A clustering approach for knowledge discovery in database marketing

M. F. Santos¹, P. Cortez¹, H. Quintela¹ & F. Pinto²
¹Department of Information Systems, University of Minho, Portugal
²Department of Computer Science Engineering, School of Business and Technology of IP Leiria, Portugal

Abstract

From a Marketing perspective, the Customer Relationship Management (CRM) can be viewed as a process, known as Database Marketing (DBM), for establishing a profitable interaction with clients. Currently DBM is mainly approached by classical statistical inference, which may fail when complex, multi-dimensional, and incomplete data is available. An alternative is to use Knowledge Discovery from Databases (KDD), which aims at automatic pattern extraction using Data Mining (DM) techniques.

This paper exploits a clustering approach for DBM, with the intention of finding a set of simple rules which explain clusters of clients with homogeneous behaviours. This strategy was applied in a domestic distribution database taken from a multinational organization. The dataset was created after a direct marketing project, where discount vouchers have been offered to thousands of potential clients, through an own branded magazine. Each coupon included a personal inquiry, with a total of five questions. The final aim was to find the adequate customer profile for each product.

In this study, a specific hygienic product was targeted, since it presented high sales. The aim was to correctly classify which clients use (or not) the voucher, using the five answers as inputs. The work involved validation and elimination of irrelevant data, extensive data pre-processing, data visualization, exploratory clustering using a Self-Organizing Map (SOM), and finally the application of a decision tree in order to achieve the set of classification rules.

Regarding the results, in 60% of the data, a predictive accuracy greater than or equal to 75% was achieved. Furthermore, the readiness of the rules favoured the interpretation of the behaviour of the clients.

Keywords: database marketing, Knowledge Discovery from Databases, Data Mining, self-organizing maps, decision trees.
1 Introduction

Due to the advances in information and communication technologies, corporations can effectively obtain and store transactional and demographic data on individual customers at reasonable costs [1]. The challenge now is how to extract important knowledge from these vast databases in order to gain a competitive advantage [2]. Firms are increasingly realizing the importance of understanding and leveraging customer level data, and critical business decision models are being built upon analyzing such data. Emphasis on customer relationship management makes the marketing function an ideal application area to greatly benefit from the use of Data Mining (DM) tools for decision support. Through DM, organizations can identify valuable customers, predict future behaviours, and make proactive, knowledge-driven decisions. This includes understanding the customers’ preferences through facts and customers’ behaviour through analyzing their transaction data. There has been much research done in this direction, and clustering transactions to learn segments has been one research stream that has generated a variety of useful approaches [3][4].

DM techniques are used in several areas, such as fraud detection [5], bankruptcy prediction [6], intensive care medicine [7], civil engineering [8], just to name a few. Their use for marketing decision support highlights unique and interesting issues such as customer relationship management, real-time interactive marketing, customer profiling and cross-organizational management of knowledge [9].

The Database Marketing (DBM) activity has changed significantly over the last several years. In the past, database marketers applied business rules to target customers directly. Examples include targeting customers solely on their product gap on marketer’s intuition. The current approach, which has widespread use, relies on predictive response models to target customers for offers. These models accurately estimate the probability that a customer will respond to a specific offer and can significantly increase the response rate to a product offering. The old model of “design-build-sell” (a product-oriented view) is being replaced by “sell-build-redesign” (a customer-oriented view). The traditional process of mass marketing is being challenged by the new approach of one-to-one marketing.

DBM departments face several types of business constraints. Typically there are:
- restrictions on the minimum and maximum number of product offers that can be made in a campaign;
- requirements on minimum expected profit from product offers;
- limits on channel capacity;
- limits on funding available for the campaign;
- customer specific ‘do not solicit’ and credit risk limiting rules; and
- campaign return-on-investment hurdle rates that must be met.

Recently, some DM methodologies and applications have been developed to explore the practices and planning methods of sales and marketing management between customers and sellers in the market [10].
In this paper, the DBM process involved a development of models to correctly classify which clients use (or not) a voucher, using five answers as inputs, predicting the customer response, enabling the commercial organization to offer products suitable to the right customers. First, a description of the adopted data is given. Then, a brief presentation of KDD is performed. Next, experiments are presented and the results analysed. Finally, closing conclusions are drawn.

2 Marketing data

In this case study two Databases (DB) of the same organization were used: one containing personal information about registered customers (including the response to the inquiries mailed in direct marketing projects with discount vouchers, and by renting of external DB), and another with the transactional data with the registered discount of the vouchers in the supermarkets.

The first phase of the project consisted in merging of the two DB, in order to make possible the DM task (64 482 customers were registered and 63 961 discount vouchers were mailed). This project focused on the Most Valuable Consumers (MVC), which received the discount voucher of a specific hygienic product (19 382 customers). The classification rule for consumers according to a pseudo-code is depicted below:

\[
\text{If } \text{household} \geq 2 \text{ and dishwater=yes and washing machine=yes} \\
\text{Then } \text{Consumer}= \text{MVC}
\]

The database of this project presented several problems, mainly in terms of the inquiries’ data (Figure 1). The inquiries DB contained several missing data, which had to be treated (described in Section 3). However, this drawback did not occur in the discount vouchers DB, due to the automatic acquisition of the data. After the merging process, a single dataset was created. Table 1 shows a synopsis of the main attributes of this data.

![Figure 1: Data acquisition schema.](image-url)
Table 1: Database attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Household</td>
<td>{Non response, 1, 2, 3, 4, 5, 6 or more}</td>
</tr>
<tr>
<td>2. Dishwater</td>
<td>{Non response, Yes, No}</td>
</tr>
<tr>
<td>3. Monthly Consumption</td>
<td>{Non response, [0…150], [151…350], [351…500], [501…650], [651…]}</td>
</tr>
<tr>
<td>(€)</td>
<td></td>
</tr>
<tr>
<td>4. Household Income (€)</td>
<td>{Non response, [0…500],[501..750], [751…1000],[1001…1500], [1501…2250], [2251…]}</td>
</tr>
<tr>
<td>5. Childs</td>
<td>{Non response, No, Yes}</td>
</tr>
<tr>
<td>6. Number of Childs</td>
<td>{Non response, 1, 2, 3, 4, 5, 6, 7, 8, 9, [10…]}</td>
</tr>
<tr>
<td>7. Voucher Use</td>
<td>{No, Yes}</td>
</tr>
</tbody>
</table>

3 Knowledge Discovery from Databases

The second step consisted in a visualization task, pursued in order to identify the missing data origin [11]. In this task, the following problems were addressed:

a) Several registers of the first six attributes contained blank values. As presented in Table 1, this is missing data, because in the domain it is considered one value (Non response) for the customers that do not respond to the addressed question. This is due to a refusal when some respondents find some questions personally or sensitive (e.g., political, religious affiliation, education level, income, age) and procedural factors (human factor) in the introduction in DB. The impact on Knowledge Discovery from Databases (KDD) process can generate serious problems and is greater when database is constantly refreshed with new data (such as this one).

b) Inconsistency in the related questions: Childs and Number of Childs (Table 2).

c) Too many classes in attribute Number of Childs. A significant bias occurs due to natural and inadequate questionnaire options. A work of levelling was done reducing the number of classes, in order to enhance the learning phase (Figure 2).

Table 2: Answers to questions Childs and Number of Childs.

<table>
<thead>
<tr>
<th>Number of childs (6)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10+ Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>9</td>
<td>4213</td>
<td>4597</td>
<td>929</td>
<td>236</td>
<td>57</td>
<td>24</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>No</td>
<td>222</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1089</td>
</tr>
<tr>
<td>Non response</td>
<td>2</td>
<td>65</td>
<td>56</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Blank</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>77</td>
</tr>
</tbody>
</table>
3.1 Pre-processing

To solve the problems of missing data addressed in previous section, some techniques were used, taking into account several aspects regarding statistical significance issues after pre-processing the data.

The blank values were considered Non Response, assuming an error in introduction of data by the operator, what reinforces the necessity of validation mechanisms in the software used.

![Distribution](image)

Figure 2: Distribution after the discretization of the attribute Number of Childs.

Attributes Childs and Number of Childs were processed together regarding the important issue of concordance between the information of one and another attribute. For example, when Childs equal to No and Number of Childs equal to Non Response (n=3004), the value of Number of Childs was changed to 0. In cases such as Childs equal to Yes and Number of Childs equal to 0 or Non Response, or Childs equal to No and Number of Childs not equal to 0 and Non Response, the cases were discarded. The number of classes of the attribute Number of Childs was reduced (Figure 2), merging the classes 5, 6, 7, 8, 9, 10 or more, into the class 5 or more. After this cleaning stage, the number of cases was reduced to 14710.

3.2 Data mining

All experiments reported in this section were conducted using the Clementine Data Mining System version 6.5 [12].

Initially, a decision tree algorithm, the C5.0 algorithm [13], was applied in the whole dataset, in order to achieve a set of rules which could explain the Voucher Use behaviour in terms of the set of 5 golden questions. However, the algorithm
produced a single naïve rule (\textit{Voucher Use} equal to \textit{No}), which presented a classification accuracy of 75\%, corresponding to the percentage of \textit{No} uses.

In order to improve the predictive results and to obtain more explainable information about the clients’ behaviour, a different strategy was pursued. A clustering approach was adopted, in order to find clusters of clients with homogeneous behaviours. This task consists of two essential modules: one is the exploratory clustering module based on a neural network, a \textit{Self-Organizing Map} (\textit{SOM}), and the other is the rule extraction module by employing a decision tree that can extract association rules for each homogenous cluster. According to different cluster’s characteristics, different marketing strategies could be adopted by making use of the set of classification rules.

![Neural network structure of SOM.](image)

**3.2.1 Self-organizing map**

The SOM architecture [14] [15] was proposed in 1982, being an unsupervised two-layer network that can recognize a topological map from a random starting point. Input nodes and output nodes are fully connected with each other, and each input node contributes to each output node with a numeric weight (Figure 3). In the SOM neural network modelling, several parameters were experimented, being the final topology set to 20 input nodes and 25 output nodes, with a map of a 5x5 grid, corresponding to 25 clusters. Figure 4 plots the number of cases contained in each of these clusters, while Figure 5 shows the distributions of each cluster.

It should be stressed that in 60\% of the data, the \textit{Voucher Use} distribution within a cluster is higher than the original dataset (75\%). This means that in such cases, the prediction given by the SOM is better than the one given by the default C5.0 naïve rule.
3.2.2 Generation of rules

The C5.0 algorithm [13] was applied to each of the 25 clusters (given by the SOM), in order to obtain a set of explanation rules. The training sets used a random sample with 2/3 of the data while the test sets contained the remaining 1/3. The set of classification rules, produced by the C5.0 algorithm, attained a 99.93% predictive accuracy to describe the SOM clustering.

As an example, the rules for Cluster 5, which are simple to understand, are depicted below:

**Rule 1**
Dishwater? Yes
Childs? Yes
4 Conclusions

In this work, a KDD approach was used in a DBM project, regarding discount vouchers of a hygienic product. To achieve this, several pre-processing tasks were required: merging personal and transactional data, reducing/eliminating inconsistencies and levelling attributes. Next, the DM phase was performed, where the use of the clustering approach allowed an improvement, in terms of predictive accuracy, in the majority of the clusters. Then, a decision tree was executed over each cluster, in order to gain a comprehensive knowledge of the customers’ profile assigned to each cluster.

The gain of this approach is in the allowed generation of homogeneous clusters according to personal and transactional characteristics and the rule extraction that characterizes each cluster, what favoured the accuracy of proposed models and the readiness of the rules. These results can help DB marketers to perform their customers/products fragmentation and one-to-one marketing.

The obtained results, although not authoritative, are encouraging. Furthermore, this study allowed the identification of important issues, which need to be improved in future DBM/KDD projects, namely: reviewing of the inquiries questions and options of response, using automatic validation of the responses (in order to reduce inconsistencies and missing data), and merging of data from more databases (e.g., credit card and geographical data).

References


