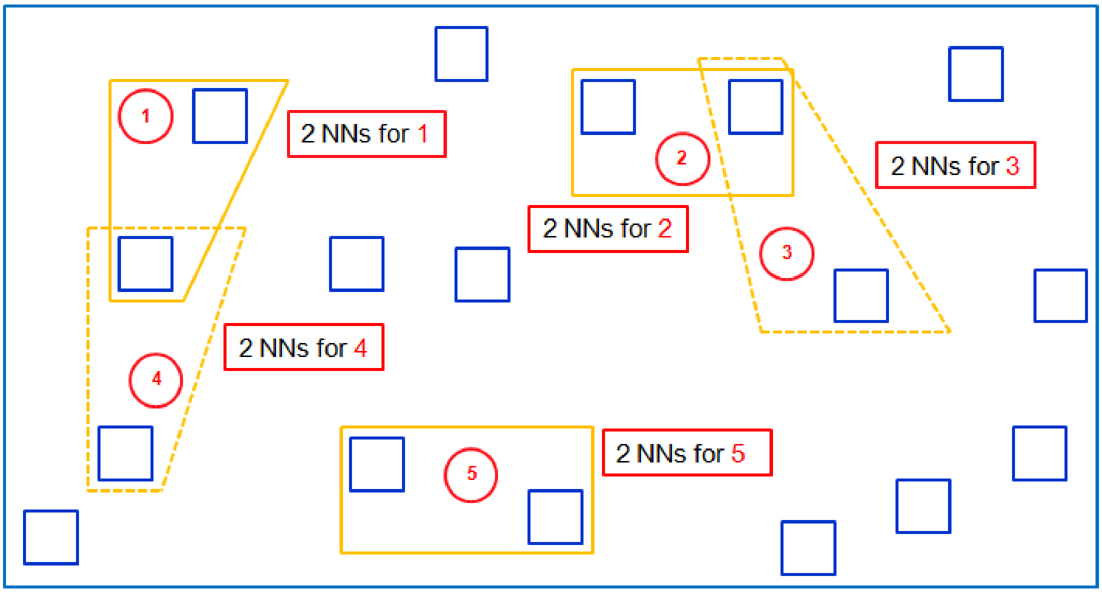
k nearest neighbours is a non-parametric method used in estimation and classification problems. It is non-parametric because it does not require the data to follow a particular distribution. Rather, it is based on the concept of local similarity, i.e. the statistical similarity between two records is based on the distance between them when their common known data are plotted. The more similar two records are, the shorter the distance between them. If two records are identical in each field, the distance between them is zero.

The method involves calculating the distance from each incomplete record to each complete record. The *k* nearest neighbour complete records, where *k* is an integer, to each incomplete record are the *k* complete records with the *k* shortest distances to the incomplete record. The missing data in each incomplete record are then imputed from its *k* nearest neighbours. Thus, imputing missing values using *k* nearest neighbours can be regarded as being based on a form of very local cluster analysis segmentation (see below for more information on cluster analysis).

A very appealing feature of using *k* nearest neighbours for imputation is that it is applied across all fields in the database in one go rather than on a field by field basis, and so it is very suitable for imputing missing data in very large databases.

Figure 1 shows how *k* nearest neighbours works geometrically for 2 (= *k*) Nearest Neighbours (NNs).

Figure 1



The **red circles** are the incomplete records and the **blue squares** are the complete records. The **orange lines** enclose the two nearest neighbour complete records to each incomplete record. The missing data in each incomplete record are imputed from its two nearest neighbours. Incomplete records 1 and 4, and incomplete records 2 and 3 each share a complete record. Nine complete records are not associated with any incomplete records.

Example

Table 1 shows a sample of 10 records from a large database. Records1, 2, 4, 6, 7 and 9 are complete and the other three records have missing data (shown by **?**).

Table 1

Record No.	Marital Status	Resid. Status	Age (yrs)	Bank (mnths)	Bank Card	Address (mnths)	Employ.	Occup.
1	M	T	28	18	Y	20	12	ES
2	M	T	26	24	Y	24	66	ES
3	D	?	26	?	N	36	?	SB
4	M	T	23	60	N	36	60	EM
5	M	?	?	132	Y	?	0	?
6	D	T	43	72	N	6	11	S
7	S	P	26	14	N	54	42	EB
8	?	?	28	12	?	?	66	SB
9	M	Z	37	126	Y	82	120	EP
10	?	?	62	?	Y	12	?	?

Table 2 shows the nearest neighbour records for each incomplete record. The nearest neighbour record to each incomplete record is its most similar complete record. This means that when the known values of an incomplete record and all the complete records are plotted, the compete record that is closest to it is its nearest neighbour.

Table 2

Record No.	Nearest Neighbour	2nd Nearest Neighbour	3rd Nearest Neighbour
3	4	7	6
5	1	2	9
8	2	7	4
10	1	2	9

Table 2 shows that for record 3 the most similar complete record is record 4, and the next two most similar complete records are records 7 and 6 respectively.

Now that the nearest neighbours in ranked order to each incomplete record have been found, the missing data can be imputed quickly and easy. This requires two questions to be answered: firstly, how many nearest neighbour records should be used; and secondly which statistic should be used to impute the missing data.

When deciding how many nearest neighbour records to use, it is very important to consider the fact that as more nearest neighbour records are used, they become increasingly dissimilar from the incomplete record. Looking at Figure 1, this means that the distances from the incomplete record to the complete records increase as more nearest neighbour records are used. Now, in general, as sample sizes increase, the confidence in the estimate of the statistic calculated from the sample increases. However, this is not true when *k* nearest neighbours is used for imputation because the nearest neighbour records to each incomplete record are ranked by *increasing* distance and therefore *decreasing* similarity to the incomplete record. The trade-off between accuracy and the number of nearest neighbour records can be investigated by running a number of simulations to find the optimal number of nearest neighbour records.

The answers to the second question are much more straightforward. Missing continuous data can be imputed from either the mean or the median of the nearest neighbour records and missing nominal data are imputed from the mode of the nearest neighbour records.

Table 3 shows the results of imputing the missing data in Table 2 using three nearest neighbours and the mean for the continuous fields and the mode for the nominal fields (the imputed values are in **red**).

Table 3

Record no.	Marital Status	Resid. Status	Age (yrs)	Bank (mnths)	Bank Card	Address (mnths)	Employ.	Occup.
1	M	T	28	18	Y	20	12	ES
2	M	T	26	24	Y	24	66	ES
3	D	T	26	49	N	36	38	SB
4	M	T	23	60	N	36	60	EM
5	M	T	30	132	Y	42	0	ES
6	D	T	43	72	N	6	11	S
7	S	P	26	14	N	54	42	EB
8	M	T	28	12	N	38	66	SB
9	M	Z	37	126	Y	82	120	EP
10	M	T	62	56	Y	12	66	ES

Example Missing Value Imputation

PAM Analytics worked with a customer relationship management consultancy to improve the quality of a large customer database of a well known high street jeweller. The database consisted of transactional data, and appended lifestyle and demographic data. It only had 19,000 complete records but about 8 million incomplete records, all at individual customer level. The missing lifestyle and demographic data meant that the database was of little commercial value to the jeweller for understanding his current customers and then designing effective marketing programmes to target potential new customers.

PAM Analytics imputed the missing data in the large pooled database. Results from marketing campaigns using the new fully populated database (containing both recorded data and imputed data) showed significant improvements on the results of previous marketing campaigns. Thus, the quality of the database was improved very significantly by imputing the missing data.

Segmentation

Many companies have customer databases to help them understand their customers' behaviour and what drives their needs and desires. The databases are very large, with possibly hundreds of thousands of records (customers) and hundreds of fields. They are formed by merging, with varying degrees of success, data from a large number of disparate databases to form a single customer view database with a wide range of information about the customers. Since the number of records in the merged databases is very large and the records show great variation, analysing the databases at record level to understand each customer's needs and wants is not feasible. These problems are overcome by segmenting the records in the databases.

Segmentation is used to classify or categorise customers in a single customer view database into a finite number of groups where the customers in each group are more similar to one another than they are to customers in the other groups. By segmenting their customers and profiling the groups, organisations can understand their customers more clearly and with better differentiation, and so are able to design appropriate and individual (for each cluster) marketing strategies. Indeed, one of the commonest reasons for segmenting customer databases is to identify the most profitable customers.

There are two types of segmentation: cluster analysis and decision trees.

Cluster analysis. Also called undirected segmentation, cluster analysis does not have a target variable. The output is a number of homogeneous groups called clusters where each cluster contains similar people or objects.

Decision trees. Also called directed segmentation, decision trees have a target variable. The output is a number of nodes where each node contains a clearly defined set of people or objects.

Cluster Analysis

Cluster analysis is based on calculating the similarities between the records in the database and then grouping them by their similarities. The more similar two records are, the more likely they are to be in the same cluster. The similarity between two records is calculated from their values and is defined as the (geometric) distance between them when their values are plotted in multi-dimensional space. Distance and similarity are inversely related because the shorter the distance between two records, the more similar they are. In the extreme case, the distance between two identical records is zero.

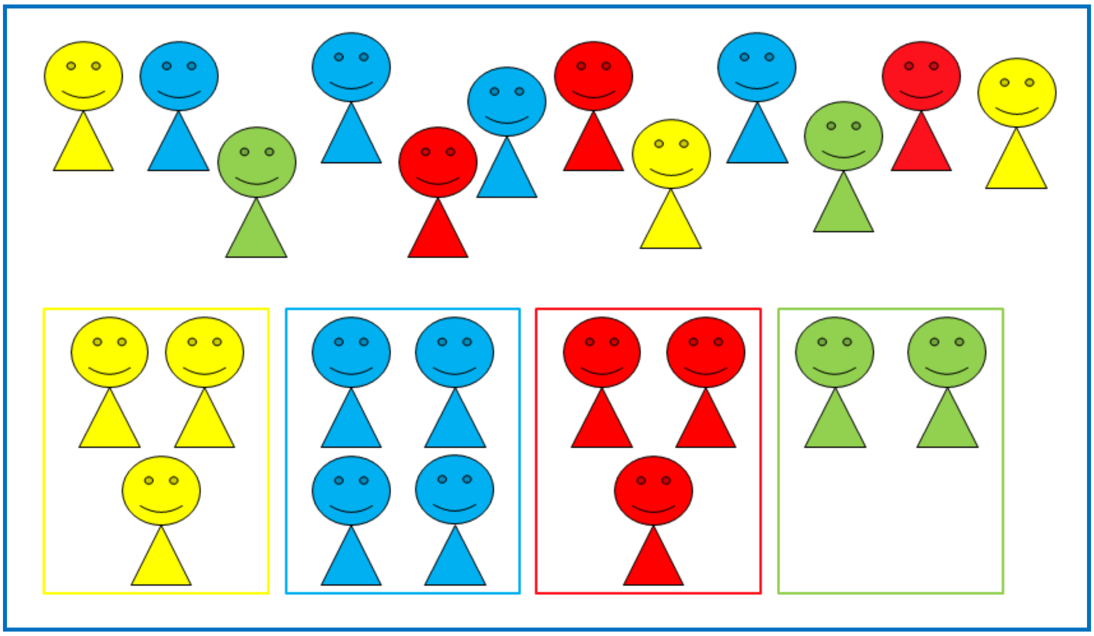
There are a number of cluster analysis methods but they are all based on the concept of similarity. The methods differ in a number of ways, including how they define distance and how the differences between the clusters are defined.

After the records have been segmented, the number of clusters is worked out empirically from the results of each stage of the cluster analysis. It is impossible to give absolute rules for the number of clusters but it is unusual to have more than about 10 clusters. There is a trade-off to be considered when working out the number of clusters. As the number of clusters increases, their internal homogeneity increases. High internal cluster homogeneity is desirable but it must be balanced by the work required to process and maintain the larger number of clusters. However, if a population can be segmented into a reasonable number of clusters and the fields used in the segmentation differentiate the fields well, the number of clusters is usually clear.

After the segmentation has been carried out, a profile of each cluster is built by analysing in each cluster the fields that were used to build the clusters. For example, one cluster may describe young struggling families (at one extreme) and another cluster may describe well-off pensioners (at the other extreme). Since people differ to varying extents as described by the clusters, the profile of each cluster dictates how the customers in each cluster should be considered and treated, and which products are most suitable for them.

Figure 2 shows the principle of cluster analysis. Consider 12 people in one group where each colour represents a different set of characteristics. Clustering them will split them into four homogeneous clusters as shown in the four coloured squares. Since each cluster is more homogeneous than the 12 people considered as one group, predictive models for each cluster will have much greater accuracy than models built for the group of 12 people as a whole.

Figure 2



It is clear that each cluster has high internal homogeneity and different clusters have high external heterogeneity *with respect to the fields that were used to define the clusters*. The fields used to define the clusters must be chosen carefully. It is best to use fields that define and characterise each record as 'an individual', show variation between the records and measure different attributes of the people or objects (ideally the fields should be independent). Suitable fields for segmenting people may be age, income, marital status, highest level of education, life stage, desires and (frequent) purchases. Examples of data that should not be used are telephone numbers because they do not define people and fields that have very low penetration, for example some hobbies and leisure activities.

Decision Trees

Decision trees is a data mining method for predicting the value of a target variable as a function of a number of input variables as in regression. The model resembles a tree with a number of nodes where each node consists of a number of people or objects that mostly have the same value or similar values of the target variable. The predictor variables and their values vary from node to node depending on how far down the tree the node is.

The values in the tree and the way the nodes are created and defined can be converted into a set of decision rules that can be programmed very easily. The inputs to the program are the target and predictor variables and the outputs are the values of the target variable in each node. The program consists of a set of decision rules that define the route from the top of the tree, i.e. the input data, to each terminal node.

There a number of decision tree models including CHAID, Exhaustive CHAID, CART (Classification and Regression Trees) and QUEST. The models differ in the type (continuous or categorical) of predictor variable they can model and the way they grow the tree.

Example Segmentation

PAM Analytics has carried out many segmentations in a range of business sectors including customer relationship management, air travel and mobile phone usage.