Self-Clustering for Identification of Customer Purchase Behaviours

Guillem Lefait[*,**,***] Gilles Goncalves[**], Tahar Kechadi[*]

[*]UCD, Belfield, School of Computer Science and Informatics, Dublin 4, Ireland
guillem.lefait@ucd.ie, tahar.kechadi@ucd.ie
[**]Univ Lille Nord de France, F-59000 Lille, France
[***]UArtois, LG2IA, F-62400, Béthune, France
gilles.goncalves@univ-artois.fr

Abstract. La segmentation d'une base client peut avoir différents objectifs et plusieurs segmentation peuvent être utiles pour décrire les clients ou pour s'adapter avec les stratégies commerciales d'une entreprise. Dans ce papier, nous présentons un schéma expérimental visant à proposer un ensemble de segmentations alternatives. Ces segmentations sont produites sur des données réelles par la transformation des données initiales, la génération et la sélection de différentes segmentations.

1 Introduction

Clustering consists in the creation of groups, such as objects inside same groups are very similar and objects of different groups are very dissimilar Xu and Wunsch (2005). Clustering is used for different objectives : to explore the data set, to condense it into a small set of representative points or to organise the data.

Segmentation is the ability to recognise groups of customers who share the same, or similar, needs McDonald (1996). Not all the customers in a broadly-defined market have the same needs, therefore the segmentation enables companies to provide specific products or services to different segments.

The use of clustering to automatically provide a segmentation is not recent and has been performed for two main objectives : 1) to identify groups of entities that share certain common characteristics and 2) to better understand buyer behaviours by identifying homogeneous groups of buyers Punj and Stewart (1983). However, there are different challenges to address when using clustering to perform the segmentation : which data to select, how many clusters to produce and how to evaluate the clustering results.

Very few solutions have been proposed to evaluate the quality of the customer segmentation. Manual investigation is often the solution used to assess the relevance of the clusters Aggelis and Christodoulakis (2005). In Cheng and Chen (2009), the segmentation result is assessed by the accuracy to predict loyalty of unknown customers. In Chang et al. (2009), sales forecasting is used on the segmentation result.

In this paper, we are interested to create, select and evaluate clustering results that will be presented to experts. The idea presented in this paper is based on the systemic search

of dissimilar but relevant clusters. This approach consists of the generation of thousands of clusters by both transforming the input and using clustering methods with different parameters. These clusters are later evaluated to estimate how useful the clustering results are, and only the most coherent and the most different models are retained. Finally, the selected set of segments is provided to experts with structural information to facilitate the exploration and the comparison of the different segmentations.

The rest of this paper is organised as follows: The components to create and evaluate the clustering results are described in the Section 2. Experimental results with discussion are provided in Section 3, and the conclusion of this paper and perspectives are given in Section 4.

2 Customer Segmentation Architecture

Our objective is to produce automatically diverse and meaningful customer segmentations. The automatic components are described in the following subsections.

2.1 Input Component

The *input* component is in charge of the data transformation. This component is composed of several data transformation methods and may reduce or increase the number of features of the original data set. Moreover, all these transformations techniques may be combined to produce multi-transformed data.

A first data set, ds_ir, is created to gather the Recency (R) value that describes the the last purchase recency, the Frequency (F) and the Monetary (M) values that represent respectively the number of purchases made and the amount of money spent in the period t by a customer.

A second set of data sets is created through the discretisation of the consumer data with the Symbolic Aggregate approXimation (SAX) method Lin et al. (2003). First, the time is discretised and data are separated into w periods where values retained are the average values of the period under consideration. Second, the values are also discretised with an alphabet of size α.

We create a last set of data sets that record the frequency transitions between the symbols identified by SAX. This data set gather relations in the purchase frequencies or the similarities in the consumption rates.

Additional methods to transform the data, such as the smoothing or feature selection methods could also be added.

2.2 Generation Component

The *generation* component is responsible of the creation of the clustering results. It is defined to accept several clustering methods with multiple parameters. When the clustering algorithm presents a stochastic behaviour, it is repeated r times.

We restrict the choice of the clustering algorithms to the algorithms that produce a hard partition because of the usage that will be made of this segmentation. Using a fuzzy clustering results may be very useful if it is combined with another learning technique or when considering one particular customer. However, it is of a limited help if an expert has to explore and analyse the partition.

2.3 Selection Component

The *selection* component's aim is to score the clustering results to keep the more relevant segmentations only.

We used three different approaches to quantify the quality of the clusters.

The first estimator, Q_1, is the coefficient of determination (R^2) and measures the proportion of variability in the data set that is explained by the model. It is defined as :

$$Q_1 = R^2 = 1 - \frac{\sum(p_i - \mu_j)^2}{\sum(p_i - \bar{p})^2} \tag{1}$$

where p represents the purchase information data and μ_j the cluster model associated to the customer i.

The second estimator, Q_2, measures the accuracy of the model given a classification task. Given the RFM Value of each customer (RFM Value = R * F * M), customers received a label (very low, low, average, high, very high) provided the quantile they belong to. The clusters are then assessed by the dispersion over the labels among the clusters. We select the F-Measure, the harmonic mean of the precision and the recall to measure the homogeneity of the clusters. The Precision $P(l, c)$ is the proportion of the customers with the label l in the cluster c. The Recall $R(l, c)$ is the number of the customers with the label l in the cluster c over the total number of customers with label l.

The last clustering quality estimator, Q_3, indicates the accuracy of sales forecast by using the segmentation. The forecast is performed with the exponential smoothing. The sales at time $t + 1$ depends on both the last real value and the last smoothed value. Given a real value v_t, a smoothed value s_t at time t and the smoothing parameter α, the forecast f_{t+1} value at time $t + 1$ is given by :

$$f_{t+1} = \alpha v_t + (1 - \alpha)s_t \tag{2}$$

The parameter α is selected for each segment by internal validation.

We also compute a global coherency estimator, G, that takes into account the quality and the consistency measures :

$$G = C_{Q_1} \times C_{Q_2} \times C_{Q_3} \tag{3}$$

G consist of a pool of measure and similarly to Williams (1999), this architecture could be extended such as each measure receive a weigth that describes its contribution in the discovery of good segments.

For all the clustering, the best results given each of the estimator Q_0, Q_1, Q_2, and G are selected and sent to the next component.

2.4 Visualisation Component

The *visualisation* component is in charge of the graphical representation of the clustering results. It has to provide information both on the clusters and on the estimated quality. Moreover, it has to provide tools to facilitate the comparison between different clustering results.

Intelligent Icons Keogh et al. (2006) are a technique that map the frequency transition between symbols into colours. Then given a frequency matrix, a squared icon can be derived to represent visually the matrix content. Recalling that we have discretised the data with SAX

(a) Customer 27 (b) Customer 94

FIG. 1 – *Intelligent Icons applied on consumer purchase data (brand 1)*

and calculated the frequency transition matrix (ds_isf), we can applied the same process to describe and visually represent the identified clusters.

The Figure 1 demonstrates the efficiency of this representation for two different customers. The purchase and non-purchase events are represented by a square composed of four pixels. Although the information retained has been divided by 15, it still allow the comparison of two consumers.

3 Experimental Results

Experiments were carried out on a data set obtained from the SLDS09 challenge [1]. This data set consists of the weekly purchase log of 10 000 customers over 62 weeks. Purchases were made for 3 brands in two different supermarkets (6 brands in total). The initial objective was to identify brand and/or supermarket specificities. The following investigation is performed only with the log of customer purchases : no additional information is known about the customers nor the brands.

Because of the lack of space, we only present and compare results with $K = 5$. Results for different number of clusters will be put online [2].

First, we present the R^2 results on the 6 brands. On average, the R^2 scores are very low, indicating that very few variance may be explained by the model.

However, we can note two distinct behaviours in the population of the identified clusters. For example, segments of the brands 1, 3, 4 and 5 are clearly separated by the purchase volume. However, when considering the brand 2 and 6, we can see that clusters seems also to have temporal specificities. This indicates that clusters may regroup individuals that made simultanous purchases (peak of one segment at Christmas in the brand 6).

We can also notice that brands 2 and 6 (where segments are partially based on simulatenous purchases) obtain a better R^2 value with segmentation on ds_is, while the others (where identified segments are separated by consumption rates) rely on the transition matrix data ds_isf.

The second result, the RFM Value classification evaluated by the F-Measure are given in the Table 1. It is interesting to note that for the brand 1, 3, 4 and 5, the segmentation that separates the best the customers with their RFM values does not use the RFM data but the transition matrix ds_isf.

When considering the internal class distribution we can see that for the best clustering given Q_2, each of the clusters contains a certain type of customers. We can also note that segments with the customers of extreme RFM values (with very low or very high value) are more consistent.

[1] Symposium Apprentissage et Science des Données 2009, http://www.ceremade.dauphine.fr/SLDS2009

[2] Extra information are available on http://www.emining.fr/data/consumer-behaviour/

Brand	1	2	3	4	5	6
Train	71.00	78.14	77.02	70.41	74.86	77.28
Test	40.96	48.56	43.41	39.94	41.56	48.47
Data	ds_isf	ds_ir	ds_isf	ds_isf	ds_isf	ds_ir

TAB. 1 – *Best F-Measure on labelled data with K = 5*

The performance of sale predictions estimated by Q_3 are very low. The difference with one segment and n is not significant, and moreover and contrarily to the two previous results, there is no correlation between good results in the training set and in the testing set. This result is very disappointing as Q_3 may be a very sensible indicator of the actual segmentation performance. This result indicates that 1) exponential smoothing is not an adapted to perform the forecast on these data or 2) the selection of the smoothing parameter α is not performed properly, *i.e.* the interval validation should be take a longer historic. The predictive power of the segmentation should then be assessed through a more reliable forecasting method, such as methods based on neural networks.

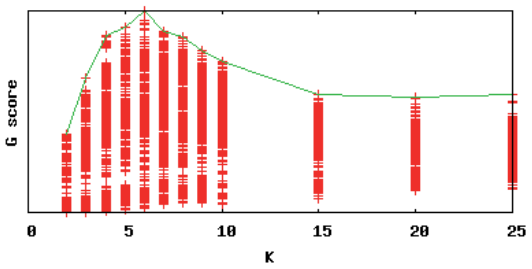

FIG. 2 – *Global score for all the segmentations made on brand 1*

The global score (Eq. 3) is the last method to perform the selection among the different clusters, with respect to all of the estimators defined. The Figure 2 shows the global score distribution for all the clustering results of the brand 1. On this figure, the best global segmentation is one of the segmentations with $K = 6$.

Using Intelligent Icons, it is easy to compare different segmentations altogether by describing the intersection of two segments. This may help to select the appropriate number of clusters : when intersection remains consistent, it is not needed to further increase the number of clusters.

4 Conclusion

In this paper we have experimentally performed a customer segmentation on purchase log data by 1) transforming the data, 2) generating diverse models, 3) selecting the most adequate

models given a set of evaluating functions and 4) creating a visual representation of the segmentations. This segmentation have been performed on a real-world data set and diverse and relevant segmentation models have been selected from a large set of clustering candidates. Our next priority is to improve the sale forecasting by using a more sophisticated approach such as Neural Networks. A second objective is to be able to iterate and favour the data transformation and the clustering methods that have been at the origin of the best clustering results.

References

Aggelis, V. and D. Christodoulakis (2005). Customer clustering using rfm analysis. Stevens Point, Wisconsin, USA. WSEAS.

Chang, P.-C., C.-H. Liu, and C.-Y. Fan (2009). Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. *Knowledge-Based Systems 22*(5), 344–355.

Cheng, C.-H. and Y.-S. Chen (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert Syst. Appl. 36*(3), 4176–4184.

Jiang, T. and A. Tuzhilin (2006). Improving personalization solutions through optimal segmentation of customer bases. In *ICDM*.

Keogh, E., L. Wei, X. Xi, S. Lonardi, J. Shieh, and S. Sirowy (2006). Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems. In *ICDM*.

Lin, J., E. Keogh, S. Lonardi, and B. Chiu (2003). A symbolic representation of time series, with implications for streaming algorithms. In *DMKD*, New York, NY, USA. ACM.

McDonald, M. (1996). The role of marketing in creating customer value. In *Marketing from an Engineering Perspective (Digest No. 1996/172)*, pp. 1/1–111.

Punj, G. and D. W. Stewart (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research 20*(2), 134–148.

Williams, G. J. (1999). Evolutionary hot spots data mining - an architecture for exploring for interesting discoveries. In *PAKDD*, pp. 184–193.

Xu, R. and D. Wunsch (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*(3), 645–678.

Résumé

Segmentation, the process of dividing customers into groups, have long being used by companies to regroup customers with same characteristics. However, with the increasing number of services available, multiple segmentations are now required to describe and fit company strategies to customer behaviours. We present an experimental scheme to produce automatically diverse and meaningful segmentations, by transforming the initial data, generating multiple segmentation and selecting a set of diverse segmentations. This paper presents experimental results on a real-world data set of 10000 customers over 60 weeks for 6 products.