VTT INFORMATION TECHNOLOGY

RESEARCH REPORT TTE1-2001-29

LOUHI-project

Data Mining Case Studies in Customer Profiling

Version 1.0

11.12.2001

Jussi Ahola and Esa Rinta-Runsala

Version history

Version	Date	Author(s)	Reviewer	Description
0.1-1		J. Ahola & E. Rinta-Runsala	C. Bounsaythip, S. Pensar	First draft to the steering group
1.0	11.12.2001	J.Ahola & E. Rinta-Runsala	S. Pensar	Final version

Contact information

Jussi Ahola, Esa Rinta-Runsala
VTT Information Technology
P.O. Box 1201, FIN-02044 VTT, Finland
Street Address: Tekniikantie 4 B, Espoo
Tel. +358 9 4561, fax +358 9 456 7024
Email: {Jussi.Ahola, Esa.Rinta-Runsala}@vtt.fi
Web: http://www.vtt.fi/datamining/

Last modified on 11 December, 2001

The information in this document is subject to change without notice and does not represent a commitment on the part of VTT Information Technology. No part of this document may be reproduced without the permission of VTT Information Technology.

UDK
Key Words Data mining, customer profiling, clustering, sequence analysis, logistic regression, decision trees

Abstract

The purpose of this document is to report the results of applying some data mining methods that were reviewed in the earlier customer profiling report [2], to real world problems. The cases in question concern customer data from a bank and a book club.

Two separate approaches are taken in the problem solving. First, an explorative analysis is applied to both cases. The goal of the analysis, as can be expected, is to give insight of the data in order to reveal or extract the relevant structure of and patterns in the data. The specific tasks include clustering of the banking data and sequence mining of the book club transactions.

Subsequently, a predictive analysis of the cases is performed. Unlike the explorative analysis, the prediction involves a target variable to be considered. The general aim of the analysis is to generate a model using examples of the past in order to predict the outcome in the future. The concrete tasks performed are to predict the profitability of the book club members with logistic regression and to classify the customers of the bank with decision trees.

Contents

1 Introduction

This research report is a direct continuation of the customer profiling overview [2] published previously in the same series. The work described in this report consists of applying the methods presented in the overview report to two real world problems. The case studies presented here include customer profiling problems of a bank and a book club. Thus, the report can be considered as an applied study of customer profiling to real world problems.

The work involves two somewhat different approaches for the customer profiling. An exploratory data analysis followed by a predictive one is applied to both cases. The main difference of the two approaches is, that the former involves *unsupervised* learning while the latter applies *supervised* learning. In other words, in the prediction the outcome variable is present to guide the learning process whereas in the exploratory analysis it is not.

The goal of the exploratory analysis is to make sense out of a data set in order to generate sensible models or to find some interesting novel patterns [4]. Thus, the data need to be summarized into an easily understandable form, while preserving the essential information in it. This requires methods that can effectively discover and illustrate the structures in the data. In this report, the exploratory analysis is applied to customer segmentation by clustering and to buying behavior analysis by sequence analysis.

The purpose of the predictive analysis is, as expected, to predict the value of the outcome variable in the future based on the examples of the past. This is done by generating a prediction model for the target variable, with one or several explaining or input variables. Here, the logistic regression is used for predicting customer profitability groups and decision trees for classifying customers into existing segments.

The report is organized as follows. Chapter 2 describes the exploratory analysis, including clustering and sequence analysis. The former concentrates on the K-means algorithm and its application to the banking data, while the latter on the multidimensional sequence mining of book club transactions with MOSE-algorithm. Chapter 3 presents prediction using logistic regression and decision trees. With the former, the main stress is on applying the multinomial logistic regression to customer data of the book club, while the latter focuses on using the decision tree algorithm C&RT with the customer data of the bank. Chapter 4 recapitulates the properties of and experiences with the used customer profiling methods.

The work done concerning this report was carried out in the TEKES-funded research project LOUHI.

2 Exploratory Analysis

2.1 Clustering

Grouping of similar observations into separate clusters is one of the fundamental tasks in exploratory data analysis. Depending on the form of the data, the clustering can be done by using central or pairwise clustering techniques. Central clustering techniques minimize the average distance between an observation and its cluster center. Thus, the clustering solution can be described by means of cluster centroids.

The other possibility to cluster observations is pairwise clustering, where the dissimilarities between the observations are exploited. In pairwise clustering, the clusters are formed by minimizing the average dissimilarity between the observations within the same cluster.

2.1.1 K-means

K-means algorithm [5] is one of the most widely used central clustering techniques. In the algorithm, the data set is divided iteratively into K clusters by minimizing the average squared Euclidean distance between the observation and its cluster center. The algorithm starts with assigning K observations as initial cluster centroids and assigning all the observations to the nearest cluster. After this new clustering, the centroids are calculated as means of the observations belonging to that cluster. The observations are assigned again to the new clusters, and new cluster centroids are once again calculated. This iteration procedure is continued until the centroids stabilize (Figure 1).

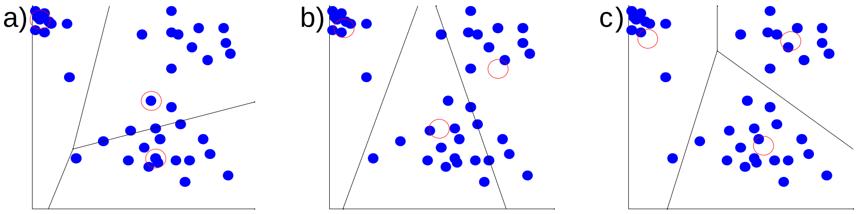


Figure 1. Illustration of K-means algorithm with 40 observations and K=3: a) observations (blue dots), the initial cluster centroids (red circles), and cluster boundaries b) cluster centroids and boundaries after the first iteration and c) the final cluster centroids and boundaries.

The K-means algorithm alone does not give any measure of the goodness of the clustering or the sufficient number of clusters. A common measure for the goodness of the clustering solution is an R-squared statistic [5], which is the rate of sum of squares between the clusters and the total sum of squares. This gives an idea how much of the variation in the data is explained with the clusters found. R-squared values near one indicate compact and separate clusters, whereas values near zero indicate bad clustering solution and/or clusterless data. The R-squared values increase with the number of clusters and there is no general rule for the optimal number of clusters, but it depends on the problem on hand.

2.1.2 Case Study: Bank

This case study focuses on the use of K-means clustering of bank customer data. The goal of the case study is to cluster bank customers into clusters with similar paying behavior. In this context, paying behavior refers to the use of cards, cash dispensers, on-line banking terminals, internet banking, direct debiting, bank transfers, withdrawals from branch offices, etc. Paying behavior of each customer is described with 11 variables, which tell the number of times the customer used each of the paying channels (cards, cash dispensers, etc.) during a period of six months. Thus, the scale of the original variables is from zero to infinity.

As we are interested in the paying habits, not the activity of paying, the absolute usage times are normalized with the total number of usage times of the customer. This gives 11 new variables describing the percentage of each of the paying channels from the total use of the customer. An additional advantage of these new variables is their common scale from zero to one giving each of the variables the same weight.

K-means clustering is done with three different values of K, namely 5, 10, and 20. The analysis was conducted with SPSS software. The algorithm is sensitive to the selection of the initial cluster centroids and therefore the clusterings are done with different initial values. Clustering with K=5 is done with 13 different initial centroid values, clustering K=10 with two different initial centroid values and clustering K=20 once. With each of the K values, the clustering with the greatest R-squared value is chosen for further analysis (Figure 2). However, the clustering solutions are very similar in all the cases where the number of clusters is the same, with a maximum of 100 customers clustered differently depending on the solution.

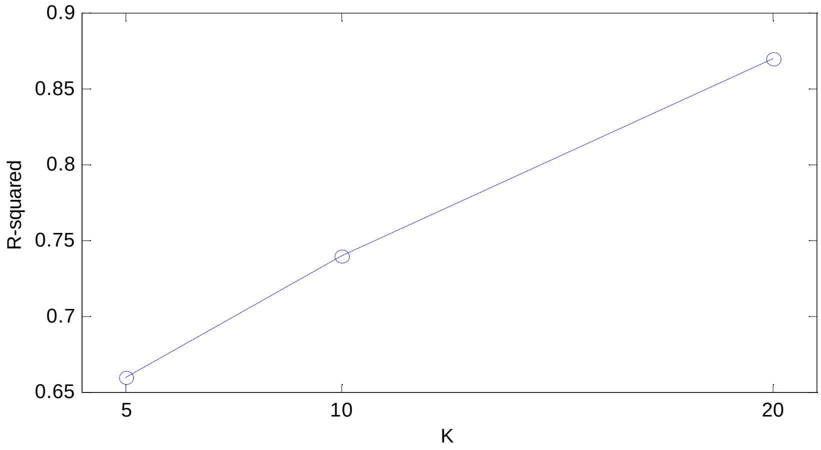


Figure 2. R-squared values of the K-means clustering solutions for K=5, 10, 20.

The clustering with K=5 results into four homogenous and one more heterogeneous clusters (

Figure 3). The homogenous clusters are named according to their most prominent feature as cash, web, ATM, and card clusters. The last cluster contains customers with miscellaneous paying habits. The differences of the clusters can be seen in Figure 4 where the deviations of cluster centroids from the total means are drawn. The deviations are measured in terms of standard deviations.

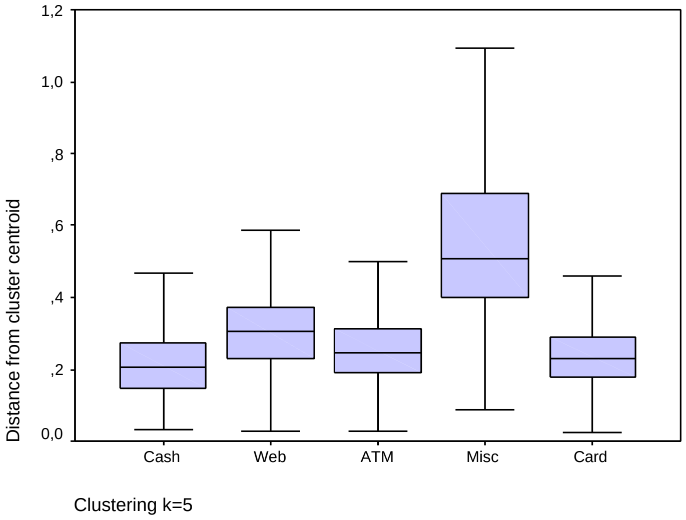


Figure 3. Distribution of distances from cluster centroids within clusters. The box indicates the upper and lower quartiles.

Increasing the number of clusters to 10 results in splitting of the miscellaneous cluster into smaller ones with more distinct features. With 10 clusters, the average distance of an observation from its cluster center is about the same in all the clusters, suggesting 10 to be a sufficient number of clusters. The clusters themselves can be described as above with their centroids' deviations from the total means (Figure 5). As can be seen from the figure, the four centroids of cash, web, ATM and card clusters remain the same and relatively near the total means. However, the clusters resulted from the splitting of the heterogeneous cluster are more deviant in nature. Their maximum deviations are more than 4 standard deviations from the total mean and this gives rise to an idea, that the miscellaneous cluster is not one big cluster of heterogeneous customers, but in fact a group of smaller but more extreme clusters.

Increasing the number of clusters to 20 splits the clusters further. In the K=20 solution, none of the original five clusters remains untouched, but all of them are split into two or more (Figure 6). However, it has to be mentioned that K-means clustering with different values of K is not hierarchical in the sense that clusters would split purely to subclusters. There are always some observations, which jump from one cluster to another, e.g., from the cash cluster to the ATM cluster when K is increased from 5 to 10. These observations lie on the border of the clusters and they do not belong clearly to either of the clusters.

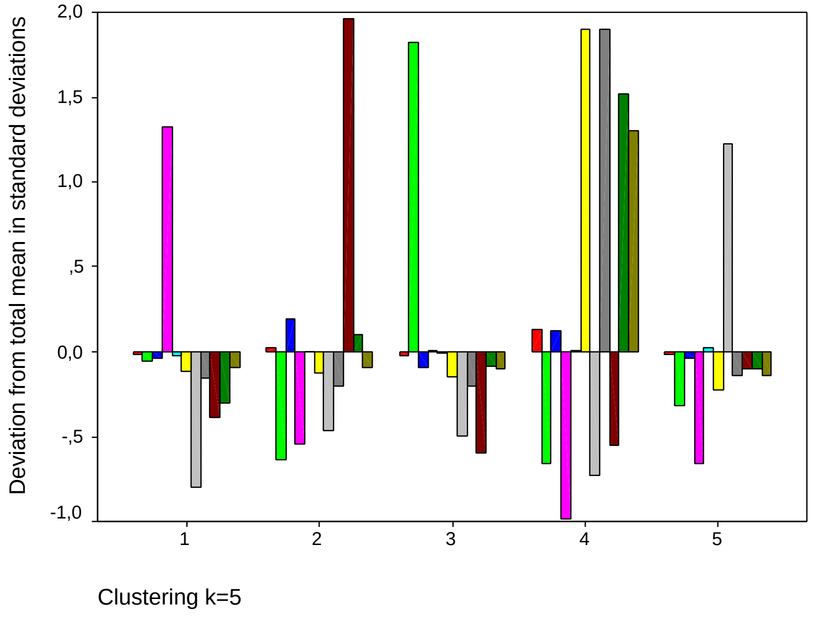


Figure 4. Distances of cluster centroids from the total means in the K=5 solution. The distance measurement unit is standard deviation. E.g., in cluster 1, usage of cash dispenser (purple bar) is more common than in general. On the other hand, the use of credit card (light gray bar) is more uncommon.

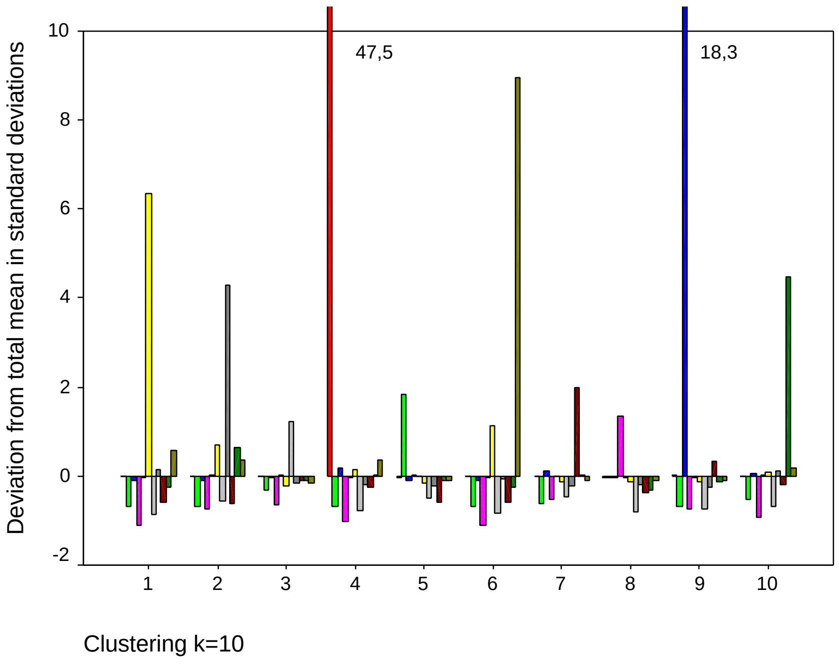
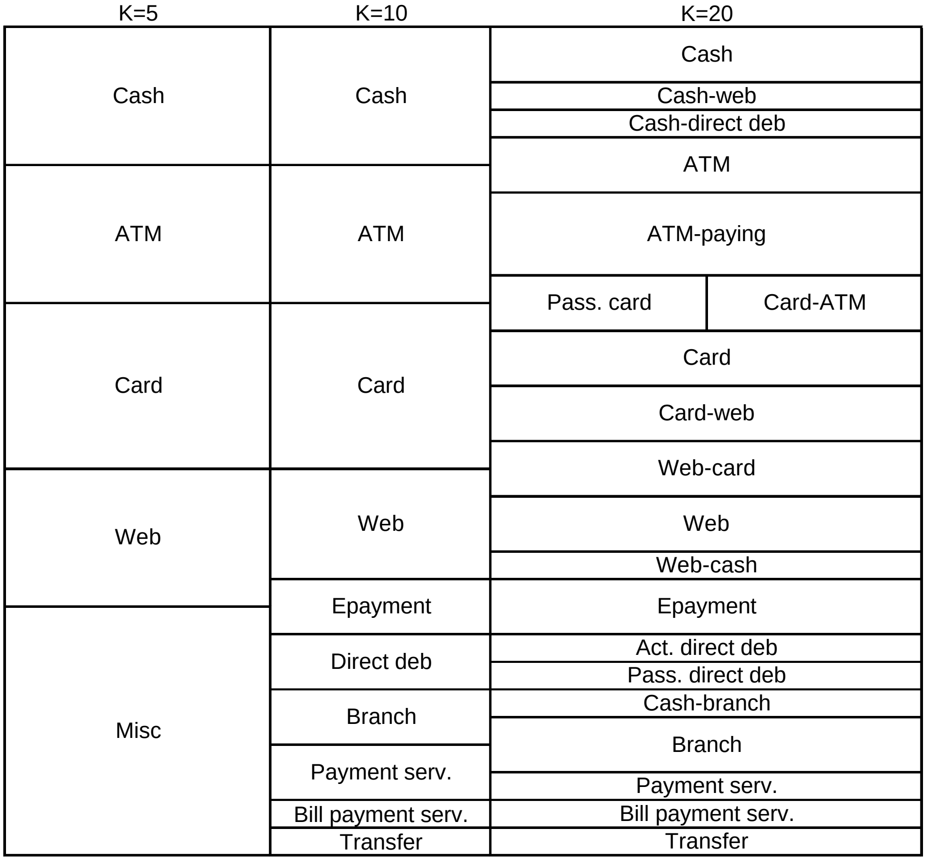


Figure 5. Distances of cluster centroids from the total means in the K=10 solution. The distance measurement unit is standard deviation.



K=5	K=10	K=20	
Cash	Cash	Cash	
		Cash-web	
		Cash-direct deb	
ATM	ATM	ATM	
		ATM-paying	
Card	Card	Pass. card	Card-ATM
		Card	
		Card-web	
Web	Web	Web-card	
		Web	
		Web-cash	
Misc	Epayment	Epayment	
	Direct deb	Act. direct deb	
		Pass. direct deb	
	Branch	Cash-branch	
		Branch	
	Payment serv.	Payment serv.	
	Bill payment serv.	Bill payment serv.	
	Transfer	Transfer	

Figure 6. The splitting of the clusters with increasing values of K.

2.1.3 Experiences

Most experiences from the case study are problem dependent. Some general notions about the use of K-means consider the determination of the number of clusters. As mentioned before, there is no clear rule for the optimal number of clusters. The R-squared measure is often useful in measuring the goodness of the clustering. However, in this case it increases nearly linear with respect to the number of clusters, so it does not provide much a good indication for the determination of optimal number of clusters. Additionally, the useful number of clusters is partially determined by the application the clustering is used later on, and this is not necessarily the same as the theoretically best number of clusters.

Next, some general notions about clustering applications and especially on K-means are discussed. A feature of K-means is its tendency towards clusters of same size. Thus, the clustering solutions achieved may be biased and give a wrong image of the data.

Why then use K-means? The reasons are practical. K-means has good scalability properties. This means the computational demands of the algorithm stay reasonable with increasing number of observations, which is very important in data mining applications. Another advantage of K-means is that it is included in many commercial data mining software packages and thus is easily implemented in real-world applications.

2.2 Sequence Analysis

The idea of sequence analysis, or sequence mining, is to discover sequential rules that identify time dependent patterns of behavior. Thus, it is basically association rule mining with time taken into account in the discovery process. An example of a sequential rule, or pattern, is presented in Figure 7. It can be seen that the rule consists of elements, which occur, in a specific order. The rule of the figure can be interpreted as: the occurrence of *A* will be followed shortly by the occurrence of *C* and *B* together, after which *D* occurs, followed by the occurrence of *F*, *E*, and *G* again together.

As illustrated in the figure, there are several constraints associated with the rules. They correspond to the computation of the frequent pattern occurrences and are used to limit the discovery of unimportant rules and to boost the operation of the sequence mining algorithm. The constraints limit:

- the maximum duration of the entire rule (*ms*),
- the maximum time difference of the independent events, within which the events are considered to occur together, i.e., to belong to the same element (*ws*),
- the maximum time difference of the consecutive elements, within which their co-occurrence is still considered interesting and thus included in the rule (*xg*),
- the minimum time difference of the consecutive elements, within which their co-occurrence is considered uninteresting and thus not included in the rule (*ng*).

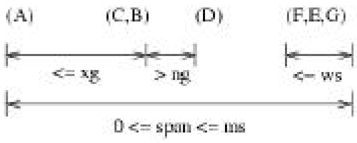


Figure 7. An example of a sequential pattern [3].

Sequence mining can be applied to any problem, which involves data with inherent sequential nature, i.e., it is composed of discrete events which have a temporal/spatial ordering. This kind of data can be obtained, e.g., from telecommunications networks, electronic commerce, www-servers of the Internet, and various scientific sources, like gene databases. More detailed information about the sequence mining can be found in [1] and [2].

2.2.1 MOSE Algorithm

The major drawback with sequence mining, as presented above, is that it does not make use of the possible attributes of the input events. That is, in many cases, the events have additional descriptive attributes and the traditional sequence mining algorithms cannot, as such, utilize the additional information they provide. Instead, the problem has to be solved by pre-processing the data, which tends to lead to an insufficient and/or inefficient solution.

The MOSE algorithm was designed to implement sequence mining for multi-attribute event data. The algorithm considers the input data to consist of sequences, each of which is identified with an *object-id*. The sequences include an arbitrary number of events, which are identified by *event-id*s and ordered temporally using a *timestamp* associated with them. Furthermore, each event may have an arbitrary number of *event attributes*, whose values are defined as discrete classes. The instances of the same event are required to have the same number of attributes, although their values are allowed to be missing.

The algorithm involves several parameters, with which the user governs the mining process. First, the way in which the number of pattern's occurrences is counted has to be determined. The available alternatives are *COBJ* and *CDIST*. The former increments the pattern counter by one, if a pattern is contained in an input sequence (regardless how many times it occurs), while the latter increments the counter by the number of the distinct occurrences of the pattern in the input sequences. Secondly, the minimum number of occurrences for the discovered patterns is defined. This so called *minimum support* (*minsup*) defines the limit for the patterns being frequent enough to be interesting, and it can be presented as either an absolute number or a relative percentage of the occurrences. Furthermore, the support can be defined in three levels, one for limiting the frequent events, attributes and sequences, respectively. Finally, the time constraints described above are also included in the pattern discovery.

Given these user-specific parameters, the algorithm results in all sequential rules that are frequent enough. The discovered rules represent different levels of generality. The most general rules are similar to those of traditional sequence mining. Thus, they are of the form: "*the occurrence of A is followed by the occurrence of B*". Let us assume that the event *A* correspond to the symptoms of cough and *B* to the raise of fever. Then the rule can be expressed with words as*: "a patient, who has a cough, will probably develop a fever in the near future*".

On the other hand, the more specific rules include some or all of the attribute values for the events. They can be presented, e.g., as: "*the occurrence of A with the first attribute value of one and the second attribute value of three is followed by the occurrence of B with the first attribute value of two"*. Applied again to the medical diagnostics, the rule could be something like "*a patient, who has a hard and mucous cough, will probably develop a high fever in the near future*". Here the attribute values "*hard*" and "*mucous*" correspond, respectively, to the categories 1 and 3 of the first and second attributes of the event *A* (symptoms of cough). Similarly, the value "*high*" corresponds to the category 2 of the first attribute of the event *B* (raise of fever).

The algorithm outputs also, for each sequential pattern, its support and *confidence*. The latter specifies, when the rule is divided into antecedent and consequent parts, the probability of the occurrence of the consequent provided the occurrence of the antecedent.

2.2.2 Case Study: Book Club

In the case study, the data consisted of the transactions of about 7000 book club members from their first six membership months. The transactions were purchases, payments, cancellations, or returns of the products, and they were ordered temporally for each customer. Thus, the used timestamp was the transaction month, and the event id was the product category. Each product had also additional attributes like the price group, sales channel and transaction type (sales, returns). Additionally, the data was divided into groups based on the customers profitability after a year of their membership. Therefore, the aim of the sequence mining was to detect the typical buying behaviors of the customers in the different customer profitability groups.

The MOSE-algorithm was used for discovering the sequential patterns from the input data sets. The maximum duration, as well as the maximum time difference between the elements, of the patterns was obviously six months and the transactions occurring within the same month were combined in the same element. So, the used time constraints were: *ms*=6, *xg*=6, *ng*=0. Furthermore, it was decided that an event or a sequence was considered interesting if it occurred in more than 65% of the customers in that profitability group. Hence, the relative *minsup* of 0.65 for all levels with *COBJ* counting was used.

As a result, when the first customer profitability group ("adequate customers"), was considered, almost 400 frequent sequential rules were discovered. Obviously, this is too much for a practical interpretation of the results. Moreover, it turned out, that there is much repetitious and unrevealing information in the rules. Thus, all the rules including redundant and non-informative attribute values could be removed. The resulting rule table is shown in Figure 8. From the rules can be drawn, e.g., following conclusions:

- The customers who are adequately profitable (after a year of membership) buy, within their first half year in the book club, mainly domestic or translated fiction or cooking and handicraft literature as an introductory offer from the backlist.
- They also often buy two translated fiction books in a month.
- After buying translated fiction or cooking and handicraft literature, probably from the back list, they frequently buy translated fiction again in the near future.

2.2.3 Experiences

The MOSE-algorithm seems to function adequately well and result in the correct sequential rules. The performed analyses indicate that both the usability and usefulness of the algorithm seem promising for the real world problems. However, the algorithm is still in the developing stage, so currently it has several deficiencies. First, the algorithm is computationally too ineffective. Both the time and memory consumption hinder applying the algorithm in the problems with large amounts of input data. In practice, the algorithm also discovers too many sequential rules. Thus, the really interesting rules are hard to detect from the extensive output rule table. Furthermore, using minimum support alone as the measure of relevance is not sufficient in many applications, e.g., when the interesting rules are those occurring rarely. Finally, at this stage of development, the constraints implemented are not versatile enough. Especially, involving event and attribute constraints should be considered, in order to mask out uninteresting or include relevant events/attributes.

```
Frequent events:
domestic fiction (78%, 100%)
translated fiction (96%, 100%)
cooking and handicraft (88%, 100%)

Frequent event attributes:
translated fiction[other order] (77%, 100%)
translated fiction[intro] (77%, 100%)
translated fiction[other order, intro] (77%, 100%)
translated fiction[bl sell 1] (77%, 100%)
translated fiction[other order, bl sell 1] (77%, 100%)
translated fiction[intro, bl sell 1] (77%, 100%)
translated fiction[other order, intro, bl sell 1] (77%, 100%)
cooking and handicraft[other order] (78%, 100%)
cooking and handicraft[intro] (78%, 100%)
cooking and handicraft[other order, intro] (78%, 100%)
cooking and handicraft[bl sell 1] (78%, 100%)
cooking and handicraft[other order, bl sell 1] (78%, 100%)
cooking and handicraft[intro, bl sell 1] (78%, 100%)
cooking and handicraft[other order, intro, bl sell 1] (78%, 100%)

Frequent sequences:
translated fiction, translated fiction (74%, 77%)
translated fiction -> translated fiction (77%, 80%)
translated fiction[bl sell 1] -> translated fiction (69%, 89%)
translated fiction[other order, bl sell 1] -> translated fiction (67%, 86%)
cooking and handicraft -> translated fiction (74%, 83%)
cooking and handicraft[other order] -> translated fiction (73%, 93%)
cooking and handicraft[bl sell 1] -> translated fiction (73%, 93%)
cooking and handicraft[other order, bl sell 1] -> translated fiction (72%, 92%)
```

Figure 8 The relevant frequent events and sequences for adequate customers.

3 Predictive Analysis

3.1 Logistic Regression

3.1.1 General

Logistic regression is an expansion of the normal regression, so it generates a model between one or more input variables and a target variable [6]. The model defines coefficients for each input, or predictor, which somehow indicate its effect on the target, or dependent variable. The difference between the normal and logistic regression is that in the latter the model output is not directly the sum of the products of the inputs and coefficients, but the logistic function of that. In more precise terms, the output (*y*) of the normal regression model with *n* inputs (x_1, x_2,..., x_n) is obtained using a formula:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n ,$$

whereas in the logistic regression, the corresponding formula is:

$$y = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}} .$$

The output of the logistic function is evidently always between zero and one and can thus be interpreted as a probability. Therefore, the model is especially useful for situations in which the presence or absence of a characteristic or outcome is to be predicted. The predictors may be numerical or categorical variables, but the target should be dichotomous. Furthermore, the coefficients cannot be interpreted, as such, as the importance measures of the inputs on the output. Instead, they can be used for determining how much the inputs increase the odds of the output occurring. For example, if one would like to predict which lifestyle characteristics are risk factors for coronary heart disease, a model could be generated with the patient data, where the inputs could be binary variables like smoking, diet, exercise, alcohol usage etc., and the output the diagnosis of the disease (yes or no). The model then gives based on the person’s habits of living, the probability for him/her to develop a coronary heart disease. Furthermore, the model coefficient can tell, e.g., how much more likely smokers are to develop the disease than non-smokers.

There exists also an expansion of the logistic regression where the dependent variable is not restricted to two categories. This *multinomial logistic regression* method is suitable for problems, where the subjects should be classified into several groups. It generates several non-redundant models. The number of models is one less than the number of the output categories, so, e.g., for five target groups, four models are generated. In practice, one category is selected as a base category, to which the other categories are compared. The exact formula for the output of the *i*th model, when category *k* is selected as the base, can be written as:

$$y = \frac{P(y=i)}{P(y=k)} = e^{b_{i0}+b_{i1}x_1+b_{i2}x_2+\ldots+b_{in}x_n}$$

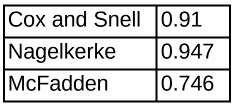
The output presents now the ratio of the probability of the subject belonging to the *i*th category compared to the probability of it belonging to the *k*th, or base, category. So, values over one indicate that the subject is more probable to belong to the category *i* than *k*. However, the interpretation of the model coefficients becomes slightly more complicated, since the comparison to the base category has to be taken into account. So, the model coefficients can now be used to tell how much the inputs increase the odds of the subject belonging to the *i*th rather than to the base category. For example, if one would like to predict, what kind of film a moviegoer is likely to see, a multinomial logistic regression can be performed with the customer query data using variables like customers age, gender, income, dating status etc. as inputs and the type of film they prefer as the output. If romance film type is used as the base category, then one model may result in a ratio, telling how much more likely a person is to watch action than romance films, and another may result in a similar ratio of drama films compared to romances. Furthermore, the model coefficients of the former model can tell, e.g., how much more likely it is for men, compared to women, to watch an action than a romance film.

3.1.2 Case Study: Book Club

In this case study, the aim is to predict the customer’s future profitability, based on his/her demographic information and buying history. Thus, the output variable is naturally the customer’s profitability group after a year of membership in the book club. The inputs include personal information like gender and postal code, and also variables preprocessed from the transaction data. Due to their dynamic nature, the transactions are not, as such, usable for the logistic regression, so they have to be transformed into static variables. Thus, for each customer the number of different purchases (or returns) is counted from the transactions of her/his first half-year as a member. The result includes the amounts of the different product types, products in different price groups, different sales channels and transaction types involved etc. Thus, the input variables include both continuous (counter) and categorical (demographic) data.

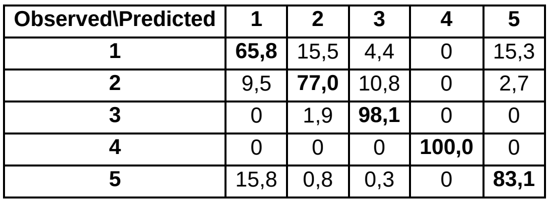
The multinomial regression tool of SPSS was used with this data. The data was first divided into independent training and test sets. The customer profitability group has five distinct values, so the multinomial regression resulted in four models. The group 5, “new customers”, is used as the base group. Thus, the output of the resulting preliminary models was the probability of the profitability group per the probability of the new customer’s group. As an example, the first model tells how likely a customer belongs to the “adequate customer” rather than to the “new customer” group.

The regression tool provides several measures of the goodness of the fit performed by the logistic regression. Two of them are shown in the tables of Figure 9 and Figure 10. The pseudo R-Square measures the variance of the dependent variable explained by the regression models. The closer the value is to one, the better the variance is explained. It can be seen from the table, that the models can explain most of the variance, which implies a good fit. The classification feature of the models can be achieved by selecting for each customer the model with the highest probability. Furthermore, if the probability ratio of all models is less than one, the base group is selected. The classification percentage matrix tells how accurate these classifications are. It is a cross-tabulation of the real and predicted customer profitability groups. The percentages of the correctly classified customers are presented on the diagonal of the table. It seems, that the misclassification percentage (on average 15.2 % of the customers are classified into a wrong category) is not very good, but reasonable for the purpose.



Cox and Snell	0.91
Nagelkerke	0.947
McFadden	0.746

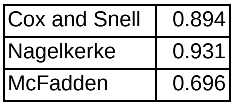
Figure 9 Pseudo R-Square the preliminary models.



Observed\Predicted	**1**	**2**	**3**	**4**	**5**
1	**65,8**	15,5	4,4	0	15,3
2	9,5	**77,0**	10,8	0	2,7
3	0	1,9	**98,1**	0	0
4	0	0	0	**100,0**	0
5	15,8	0,8	0,3	0	**83,1**

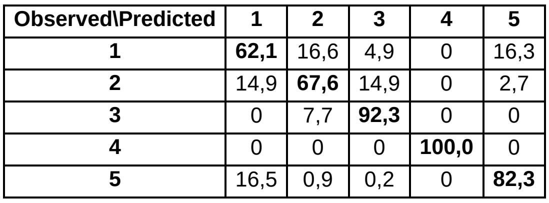
Figure 10 Classification percentage matrix of the preliminary models.

The tool also provides extensive statistics of the input variables for each model. From the analysis point of view, an interesting statistics is the variable significance. Technically, it defines the significance level for the Wald statistic of the variable in that model. In practice, it measures whether the coefficient of the variable really differs from zero. Traditionally, only the variables having a value less than 0.05 are considered to be significant. This property can be used also for reducing variables from the models. That is, the variables having the significance over 0.05 *in every model* can be removed from the data and a new regression can be performed with the resulting new data. In practice this leads to an iterative process, where the procedure is repeated until no variables can be removed from the data. This procedure is used here to reduce the dimensionality of the input data. So, in the final models there are less than 90 variables (including categorical variables, whose all different categories were coded into separate variables), while the preliminary models included about 700 of them. In Figure 11 the pseudo R-square and in Figure 12 the classification table of the final models are presented. When compared to the statistics of Figure 9 and Figure 10 respectively, it can be stated that the diminished dimensionality does not substantially deteriorate the resulting models.



Cox and Snell	0.894
Nagelkerke	0.931
McFadden	0.696

Figure 11 Pseudo R-Square of the final models.



Observed\Predicted	1	2	3	4	5
1	**62,1**	16,6	4,9	0	16,3
2	14,9	**67,6**	14,9	0	2,7
3	0	7,7	**92,3**	0	0
4	0	0	0	**100,0**	0
5	16,5	0,9	0,2	0	**82,3**

Figure 12 Classification percentage matrix of the final models.

Furthermore, the resulting significant variables for the first model are presented in Figure 13. The figure presents the variables, their explanation, significance, and effect on the model output. For the transactional variables (*KKIRJA...KALENTER*), the interpretation of the effect column is, that increasing their count by one, changes the probability ratio (member is an adequate per he/she is a new customer) by the amount shown in the *Effect*-column. In other words, it tells how much more (or less) likely the customers belonging to the first profitability group than the base group becomes. The values greater than one indicate that the increase in the variable value increases the probability of the first group, while the values less than one indicate increase that of the base group. Thus, it can be stated based on the figure, e.g., that customers who buy CD-ROMs within their first six months of the membership, are 11.435 times more likely to become adequate customers than remain new ones after the whole year. On the other hand, for categorical values the interpretation is not quite as straightforward, because they are coded against their base category. That is, e.g., the gender variable “female” is defined to be the base category, so the interpretation has to be done against it. Thus, it can be said, that males are 1.374 times more likely than females to belong to the adequate-group that to the new-group. Furthermore, the model claims that customers having the postal code 705 are (1/8.12e-2=) 12.315 times more likely than those having the base postal code (999 = other postal code) to remain new than become adequate customers. Similar analysis was also performed for other three models resulting in a set of significant variables and their effects on the customer profitability group after their one year of membership.

Finally, the independent test set is used for validating the performance of the models. It turns out that the resulting classification behavior resembled that of the training data. Although the classification error was considerably larger, the misclassifications tended to fall pretty much into neighboring categories. However, one must take into account, that the size of the test set was not large enough, which increases its misclassification percent incorrectly.

Variable	Explanation	Sig.	Effect	Variable	Explanation	Sig.	Effect
Intercept	intercept term	0		[POSTINUM=55]	postal code = 055	0.024	0.52
KKIRJA	ordered as book of the month	0	0.424	[POSTINUM=152]	postal code = 152	0.036	4.148
NORMHIN	normal price	0	1.582	[POSTINUM=205]	postal code = 205	0	0.217
BONUSHIN	bonus price	0	0.508	[POSTINUM=207]	postal code = 207	0.001	0.228
TARJHIN	bargain price	0.001	1.307	[POSTINUM=300]	postal code = 300	0	0.29
VAIHTEHT	alternative price	0	0.618	[POSTINUM=331]	postal code = 331	0	4.434
KKMYYNTI	book of the month purchase	0	4.316	[POSTINUM=332]	postal code = 332	0.034	0.474
BLMYYNTI	backlist purchase	0	1.536	[POSTINUM=335]	postal code = 335	0.022	0.496
LAPSNUOR	children's and youth literature	0.037	0.887	[POSTINUM=338]	postal code = 338	0.001	0.294
TAIDE	art literature	0.041	0.487	[POSTINUM=339]	postal code = 339	0.026	0.339
UUSMEDIA	cd-roms	0.04	11.435	[POSTINUM=340]	postal code = 340	0	0.504
SANAKIRJ	dictionaries	0.018	0.839	[POSTINUM=404]	postal code = 404	0.044	9.69E-02
TAVARAT	wares	0.001	0.74	[POSTINUM=410]	postal code = 410	0.045	0.653
KALENTER	calendars	0.014	0.628	[POSTINUM=450]	postal code = 450	0	2.75
[SEXC=1]	gender = male	0.003	1.374	[POSTINUM=530]	postal code = 530	0	0.344
[SKEY=00E]	skey-code = 00E	0.039	1.536	[POSTINUM=570]	postal code = 570	0.019	0.428
[SKEY=00G]	skey-code = 00G	0.005	1.23	[POSTINUM=650]	postal code = 650	0.002	0.367
[POSTINUM=2]	postal code = 001	0	0.247	[POSTINUM=705]	postal code = 705	0.01	8.12E-02
[POSTINUM=3]	postal code = 003	0	0.147	[POSTINUM=810]	postal code = 810	0.048	0.586
[POSTINUM=8]	postal code = 008	0.001	0.355	[POSTINUM=840]	postal code = 840	0.035	0.565
[POSTINUM=23]	postal code = 023	0.002	0.297	[POSTINUM=901]	postal code = 901	0	0.32
[POSTINUM=27]	postal code = 027	0.036	0.449	[POSTINUM=910]	postal code = 910	0.023	0.616

Figure 13 The significance and effect of the interesting variables of the first model.

3.2 Decision Trees

Decision trees are intuitive and simple tools for classification of observations. The general idea and properties of decision tree methods are presented in [2].

3.2.1 C&RT Algorithm

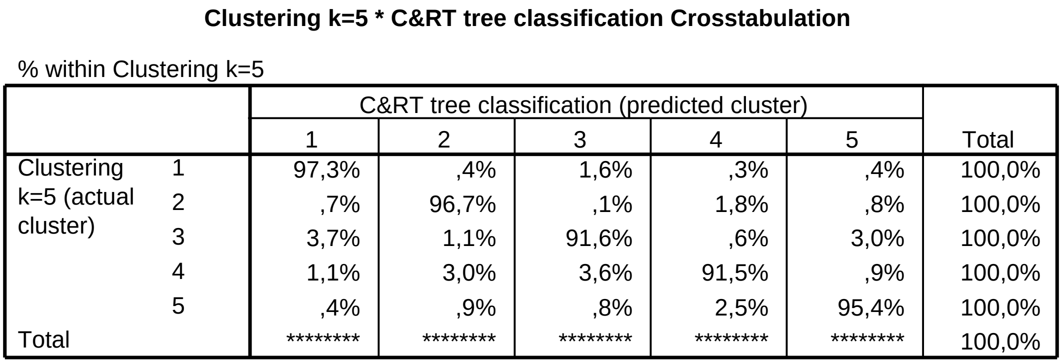
Classification and regression trees (C&RT) [8] are binary trees that split the nodes according to a certain impurity measure and thus result in more and more homogeneous branches. The impurity measure used in the case study is Gini index [6], which is calculated from the category probabilities of the node. Here categories refer to the different classes the observations may belong to. The Gini index equals zero when all the observations in the node belong to the same category. The maximum value 1-1/*k* is reached, when all the *k* categories are of the same probability, i.e., there is the same number of observations in each category in the node.

When splitting a node, the observations of the same category are tried to put into the same child node. Thus, the impurity of the child nodes is lower than that of the parent node. The split of a node is made in such a way that this impurity decrease due to the split is maximized. A node is split further and further until one of the stopping rules is met. The stopping rules include minimum change in the nodes' impurity with respect to the parent node, maximum depth of the tree, minimum number of cases in the parent node, and minimum number of cases in the child node.

3.2.2 Case Study: Bank

The banking case study involves classification of customers into one of the five clusters specified with K-means algorithm in Section 2.1.1. The predictors are the customer's percentages of using of different paying channels. As the training algorithm is computationally quite demanding, only part of the customers, about 70 000, are used in the training of the tree. The software used is SPSS AnswerTree and the C&RT tree resulting from the training is in the Appendix 1. The classification error of the tree is 4.8%, i.e., 4.8% of the customers are classified into a wrong class. This classification is done with a test set not used in the training of the tree to get an unbiased picture of the classification capabilities of the tree.

When applying the tree to the whole customer data set available, the classification error is about the same, 4.7%. The cross-tabulation of actual and predicted classes for the whole data set is in Figure 14.



Clustering k=5 * C&RT tree classification Crosstabulation

% within Clustering k=5

		C&RT tree classification (predicted cluster)					Total
		1	2	3	4	5	
Clustering k=5 (actual cluster)	1	97,3%	,4%	1,6%	,3%	,4%	100,0%
	2	,7%	96,7%	,1%	1,8%	,8%	100,0%
	3	3,7%	1,1%	91,6%	,6%	3,0%	100,0%
	4	1,1%	3,0%	3,6%	91,5%	,9%	100,0%
	5	,4%	,9%	,8%	2,5%	95,4%	100,0%
Total		********	********	********	********	********	100,0%

Figure 14. Classification results of C&RT tree.

The advantage of decision trees is their ability to produce understandable and easily implementable rules for classification. For example, the rules for the right most leaf node of the tree are:

1. Card use percentage of total use > 34%,
2. Web use percentage of total use > 40%,
3. Card use percentage of total use > 43%,

resulting in a node with 86% of cases in cluster 5, 14% of cases in cluster two, and 0% in other clusters.

3.3 Comparisons and Experiences

The case of an uneven amount of data samples in the target categories brings about problems in both multinomial logistic regression and C&RT. The categories with a large amount of data tend to dominate the model generation, which often results in a high misclassification rate for the more rare categories. Fortunately, the problem can be tackled by either resampling or weighting the data. The latter is usually an advisable approach, since the former involves undesired loss of information.

Relating to the preceding, the small sample sizes degrade the classification accuracy of the presented methods. Since the methods are statistical in nature, the amount of samples in each target category should be sufficiently large. Thus, the amount of data should be enough for generating a representative test set for the validation of the classification result. Yet, too large sample amounts may yield to performance problems with the used algorithms.

Finally, the categorical variables may be problematic for the methods. Especially this takes place, when the amount of categories is extensive, as is the case, e.g., with the postal codes. The C&RT algorithm handles the situation with high cardinality input variables somewhat better than the logistic regression, but still the problem solving is strongly complicated. There are a couple of practicable approaches proposed in literature to overcome this problem. However, the case of high number of categories in the target variable is far more difficult and cannot often be solved at all.

4 Conclusions

In this report, some customer profiling methods were used in two real world applications. Two approaches for the customer profiling were applied: explorative and predictive analysis. Both approaches involved the data mining methods of their own. In general, all the methods used performed fairly well for the tasks in question, both in the bank and book club cases. Furthermore, all the methods, except the MOSE-algorithm, were more or less basic methods, which are implemented in several commercial and freeware software tools. In this work mainly statistical software package SPSS was used.

From the experiences of applying different data mining methods to customer profiling, raised several aspects to be considered in the planning and executing mining of real world problems.

First of all, as always, the good quality of the data is an absolute necessity for successful customer profiling. Albeit there are differences in the robustness of the methods, the low-grade data usually leads to bad results. Thus, the preprocessing of the raw data into a clean and proper data set takes a major portion of the whole mining process.

Often, the amount of data corresponding to the normal or average cases is considerably large compared to the few anomalies present. These special cases might be outliers or, on the other hand, they might be just the interesting cases that are needed to be detected. Thus, the preprocessing and selection of the mining method should depend on what is desired to be discovered.

Large data sets, containing hundreds of attributes and millions of samples, proved to be a little problematic overall. Their processing involved both the extensive computational load and memory usage, which often became even intolerable. Again, different methods have different performance characteristics, but often some kind of adjustments of the data or the methods is required in order to complete the desired profiling task.

On the other hand, the statistical nature of the used methods requires sufficient sample sizes to produce a reliable outcome. This is especially essential when the data includes categorical attributes with rather high cardinality. In the worst case, the target variable comprises of a large number of categories. Thus, it should be assured that there are enough samples in each category, in the both training and test data.

Finally, the interpretation and validation of the mining results forms the essential part of the customer profiling process. This clearly requires expertise of both the used methods and the application area. By validating, problems in the mining process may be detected and the correctness and usefulness of the results can be confirmed.

References

[1] Jussi Ahola. *Mining Sequential Patterns*. Research report TTE1-2001-10, VTT Information Technology, May 2001.

[2] Catherine Bounsaythip and Esa Rinta-Runsala. *Overview of Data Mining for Customer Behavior Modeling*. Research report TTE1-2001-18, VTT Information Technology, June 2001.

[3] Malesh Joshi, George Karypis, and Vipin Kumar. *A Universal Formulation of Sequential Patterns*. Technical Report No. 99-021, Department of Computer Science, University of Minnesota, 1999.

[4] Samuel Kaski. *Data Exploration Using Self-Organizing Maps*. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, 1997.

[5] Subhash Sharma. *Applied Multivariate Techniques*. John Wiley & Sons, Inc., 1996

[6] *AnswerTree 2.0 User's Guide*. SPSS Inc., 1998.

[7] *SPSS Regression Models 9.0.* SPSS Inc., 1999.

[8] Breiman, L; Friedman, J.H.; Olsen, R.A. & Stone, C.J. *Classification and Regression Trees,* Belmont, CA, Wadsworth, 1984.

Appendix 1

Decision tree of the banking case study.