USING THE XLMINER TOOL FOR DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT

I. Polaka and A. Sukov

Keywords: data mining software, classification, clustering, associative rules, XLMiner

1. Introduction

Data mining is a set of methods that helps an expert to obtain new, previously unknown significant information from available data. To apply data mining possibilities practically software that deploys data mining algorithms is needed. This software may also be integrated into other tools. One of these tools is the QuantLink XLMiner tool for Microsoft Excel. In this paper the data mining possibilities of XLMiner are examined. More precisely – this tool is used for extracting unknown client data, rules and patterns, as well as for extracting information that is significant for client relationship management.

2. The XLMiner software

XLMiner is Microsoft Excel plug-in software [1], a product of QuantLink. It works with Microsoft Excel (English versions 2000, 2003 or 2007) environment. This tool offers solving many data mining tasks using the most popular methods [2]. Available tasks include classification, clustering, time series analysis and associative rules. XLMiner also offers a possibility of sampling data from external data bases or MS Excel spreadsheets, as well as partitioning data set into training, validation and test data sets and performing data pre-processing. The tool also includes some visualization options and allows using a pre-created model on new data.

The tool works with the MS Excel environment but the work with the supported methods is carried out using a standard three step wizard that guides the user through the stages of input data information, method parameters and output data information.

Figure 1. Interface of XLMiner – a standard wizard
Results and model data are shown in specially designed *MS Excel* worksheets that make further analysis faster and easier. The tool also outlines the ranges of spreadsheets that will be needed for further work with this tool and this data (e.g. applying other methods).

Most of the computer users are already familiar with the *MS Excel* interface therefore learning to work with *XLMiner* is quick and easy, it doesn’t require an extended course on one program if users have some knowledge about data mining and the methods that are used in this tool.

### 3. Client information analysis

In the recent years businesses have moved from production oriented strategies to those that put customer satisfaction as their goal. Product and service sales are now more oriented towards clients’ needs. Therefore it is necessary to know your customer. It is made possible by analyzing available customer data – transactions and other committed information like filled in questionnaires etc. This information allows us to forecast further demand, personalize advertisement campaigns, cross-sell new or related products or services etc. In the experiments conducted, the data was provided by Southern-German Class lottery [3]; it holds personal information about the client and his living environment, and also the data that describes playing habits of that person. The goals of the experiments were to forecast for how long a person will be playing in this lottery (extract patterns that could help to point out potential loyal customers), to determine the age of the most active customer group and also to extract other significant rules that would help with the customer relationship management.

The data set used in the case study contains 72 attributes and more than 100 000 records. This data set was used for the classification task to determine the activity class to which a person would belong based on the information which was known previously about this customer. The results of classification will help to forecast the client flow. This data set was also used for clustering to determine the dependencies between the age of a person and the number of tickets bought by this person. This information may prove helpful in further market segmentation. To extract more relevant patterns that exist among the features of customers, the data was mined for associative rules.

#### 3.1. Forecasting client activity class

*XLMiner* offers to perform data sampling and partition using a data set that holds up to 60 000 records and 200 attributes. This includes data sampling to reduce the size of a data set (numbers of records) by choosing a subset of records that would represent the whole set. It can be done by random partitioning or partitioning with oversampling setting a desired support limit for a class.

Data set is larger than the limitations set by *MS Excel* or *XLMiner*, for this reason random partitioning was used to reduce the number of records. But the resulting data set would still have more attributes than it is allowed in this tool (30 attributes are allowed). Therefore a smaller subset of the attributes is needed, that doesn’t lose much information in the reduction process. Therefore a tool named *WEKA* [4] was used to pick a subset of 13 best attributes.

The data set was also divided into training and test data sets. Usually data sets are divided into training set holding 70% of the data and test set holding 30% of the data [5]. This ratio was also used in this case study wherever such division was necessary (e.g. classification and forecasting).

*XLMiner* offers solving a classification task using the following methods: discriminant analysis, logistic regression (not to be confused with logic regression, which works with
binary data and classifies data using the Boolean algebra approach), classification trees (CART), naive Bayes classifier, neural networks (multilayer feedforward architectures) and k-nearest neighbours. Discriminant analysis does not work with the needed amount of data and logistic regression does not support 5 classes in the data in this tool (it only classifies 2 classes), therefore this task was solved using other methods and attributes of data types that are supported by these methods (e.g. continuous data was not used with naive Bayes classifier). The best method for this task was chosen using the total classification error and the wrong classifications into most important class (that is class 4) (see Table 1).

Each client was assigned an activity class which shows how long a client will participate in the lottery, which means we can forecast whether this client will be a loyal customer and for how long he will participate). We consider the following classes that show 5 common patterns of client behaviour:
- “0” no tickets were bought,
- “1” only the first ticket was bought,
- “2” no more than two tickets were bought,
- “3” client bought all tickets for one season,
- “4” client played during the first season and bought tickets for the second season (we consider this type of people ‘loyal customers’).

The importance of this task is its contribution to forecasting client behaviour for the company, it helps to forecast the number of players, money flow etc. If there is information about prospective customers (that are not clients of this lottery yet), it can also provide information for selective advertising campaigns, distributing the advertisements to prospective customers that will most likely become loyal customers.

An analysis of the classifiers’ results enables us to evaluate the performance of the classifiers. It was done by evaluating the total classification error and the rates of the wrong classifications into each class (Table 1).

<table>
<thead>
<tr>
<th>Class</th>
<th>Classification tree</th>
<th>Naïve Bayes</th>
<th>$k$-nearest neighbours</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>38.85</td>
<td>39.57</td>
<td>42.34</td>
<td>23.64</td>
</tr>
<tr>
<td>1</td>
<td>100.00</td>
<td>100.00</td>
<td>99.33</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>100.00</td>
<td>99.25</td>
<td>97.26</td>
<td>100.00</td>
</tr>
<tr>
<td>3</td>
<td>100.00</td>
<td>48.48</td>
<td>78.89</td>
<td>100.00</td>
</tr>
<tr>
<td>4</td>
<td>13.67</td>
<td>33.48</td>
<td>26.89</td>
<td>29.07</td>
</tr>
<tr>
<td>Total</td>
<td>45.44</td>
<td>47.75</td>
<td>49.24</td>
<td>49.24</td>
</tr>
</tbody>
</table>

Some classifiers did not classify classes 1 to 3 but they were not the most important for this task. The most important was to determine the clients that will play in this lottery regularly (4th class). The classifier that had the lowest error for this class was classification tree. The same method also had the lowest total classification error. This means that the most appropriate method to solve this task is to build a classification tree. The use of this method gives us 86 right answers out of 100 when we want to know whether the client will become a loyal customer.

In this case study there were several inconveniences caused by data set limitations posed by software. Not only was it the record number limitation by Microsoft Excel, but also some unexpected limitations from the XLMiner tool. The work with 60 000 records (XLMiner limit) could not be completed even if the data set had less than 200 attributes that were
allowed. The set that did fit all the limitations had 19,000 records while pre-processing the data and 10,000 records in the training set for classification.

The results of the classification were overall worse than previously assumed – the error rate of the best classifier was above 40% and that is big enough for the results to lose their impact on business decision making. But the classification error of the most important class (class 4 which means that the client will be a loyal customer and will take part in the lottery for a long time) was better – it was 14% for the classification tree. This means that it can be used for further plans and business strategies.

3.2. Segmentation of participants age and number of tickets bought

To determine the age at which people tend to play more, client data was clusterized using attributes Age and Number of tickets. Relations extracted here will help to determine the best segment of population that will be the right recipient (i.e. the one who will participate). At the beginning of the work, a graph was plotted to show the data relations in a two dimensional data chart, to visualize the data and see the possible patterns (see Figure 2).

![Figure 2. Data graph showing client’s age and number of tickets bought](image)

It is obvious that there is a group of records that stand out in the graph – they are all with the age value of 0. This means that these are the records that did not have the age specified. These records were excluded from the data set because they didn’t hold any significant information that would contribute to the solution of this task greatly.

In the hierarchical clustering clusters are built from records or subclusters until there is one cluster left that holds all of the records [6]. After the work is finished, the best number of subclusters is determined (usually using a dendrogram which shows merging of subclusters and distances at which it happens), that means the number of clusters is chosen where there is the longest distance among clusters. In this case there were three clusters. Data division was based on the number of tickets played in this case, attribute age did not contribute much (one group played 5-15 tickets, another group played 20-25 tickets and another group played 30-35 tickets).

Partitional clustering means assigning each record to the cluster with the closest centroid (assigning records to clusters of previously specified number) [6]. Trying different
numbers of classes and coordinates of the centroids, the most typical and informative result was with 4 clusters. This method also showed different end clusters if compared to the hierarchical clustering. In this clustering, cluster centers, numbers of records in clusters and average distances in clusters were as shown in Table 2.

Table 2

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Age</th>
<th>Number of tickets</th>
<th>Records</th>
<th>Average distance in cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster-1</td>
<td>70</td>
<td>6</td>
<td>1335</td>
<td>8.784833</td>
</tr>
<tr>
<td>Cluster-2</td>
<td>56</td>
<td>1</td>
<td>1386</td>
<td>6.865723</td>
</tr>
<tr>
<td>Cluster-3</td>
<td>50</td>
<td>20</td>
<td>432</td>
<td>11.288745</td>
</tr>
<tr>
<td>Cluster-4</td>
<td>28</td>
<td>1</td>
<td>817</td>
<td>7.640936</td>
</tr>
</tbody>
</table>

We can see that people of age 20-40 (Cluster-4) do not play often and end soon (1 to 8 tickets), and the percentage of this type of participants is only 20. Participants that have reached the age of 40 (up to 60 year olds) can be divided into two groups – those who buy 15-45 tickets (and more (around 11%)) and those who buy relatively few (4-8) tickets (~35% of the participants).

Another group is people who have reached the age of 70 – they don’t buy as many tickets as group 40-60 does but the still buy more tickets than the young participants. So we can see that older people buy more tickets, and most of the tickets are bought by the participants who are 40 to 60 years old. So people of this age will most likely answer the advertisements and will be willing to try demo or introduction packages.

Also in the clustering tasks there were some unexpected limitations concerning the data set. The maximum number of records in this task is 4000. And even if the data can be clusterized, the visualization could be better – the results could be better interpretable and easier to understand, if there were more visualization tools for clusterization. Even the dendrogram shows only up to 30 subclusters.

3.3. Discovering other significant patterns

If we extract new interesting relationships about people’s lifestyle and environment, it can be used for more efficient advertisement campaigns. Therefore data was also mined for association rules. In this task frequent patterns in the data can be found. They can be expressed as relations between objects of the data set in the form ‘If A, then B’ where A and B can be sets of objects [5].

Data set was reduced because the XLMiner tool brought up an associative tree building error for the maximum supported data set. Attribute set was reduced excluding some attributes that describe others (e.g. the evaluated frequency of each car (by producers) in the client’s neighbourhood). The rest of the attributes were transformed to have unique values compared to each other so that they can be distinguished. In this case transformation was carried out with MS Excel function, adding a unique position before the values for every attribute.

Some interesting associative rules were found in the process. For example, if a person chose mail as a primary mean of communication, complaints arrived only in 2% of the cases. Also the participants who live in the old Federal States, participants who live in houses with no offices, and those who are driving Volkswagen or Seat, preferred to choose mail as a primary mean of communication.
And there was also discovered a sure rule (559 people of 559 matched) – if the participant’s purchasing power was evaluated 10 (maximum points), this person lived in the old Federal States. It is not surprising because economic growth of the East Germany was slower in the Soviet Union years compared to the growth of West Germany and this means that the company would benefit more from more expensive projects in the western part of the country.

4. Other options of the software

**XLMiner** has the ability to solve forecasting tasks using multiple linear regression, k-nearest neighbours, regression trees or neural networks (multilayer feedforward architectures). It can also analyze time series – autocorrelation function and partial autocorrelation function for data exploration and finding patterns in data, and autoregression and moving average integrated (ARIMA) for forecasting. **XLMiner** also supports smoothing functions – exponential, double exponential, moving average and Holt-Winters smoothing.

**QuantLink** also offers a tool named **XLMinCalc**, which allows using models made with **XLMiner** on other data without a need of **XLMiner** on this machine.

5. Results

This paper discusses the possibilities of **XLMiner** in solving data mining tasks, concentrating on tasks that are relevant for client relationship management. **XLMiner** can solve classification, forecasting, clustering, time series analysis and associative rules tasks. And the classification task is the one that is most emphasized. This task can be solved with various methods that work with both categorical (naive Bayes classifier) and continuous data (k-nearest neighbors and others) and the results are comprehensive and more processed. For time series analysis and associative rules there are some of the most popular methods but these tasks are not the main application of this tool, although they can be used successfully as secondary analysis tools. The same is with clustering – some of the methods are offered for this task but result analysis is quite poor and so is visualization.

This tool also lacks some data pre-processing utilities. In this case study some tools for attribute set analysis and reduction (subset selection) would have been handy but they are offered only for some regression methods (as built in functions).

All of the data set was not used in any task because of the **XLMiner** limitations. So we have to conclude that this tool is more suitable for tasks with smaller data sets and as an illustration in the learning process for the tasks described previously.

Acknowledgements

Thanks to Dr.sc.ing., Assoc.prof. Ludmila Aleksejeva for help and support.

References


Inese Polaka, B.sc.ing., Master’s student, Department of Modelling and Simulation, Riga Technical University, 1 Kalku Street, Riga, LV-1658, Latvia, e-mail: Inese.Polaka@gmail.com.

Anatoly Sukov, M.Sc.ing., Lecturer, Department of Modelling and Simulation, Riga Technical University, 1 Kalku Street, Riga, LV-1658, Latvia, e-mail: Anatolijs.Sukovs@cs.rtu.lv.

Polaka Inese, Sukovs Anatolijs. Rika XLMiner izmantošana datu ieguves klientu attiecību vadībā
Rakstā tika apskatītas dažādā datu ieguve iespējas, kas ir svarīgas klientu attiecību vadībā. Datu ieguve tika veikta loterijas klientu datos, izmantojot Microsoft Excel programmu un QuantLink XLMiner pievienojumprogrammu. XLMiner atbalsta ūdu datu ieguves uzdevumu risināšanu kā klasifikācijai, klastēšanai, laika rindu analīze un asociatīvo likumu meklēšanu, lielāku uzvaru liekot uz klasifikācijas uzdevumu, piedāvājot lielāku metožu klāstu un plašāku rezultātu analīzi. Šā rīka piedāvātās iespējas tika izmantotas, lai no esošiem loterijas klientu datiem izgūtu klientu attiecību vadības procesiem svarīgu informāciju. Tika veikta klienta aktivitātes klasēs noteikšana ar dažādu piedāvāto klasifikatoru palīdzību (klasifikācijas koki, naivais Baijesa klasifikators, k-tuvākie kaimiņi un neironu tīkli), nosakot ari labāko klasifikatoru. Tika veikta ari klientu segmentācija pēc vecuma un nopirktu bilēju skaita, izmantojot hierarhisko un sadaļošo klastēšanu, nosakot loterijas mērķauditoriju, t.i. visvairāk spēlējošo vecumu grupu. Tika veikta ari cita zīmīgu likumu meklēšana ( asociatīvie likumi).

Polaka Inese, Sukov Anatoly. Using the XLMiner tool for data mining in customer relationship management
This paper describes some of the data mining tasks that are important for customer relationship management. Data used in mining was lottery client data and it was carried out with Microsoft Excel software and QuantLink XLMiner plug-in software. XLMiner offers to solve data mining tasks like classification, clustering, time series analysis and associative rules, emphasizing the classification task by offering more methods and broader results analysis. The possibilities offered by this tool were put to use to extract data that is significant for customer relationship management from the available data. The following tasks were solved: assigning an activity class to each client, using some of the offered classification methods (classification trees, naive Bayes classifier, k-nearest neighbours and neural networks), client segmentation by age of customers and number of lottery tickets bought, using both the hierarchical and the partitional clustering to determine the age group that contains people who are most likely to play more. The data was also mined for other significant rules (association rules).

Полика Инес, Суков Анатолий. Использование инструмента XLMiner для добычи знаний в задачах управления взаимоотношениями с клиентами
В статье рассмотрены различные возможности добычи знаний, которые важны для управления взаимоотношениями с клиентами. Добыча знаний была произведена на данных клиентов лотереи, используя среды Microsoft Excel и специальную программу QuantLink XLMiner. XLMiner поддерживает решение таких задач добычи, как классификация, кластеризация, анализ временных рядов и поиск ассоциативных законов, уделяя особое внимание задаче классификации, предлагая более широкий выбор методов и более полный анализ результатов. Возможности этого инструмента были использованы для того, чтобы из имеющихся данных о клиентах лотереи получить важную информацию для процессов управления взаимоотношениями с клиентами. Проведено определение класса активности клиента с помощью различных предложенных классификаторов (деревья классификации, наивный классификатор Байеса, k-ближайшие соседи, нейронные сети), выбран наилучший классификатор. Проведена сегментация клиентов по возрасту и количеству купленных билетов, используя иерархическую и разделяющую кластеризацию, определена целевая аудитория лотереи, то есть наиболее ищущая возрастная группа игроков. Также произведён поиск других показательных законов (ассоциативных законов).