Evaluation of Clusters of Credit Card Holders

Maria Cristina M. S. G. Martins
Habber Tec Portugal, Business & Intelligence Consulting
mcris.martins@gmail.com

Margarida G. M. S. Cardoso
Department of Quantitative Methods. ISCTE Business School
margarida.cardoso@iscte.pt

Resumo

Este trabalho centra-se na avaliação de um agrupamento de clientes de cartões de crédito de uma instituição financeira portuguesa, mediante um processo de validação cruzada, transpondo um procedimento comum no âmbito da aprendizagem supervisionada para a análise de agrupamento (uma metodologia de aprendizagem não supervisionada). Este procedimento de validação cruzada que é proposto é, ainda, trabalhado de modo a adequar-se às condições da amostra de dados usada – conjunto de dados de grande dimensão e utilização de variáveis mistas (numéricas e categoriais). Esta metodologia permite não só a avaliação da solução de agrupamento, mas também ajuda à caracterização dos grupos obtidos. Para além disso, fornece regras de classificação para novos clientes de cartões de crédito. Face aos resultados obtidos, a estabilidade interna é verificada para uma solução constituída por cinco grupos de clientes. Finalmente, são obtidos os perfis dos grupos constituídos sendo, ainda, apontadas possíveis estratégias, no contexto de negócio, a estudar para cada um deles.

Palavras-Chave: análise de agrupamento, avaliação de agrupamentos, estabilidade interna

Abstract

This work is focused on the evaluation of a clustering of credit card holders of a Portuguese financial organization, using a cross-validation procedure which is imported from supervised learning and used for evaluating results yielded by cluster analysis (an unsupervised technique). The proposed approach is conceived to deal with the particular sample characteristics – it handles a large data set and mixed (numerical and categorical) variables. This approach provides both the evaluation of the clustering solution and helps characterizing the clusters. Furthermore, it provides classification rules for new credit card holders. According to the obtained results, the internal stability is verified for a solution with five clusters. Finally, this work presents the profiles of the credit card holders' clusters and suggests some possible strategies to study in each of them, in the business context.

Keywords: clustering, clustering evaluation, internal stability

1 Introduction

When dealing with the subject of evaluation within a supervised analysis one typically recurs to an error function that links the available observations of the target variable and the estimates provided by a proposed (supervised) model. Naturally, there are no available target observations when unsupervised analysis is used, clustering analysis in particular.

In general, the evaluation of clustering analysis results is an attempt to check the quality of the obtained clustering using some indicators of some desirable properties of a clustering solution. This evaluation may be focused on the properties of compactness and separability using specific indices (e.g. the Calinski and Harabasz index, [Calinski and Harabasz 1974]). In addition, it can also address the stability property (internal stability) of the proposed clustering solution, [Gordon 1999]. Under this property, it is assumed that small changes in the clustering procedure should result in approximately the same solution [Milligan 1980].

A cross-validation procedure inspired in the traditional cross-validation procedure used in supervised learning can be used to evaluate the stability of a clustering solution, based on the comparison of two clustering structures formed in a holdout sample (it was first introduced in [McIntyre and Blashfield 1980]).

This work presents a cross-validation methodology to evaluate a clustering of credit card holders (private customers, in particular) of a financial organization operating in the Portuguese market.

The objectives are two-fold:

- On one hand the clustering solution should add a better insight to the market of credit card holders and so helping to support future marketing decisions. As in other service industries, segmentation is a key tool for marketing planning especially in today's highly competitive environment;
- Finally, the proposed methodology should be able to deal with similar applications in diverse contexts.

2 Methodological Approach

2.1 The proposed cross-validation procedure

Stability is a desirable property of a clustering solution, [Gordon 1999]. A stable solution should remain approximately the same when minor changes are made to the clustering procedure. These minor changes may refer, for instance, to the parameterization of the clustering algorithm, to the introduction of some noise in the data or to the consideration of alternative clustering base variables.

A cross-validation approach may address the stability of a clustering solution. It is imported from the supervised analysis and relies on the comparison of clustering structures obtained from sub-samples drawn from the same original sample: a training and a test sample are considered ([McIntyre and Blashfield 1980] and more recently [Cardoso 2007]).

The proposed general cross-validation procedure is described in Table 1. First, we split the original sample into two sub-samples – training and test samples. Second, we cluster the training sample using an appropriate algorithm (as mentioned below). Then, we train a classifier based on the clusters' labels obtained. The results from the classification enable the allocation of new elements (credit card holders) to the clusters and the classifier can be then applied to the test sample to produce clusters that mimic the training sample's clusters. An alternative clustering may be obtained in the test sample using the same algorithm that was applied in the training sample. Finally, we can compare the two clustering structures obtained from the test sample and calculate the stability indicators' values. In fact, the final cross-validation results rely on the indices values concerning the association and the agreement between the two partitions in the holdout sample that supports the evaluation of the clustering solution stability.

Furthermore, it should be noted that the same procedure can be implemented using the training set as the holdout sample (inverse cross-validation).

Table 1. Clustering cross-validation

Step	Action	Output
1	Perform the training-test sample split	Training and test samples
2	Cluster the training sample	Clusters in the training sample
3	Build a classifier in the training sample supervised by clusters' labels; use the classifier in the test sample.	Classes in the test sample
4	Cluster the test sample	Clusters in the test sample
5	Obtain a contingency table between clusters and classes in the test sample and calculate indices.	Indices of association and agreement values, indicators of stability

2.2 The clustering procedure

In steps 2 and 4 (Table 1) a clustering analysis is performed. It is aimed to divide an heterogeneous data set into homogeneous clusters, being the concepts of homogeneity-heterogeneity based on measures of dissimilarity between the values for the attributes of the individuals.

In the present application, a large data set is considered (19 220 credit card holders characterized by multiple attributes with different measurement levels).

The Two-Step algorithm, [Chiu, Fang, Chen, Wang and Jeris 2001], based on BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies, [Zhang, Ramakrishnan and Livny 1996], is chosen to deal with the application in question, essentially for two reasons: be able to deal with large databases due to its incremental nature; and handle mixed variable types using the log-likelihood distance measure that is adequate for dealing with the mixed type of variables.

The Two-Step algorithm has two stages. First, a pre-clustering is performed according to an incremental procedure which produces several small subgroups (called Cluster Feature Entries) and registers their corresponding characteristics originating a tree structure called CFT-Cluster Feature Tree. The second step operates on the Cluster Feature Entries that were yield by the first step and clusters them using a traditional agglomerative clustering algorithm.

Finally, the information criteria – such as BIC – Bayesian Information Criterion, [Schwarz 1978], or AIC – Akaike's Information Criterion, [Akaike 1973] – can be used to automatically help determining an appropriate number of clusters.

2.3 The supervised classifier

In the two-fold validation procedure – step 3 – a supervised classifier is used to learn the clusters that were derived by the clustering algorithm. The CART- Classification and Regression Trees algorithm, [Breiman, Friedman, Olshen and Stone 1984], may be used for this purpose. It is a well known classifier, able to deal with large databases and mixed variables types.

CART methodology relies on binary recursive partitioning of the base data for the construction of a classification tree. The tree is built from a data set gathered in the root tree node. Each node is split into two descending nodes using a splitting variable (one of the predictor variables). The selection of the splitting variable searches for the decreasing of the within-nodes diversity and for the increasing of between nodes diversity, each partition obtained producing a tree with less diversity than the immediately preceding tree. The predictions are finally assessed in each terminal node of the tree using the corresponding modal classes.

2.4 Indices of association and agreement

Having built two clustering solutions in the holdout sample (steps 3 and 4 in Table 1) one has to decide upon the stability of the clustering solution at hand. The indices of association and stability between classes and clusters in the holdout sample may be used as indicators of stability.

The indices of association and agreement can be written based on a contingency table between the two partitions (with K and Q components) being considered (n_{kq} representing the table cells and $n_{k.}$ and $n_{.q}$ representing the table's row and column totals, respectively). Naturally, the number n of observations is given by

$$n = \sum_{k=1}^{K} \sum_{q=1}^{Q} n_{kq} \,. \qquad (\mathbf{1})$$

The Cramer's V is a measure of association based on the well-known Chi-Square statistic - χ^2 - for testing independence between partitions. It ranges from the value zero, when there is no association, to the unit value, when the association is perfect, [Siegel and Castellan 1988], and is given by:

$$V = \sqrt{\frac{\chi^2}{n\,(m-1)}}\,. \qquad (\mathbf{2})$$

where m=min{K,Q}.

When K=Q and after matching the two partitions' clusters, one can consider two alternative measures of association: the Percent Agreement and the Cohen's Kappa [Cohen 1960]:

$$perc = \frac{\sum_{k=1}^{K} n_{kk}}{n}\,. \qquad (\mathbf{3})$$

$$Kappa = \frac{\sum_{k=1}^{K} n_{kk} - \left(\sum_{k=1}^{K} \frac{n_{k.} n_{.k}}{n}\right)}{n - \sum_{k=1}^{K} \frac{n_{k.} n_{.k}}{n}}\,. \qquad (\mathbf{4})$$

The maximum value of Kappa is 1, if the association between the partitions is perfect. If there is no association (other than what would be expected by chance), then Kappa ≤ 0.

The indices of paired agreement quantify the similarity between partitions based on paired comparisons. The Rand index, [Rand 1971], quantifies the percentage of pairs of observations that both partitions agree to cluster together and also to separate:

$$Rand = \frac{\binom{n}{2} + 2\sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2 - \left(\sum_{k=1}^{K} n_{k.}^2 + \sum_{q=1}^{Q} n_{.q}^2\right)}{\binom{n}{2}}\,. \qquad (\mathbf{5})$$

The Rand index does not account for agreement by chance. The adjusted Rand index, [Hubert and Arabie 1985], overcomes this limitation. It has a null value when the agreement between partitions does not exceed the agreement by chance.

$$adj-Rand = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{k=1}^{K}\binom{n_{k.}}{2} + \sum_{q=1}^{Q}\binom{n_{.q}}{2}\right] - \sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}\Big/\binom{n}{2}}\,. \qquad (\mathbf{6})$$

3 Data Analysis

3.1 Clustering base variables

The first step in the data analysis process is concerned with the selection of the clustering base variables. Some general criteria may be used for this selection: the variables should not exhibit

(too much) missing values (to ensure basic data quality) and they should exhibit enough diversity or dispersion in the available data base (to ensure pertinence for the purpose of clustering). Furthermore, considering the business point of view, key attributes should be considered that:

- characterize the behavior and/or the customer's value; and
- can be used in an operational way, for targeting marketing.

After preliminary data analysis and some discussion with the financial company experts – the support and approval from the business experts being very important in the current application – the variables for segmentation were selected (Table 2).

Table 2. Clustering base variables

Customer seniority	Number of defaults with more than 1 month in the last 24 months
Indicator of customer state	Stage of defaults in the last 24 months
Stage of customer with respect to its transactional activity	Proportion of months with card use in the last 24 months
Indicator of internet use as an interface for account management in the last 24 months	Number of purchases in the last 12 months
Sum of the average monthly credit balances in the last 12 months	Value spent in shopping in the last 12 months
Proportion of months with credit balance in the last 24 months	Number of cash advances in the last 12 months
Sum of the average profitability in the last 12 months	Amount of cash advances in the last 12 months
Sum of the average revolving credit balance in the last 12 months	Proportion of leisure expenses in the last 24 months
Proportion of months with revolving credit balance in the last 24 months	Proportion of restaurant expenses in the last 24 months
Number of personal credit contracts in the last 24 months	Proportion of book and video expenses in the last 24 months
Number of directed credit contracts in the last 24 months	Proportion of personal expenses in the last 24 months
Total payments in the last 12 months	Proportion of supermarket expenses in the last 24 months
Most frequently type of payment in the last 24 months	Proportion of travel expenses in the last 24 months

The clustering base variables include different types of information about the customer, such as general customer information, information about his relationship and interaction with the organization, information about balances, profitability, revolving credit, personal credit, directed credit, payments, defaults, card use, purchases, cash advances and some categories of expense. This information was considered enough to characterize the credit card holders taking into account the experts´ knowledge and the focusing on a behavioral point of view.

3.2 The clustering structure: evaluation and profiling

Having decided upon the clustering base variables, the data set was created and the original sample randomly partitioned into a training (11 532 observations) and a test sample (7 688 observations).

The clustering algorithm – Two-Step as previously mentioned – was applied to the training data originating 5 clusters (see Table 3). An outlier "cluster" (denominated "-1") was also identified, because we admitted its existence in the algorithm parameterization.

Table 3. Clusters distribution in the training sample

Cluster	Number	Percentage
1	1705	15%
2	1436	12%
3	3282	28%
4	2984	26%
5	2103	18%
-1	22	0,2%
	11532	100%

The clustering solution with 5 homogeneous clusters as delivered by the Two-Step algorithm was then the focus of the two-fold validation procedure (Table 1).

For the evaluation of the solution with 5 clusters (outliers were discarded) two partitions were obtained in the test set: one trying to import the exact structure built in the training set by means of a supervised classifier (CART) and the other originated directly from the test sample, using Two-Step.

The results obtained in the cross-validation procedures are shown in Table 4. They show good levels of association and (paired) agreement between partitions in the holdout(s) sample(s).

Table 4. Results from cross-validation

Indices		Test sample as holdout	Training sample as holdout
Association	Cramer's V	0,765	0,746
	Percent agreement	0,781	0,763
	Cohen's Kappa	0,723	0,701
Agreement	Rand	0,706	0,680
	Adjusted Rand	0,570	0,529

Once the stability of the 5 credit card holders clusters proved, the analysis proceeds with their profiling. The 5 most discriminating variables according to the CART measure of predictors importance are: Stage of customer with respect to its transactional activity, Proportion of months with card use in the last 24 months, Number of purchases in the last 12 months, Value spent in shopping in the last 12 months and Sum of the average monthly balances in the last 12 months.

In addition, the variables that most contribute to differentiate each cluster were identified by means of chi-square tests (for nominal variables) and t-tests (for metric variables). Finally, the clusters were profiled and named after their characteristics (
Table **5**).

Table 5. Clustering solution

Cluster	Designation	Summary description	Main characteristics
1	**Heavy users**	Customers with an heavy use of the card and other credit elements, especially revolving credit	High use of the card, high use of revolving credit, high profitability and high proportion of months with balance
2	**Credit oriented with some default**	Customers with high revolving credit use, moderate card use and high default	Moderate use of the card, good profitability, high proportion of months with balance, some relevant default
3	**Moderate users**	Customers with moderate card and revolving credit use and with no default	Moderate use of the card, with moderate revolving credit use, some profitability and without risk
4	**Debit oriented users**	Customers with frequent card use but low revolving credit use	High use of the card, with high number of transactions, low use of revolving credit, some profitability, high proportion of months with balance, low number and value of cash advances, very low default
5	**Light users**	Customers with very low card use	Very low use of the card or other credit element, very low use of revolving credit, low profitability, low payment values

3.3 Possible strategies for each cluster

Once obtained the clusters' profiles, it is important to discuss some of the actions that could be implemented in a practical context, as a result of this research work. **Table 6** shows some principal strategies and possible practices to follow for each cluster found. Based on the most important characteristics of clusters and on the business knowledge, derived from the authors' experience and according to some business experts, we could identify the relevant business issues regarding each cluster. As a consequence, we were able to define some orientations to the customer relationship management so that the organization could satisfy its customers, their needs and expectations, and, at the same time, increase its own performance.

Table 6: Principal strategies and possible practices to follow in each cluster

		Principal Strategies			Possible Practices to follow		
Cluster 1: Heavy users	**14,80%**	Offers dinamization	Retention	Design of specific offers	Define triggers to detect a prospective over-leverage situation	Detect previously the lost of involvement	Consider increasing credit limit
Cluster 2: Credit oriented with some default	**12,50%**	Risk control	Stimulation to the personal credit	Detect previously the evolution of defaults	Convert revolving credit in personal credit	Offer transaction stimulation in the leisure category	
Cluster 3: Moderate users	**28,50%**	Stimulation to the revolving credit	Stimulation to the card use	Target campaigns to stimulate the use of revolving credit	Target campaigns to stimulate the use of personal credit	Target campaigns to stimulate the on-going card use	
Cluster 4: Debit oriented users	**25,90%**	Image Promotion	Stimulation to cross-selling	Target charm offers to promote a good image of the organization	Offer non-financial products or services (e.g. convenience)	Create partnerships to develop specific offers to this segment	
Cluster 5: Light users	**18,20%**	Stimulation to the card use	Cross-selling promotion	Promote the free or low cost use of some credit products	Target campaigns to stimulate the card use		

4 Conclusions and Further Research

The objective of the present work was the development of a methodology to build and evaluate clusters of credit card holders.

Evaluation was intended to be focused on the property of internal stability, since the clustering procedure itself provides specific criteria to deal with the properties of compactness (intra-clusters homogeneity) and separability (inter-clusters heterogeneity). Therefore, the clustering solution evaluation rests in a two-fold cross-validation procedure which is imported from supervised learning to the field of clustering.

The cross-validation methodology to apply to the credit card holders data had to deal with a large data base and mixed clustering base variables types. It was found appropriate therefore to use the Two-Step clustering procedure and a CART tree as a supervised classifier.

Finally, the cross-validation approach rested in several indices of association and agreement values (e.g. the adjusted rand index [Hubert and Arabie 1985]) regarding the comparison of two partitions obtained in the holdout sample.

A clustering structure with 5 clusters was obtained. It relies on 26 attributes considered relevant by experts and selected according to some quantitative criteria (missing values and diversity).

The clusters originated by Two-Step and evaluated by means of the two-fold cross-validation procedure are: "Heavy users"; "Credit oriented with some default"; "Moderate users"; "Debit oriented users"; "Light users".

In addition to the proposed methodology direct results, two additional advantages of the proposed validation approach were identified:

- it supports the clusters' characterization since it yields the relative importance of discriminating attributes between clusters (as measured by CART);
- it provides classification rules (CART rules) thus enabling the classification of new customers (card holders) in one of the clusters provided that the attributes considered are available.

Substantive results include a complete clusters' characterization and so originating an improved insight of this market. This, in turn, supports better marketing strategies directed to the identified clusters, thus potentially improving services levels and profitability.

Naturally some limitations can be pointed out to the present work. On one hand, the clustering base variables were limited to the ones available in the company's database. On the other hand, the observed values were collected in a specific point in time (September, 2007) albeit referring to the past / historical behavior of all customers. As a consequence, the obtained clustering structure may register some temporal changes, including the possibility of customers switching from one cluster to another, despite the eventual stability of the clustering base attributes.

Nevertheless, the methodology adopted in the present work is a useful tool for clustering the financial organization customers (credit card holders). Naturally it deserves periodical updates which (for short periods in time) can take advantage of the incremental nature of the adopted clustering algorithm. In the near future it should be interesting to further develop the proposed methodology to deal with dynamical stability (besides the internal stability addressed in this paper) comparing clustering solutions referred to different points in time.

References

Calinski and Harabasz, 1974, A dendrit method for cluster analysis, Communications in Statistics, 3, p. 1-27.

A. D. Gordon, 1999, Classification, Chapman & Hall/CRC.

G. Milligan, 1980, An examination of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrika, 45, p. 325-342.

R. M. McIntyre and R. K. Blashfield, 1980, A nearest-centroid technique for evaluating the minimum-variance clustering procedure, Multivariate Behavioral Research, 2, p. 225-238.

G. Schwarz, 1978, Estimating the Dimension of a Model, The Annals of Statistics, 6, p. 461-464.

H. Akaike, 1973, Maximum likelihood identification of Gaussian autorregressive moving average models, Biometrika, 60, p. 255-265.

L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, 1984, Classification and Regression trees, Wadsworth.

S. Siegel and N. J. Castellan, 1988, Nonparametric Statistics for the behavioral sciences, McGraw-Hill.

J. Cohen, 1960, A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, p. 37-46.

W. M. Rand, 1971, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66, p. 846-850.

L. Hubert and P. Arabie, 1985, Comparing partitions, Journal of Classification, 2, p. 193-218.