H.1.1 Database Marketing and Web Mining

Abstract

The four customer-related key disciplines in marketing are attraction, retention, cross-sales, and departure. The same holds for database marketing and its electronic commerce equivalent in web mining. The case study that is presented tackles the problem of cross-sales in the financial sector in which a particular service had to be cross-sold to the existing customer base. The techniques which were applied are *characteristic rule discovery* [link to Section C5.2.2] and *deviation detection* [link to Section C5.3.1]. We discuss the effect of domain knowledge on the interestingness value of the discovered rules and study techniques for refining the knowledge to increase this interestingness measure. We also investigate the use of externally procured lifestyle and other survey data for data enrichment and discuss its use as additional domain knowledge. The same scenario is then mapped onto its electronic commerce counterpart, where we used log files to discover navigational behavior in order to model potential cross-sellers.

Keywords: Cross-sales, web mining, discovery of sequences, characteristic rule discovery, interestingness measures

H.1.1.1 Project Overview

The project covered all the typical phases of a data mining undertaking, viz. from data warehousing issues, via domain knowledge incorporation, pattern discovery, evaluation, visualization to the application of results.

The 6 month project with an effort of 15 person months involved three different expertise. The domain knowledge was provided by a financial marketing adviser from the bank, the data expert was represented by the IT department of the financial institution, and the data mining expertise was provided by the Northern Ireland Knowledge Engineering Laboratory.

Although the data mining process we followed was borrowed from Anand and Büchner (1998), it is very similar to the one described in Part C of this handbook. As a result of this, our process has been

mapped, where appropriate, onto the one used in this handbook. The following sub-section describes the work undertaken by us within the various stages of the process.

H.1.1.2 KDD Process

H.1.1.2.1 Business Problem

The business problem we were confronted with was that of cross-selling household insurances to existing customers in a banking database. The problem is depicted in Figure H.1.1.1 below. The overall objective was to discover characteristics of current household insurance customers, which could then be used to target all other customer segments, in order to classify them into potential promotion targets and unlikely purchasers (Anand et.al. 1997).

Bank Customer Base   **Bank Services**

Customers who have bought Houshold Insurance

Customers who have bought Mortage Protection

Customers who have bought Indemnity Bonds

Customers who have bought Commercial Contents Insurance

Commercial Contents

Indemnity Bonds

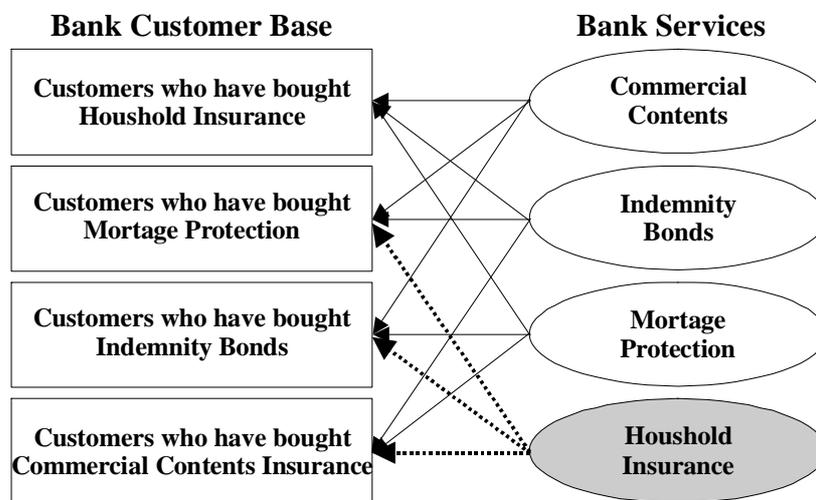Mortage Protection

Houshold Insurance

Figure H.1.1.1 The Cross-Sales Problem

H.1.1.2.2 Motivation for a Data Mining Solution

According to Anand et. al. (1998), there are four component tasks that can be identified for the cross-sales problem. They are to

- find the sets of customer characteristics that identify, in the customer base, those customers that are most likely to buy a particular product (in this case Household Insurance);

- choose the best of these sets of characteristics, to identify customers to target in a marketing campaign of some sort (for example a Mail shot);

- carry out this marketing campaign and analyze the results to see if a high 'hit rate' was achieved;

- feed back results into the customer database, to carry out refinement of the rules used for targeting customers with the product.

Of these four tasks, tasks one and two can be identified as data mining tasks, where we were trying to discover the best sets of attributes within the banks database that identify customers of household insurance. The latter two tasks are of pure marketing nature and will not be discussed any further.

The categorization of the first task as having either a classification or characteristic rule discovery goal depends on the data available within their customer database. For a classification goal the bank must have data on three types of customers, namely those that have a household insurance, those that refused to purchase a household insurance, and those that did not have a household insurance, but did not refuse to purchase it. The first type of customers form the positive examples, the second type of customers form the negative examples and the third type form the target data set. If, however, the bank only has data on two types of customers, namely those that have a household insurance and those who do not, i.e. positive examples and a target data set, rather than a classification goal we have a *characteristic rules* discovery goal [link to Section C5.2.2]. These rules define those characteristics that are prevalent in a particular group of records which in our case is the group of records that pertain to customers who have a household insurance. Given these rules, customers in the target data set with similar characteristics can be targeted with sales campaigns.

The second task is associated with filtering the discovered characteristics using some kind of domain specific mechanism so as to choose the most relevant characteristics to use for targeting customers. The technique used would clearly depend on the type of goal associated with the first task, which in turn depends on the availability of data. Thus, we postpone further discussion on the second task until Section H.1.1.2.7.

H.1.1.2.3   The Available Data

With respect to the data held by the bank, two types of the customer data were identified as being relevant to the cross-sales problem. These were the personal information about the customer, for example, demographic information, sex, occupation and marital status and transactional information on the different accounts held by the customers. An important aspect of the data was identified at this stage. While data on customers who bought household insurance was available from the Bank's databases, no information was available on customers that did not require a household insurance product at all, or had not taken up the household insurance product with the bank but had household insurance with a competitor organization. This fact confirms that the classification of the first data mining task identified earlier had a characteristic rule discovery goal as opposed to a classification goal.

Three externally available data sets were believed to be relevant to the cross-sales problem, as they provided information about the banks customers not stored in the banks database. The Robson's Deprivation Index for Northern Ireland provided information about the level of deprivation of the area in which the customer lived while the Acorn classification data and lifestyles survey data provided data such as average income, employment rates, average family size of the area in which the customer lived. Each of these data sets provided information summarized to the enumeration district (artificial geographical boundaries defined by the government based on population density). The bank identifies the geographical location using postal code and not enumeration district.

H.1.1.2.4   Background Knowledge

We formulated various types of domain knowledge in collaboration with the domain expert, all of which were either classified as *taxonomies* [link to Section C7.1] (in form of concept hierarchies), environment-based *constraints* [link to Section C7.2], and *user preferences* [link to Section C7.4] (represented as syntactic constraints and inter-attribute dependency constraints). At a later stage some of this domain knowledge was refined using *previously discovered knowledge* [link to Section C7.3].

H.1.1.2.5   Preprocessing and Data Extraction

In order to create a knowledge discovery view which encompassed relevant customer information, accumulated transactional data, as well as connected external sources, various data preprocessing and extracting steps had to be performed. The preprocessing mainly constituted the removal of outliers, conversion of continuous attributes to discrete attributes and heterogeneity resolution (achieved through data aggregation and spatial joins) among the three different data sources. Data on 60,000 customers was used in the data mining project, 430 of whom were existing household insurance customers.

H.1.1.2.6   The Discovery Mechanism

An association rule discovery algorithm based on evidence theory, the EAR (evidence-based association rule) algorithm was used to mine the clients data. The EAR algorithm is a generalization of earlier association algorithms and therefore allows the incorporation of support and uncertainty thresholds and syntactic constraints. In addition to the simple syntactic constraints of the type defined by (Agrawal and Srikant 1994), EAR allows the definition and consequently the incorporation of inter-attribute dependency constraints. For example, a rule that contains an expression pertaining to the account average balance of a customer is only valid if it also contains an expression regarding the account type as well. In addition to this, the EAR algorithm can discover knowledge from multi-valued attributes rather than just binary attributes as in the case of previous algorithms. It allows the incorporation of domain knowledge and can handle missing values in the data. The EAR algorithm requires attributes to be discrete. Therefore, interval bands were provided by the domain expert for continuous variables. Also domain specific hierarchies were provided by the domain expert for a number of other attributes.

Through knowledge discovered at various intermediate stages, these taxonomies were refined to achieve more interesting knowledge.

H.1.1.2.7   The Results

In addition to domain knowledge incorporation, the number of rules generated was regulated by setting a threshold on the support of the rule in the data set and by defining an interestingness measure. The interestingness measure used is normally dependent on the problem at hand. In cross-sales, we clearly do

not want to base the targeting of a product customer characteristics that are actually characteristics of the banks customers in general. Thus, the interestingness measure we used was based on the deviation of the characteristic rules discovered for the customers of the product being targeted from the norm. The norm in our case was defined as the support for these customer characteristics within the complete customer base of the bank, that is a characteristic rule is interesting if it is a characteristic of the customer of a product rather than the customer base in general. Thus, we defined the interestingness measure for customer characteristics, c, as:

$$Interest_c = \frac{S_p - S_o}{\max\{S_o, S_p\}}$$

(H.1.1.1)

where, S_p is the support for the characteristics c in the positive example data set and S_o is the support for the characteristics c in the complete customer base. The expression in the denominator is called the normalizing factor as it normalizes the interestingness measure onto the scale [-1, 1].

Example characteristic rules discovered are shown below:

> *if Household Insurance = Y*
> *then Occupation = SKILLED*
> *with support = 26.51% and interest 0.53*
>
> *if Household Insurance = Y*
> *then Occupation = SKILLED and Status = Hon-Commits*
> *with support = 21.86% and interest 0.68*
>
> *if Household Insurance = Y*
> *then Occupation = SKILLED and Status = Hon-Commits and Net Credit Turnover > 4000*
> *with support = 12.79% and interest 0.74*
>
> *if Household Insurance = Y*
> *then Occupation = SKILLED and Status = Hon-Commits and Net Credit Turnover > 4000*
> *and Account Type1 = CURRENT*
> *with support = 12.56% and interest 0.74*
>
> *if Household Insurance = Y*
> *then cus_nodeps = 0_Dep and cus_yrnetavgbal = Zero_1500 and CHILDREN = 4*
> *with support = 10.93% and interest = -0.77*

A positive interest measure suggests that customers with the given consequent are likely to buy household insurance, whereas a negative interest indicates that customers with the given characteristics are less likely to purchase the product.

The effect of the Interest value threshold on the number of rules discovered is shown in Table H.1.1.1. The rules were constrained to a maximum size of seven consequent attributes.

Interest Threshold	Number of Rules
0.84	7
0.80	217
0.70	1178
0.50	1739
0.00	3737

Table H.1.1.1 Number of rules discovered versus the interest measure

Figure H.1.1.2 presents some of the rules graphically. The oval nodes represent attributes used to specialize the rule specified by the path from the root node to the node preceding the oval node, while the rectangular nodes represent the specialization attribute value. The numbers shown in the rectangular node are the support and interest of the rule. The light gray nodes represent rules that have an interest value less than or equal to another rule that the present rule is a specialization of, while the darker gray nodes represent rules where the specialization attribute has improved the interest value of the rule. Non-shaded nodes represent rules where the specialization has decreased the interest in the rule.
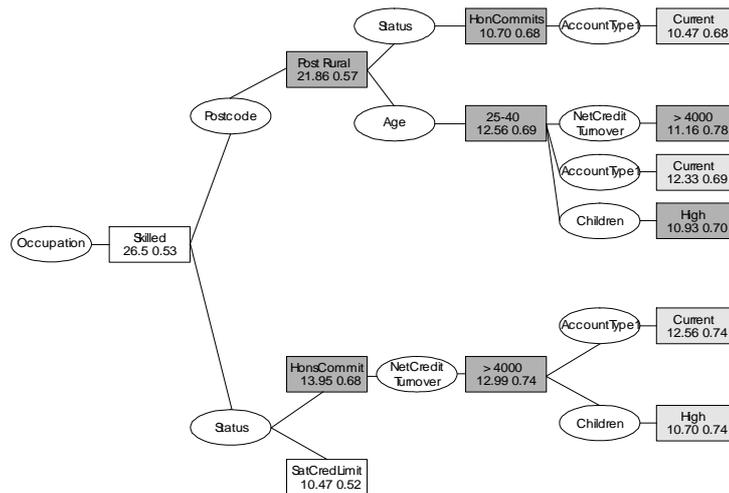

Figure H.1.1.2 Characteristic Rule Visualization

H.1.1.2.8   Applications of the Discovered Knowledge

Before the discovered knowledge was ready for the application in the business context of the bank, two knowledge post-processing steps were required. The concept hierarchies were refined to improve the interestingness of the resulting rules while the attribute relationship rules and constraints were also tuned to improve the quality of the rules discovered. Based on the refined domain knowledge the same process, as described above, was performed and the results were then used in order to classify the customers into potential promotion targets and unlikely purchasers.

In order to apply the discovered results, we ranked the potential customers according to their associated interest values. From all the rules that applied to the customer, the interest value associated with the customer was that of the rule with the lowest interest measure. The higher the (positive) interest value associated with the customer, the more likely the customer is to buy a household insurance.

H.1.1.3   Discovering Internet Marketing Intelligence

From a high-level marketing point of view, the four stages in a marketing life-cycle (attraction, retention, cross-sales, and departure) are identical in traditional retailing and electronic commerce scenarios. However, on a more detailed examination, there are some subtle differences which require additional efforts in order to apply marketing mechanisms on Internet data (Mulvenna, Norwood and Büchner 1998).

The main differences are anchored in the data that is available for knowledge discovery, as well as the type of domain knowledge that can be incorporated. In addition to traditional customer and transaction / purchasing data we can also collect more sophisticated behavioral data in form of browsing information, which is stored in web log files. Additionally, a different type of domain knowledge is available, which represents the topology (structure) of an online trading site. Considering the supplementary elements of electronic commerce business models, we are now in a position to briefly describe a cross-sales scenario based in web data.

The project has been carried out with one of the biggest Irish online book shops, where currently about 2% of the overall sales are from Internet users. The objective of the feasibility study was to establish the usability of existing customer, transactional and browsing data, in order to discover Internet marketing

intelligence. The two most interesting types of marketing knowledge were that of attracting new customers and, naturally, that of cross-sales.

The data preparation proved to be the most time-consuming task, since it involved not only the collection of customer information, transactional data, as well as common and cookie log files, but also the organization in a data warehouse in order to build a web log data cube (Büchner and Mulvenna 1998). Further activities, which were distinct to traditional retail data, were the performing of reverse DNS lookups on unresolved URIs and the handling of relatively large numbers of unknown records, often caused by broken links.

In addition to the traditional cross-sales mechanisms outlined in sub-sections H.1.1.1 and H.1.1.2, a more Internet-specific, and thus less generic, approach was performed for rule discovery. The types of domain knowledge that was considered was the above mentioned network-like electronic shop topology as well as user-defined hierarchies (generalizations of book topics). The approach we chose was to discover typical sequences that occur before existing customers purchase another product. The unique concept of storing browsing patterns allowed that type of discovery, where navigational behavior can be used as a key factor. In order to discover sequential patterns [link to Section C5.2.4] from web log files, Agrawal's family of sequential apriori algorithms has been extended, which are also known as AprioriAll (Agrawal and Srikant 1995). The main extensions are a more efficient implementation for large amounts of data, as well as the support for multiple values in discovered sequences. Both extensions were relevant for the discovery of behavioural navigational patterns for electronic commerce purposes in Internet data, because the amounts of data to be dealt with are usually rather large, and forward and backward browsing on electronic shops as well as double hits have proven important.

The results we found contained a number of sequences pertaining to customers who had bought at least one item in the bookshop within the previous 3 months and their most typical behavior before they purchased another item. These rules were than mapped onto the most recent log files in order to rank the most likely customers for cross-selling.

The application of the discovered knowledge is similar to traditional retailing, with an essential difference. The targeting can be performed online and for each individual customers, which ideally leads to

a one-to-one marketing scenario. For instance, a customer who has all the attributes for a potential cross-selling activity, can be displayed a special tailored offer dynamically. Furthermore, the reaction on this offer can be monitored and fed back in the KDD process.

References

Agrawal, R. and R. Srikant. 1994. "Fast Algorithms for Mining Association Rules in Large Databases", *Proc. 20th Int'l Conf. on Very Large Databases*, pp. 487-499.

Agrawal, R. and R. Srikant. 1995. "Mining Sequential Patterns" *Proc. Int'l. Conf. on Data Engineering*, pp. 3-14.

Anand, S. S. and A. G. Büchner. 1998. "Decision Support using Data Mining", London: Financial Times Pitman Publishers.

Anand, S. S., A. R. Patrick, J. G. Hughes and D. A. Bell. 1998. "A Data Mining Methodology for Cross-Sales", *Knowledge-based Systems Journal* **10**: 449-461.

Anand, S. S., J. G. Hughes, D. A. Bell and A. R. Patrick. 1997. "Tackling the Cross-Sales problem using Data Mining", *Proc. 1st Pacific-Asia Conference in Knowledge Discovery and Data Mining*, pp. 25-35.

Büchner, A. G. and M. D. Mulvenna. 1998. "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", *ACM SIGMOD Record* **27**(4).

Mulvenna, M. D., M. T. Norwood and A. G. Büchner. 1998. "Data-driven Marketing", *Int'l Journal of Electronic Markets* **8**(3): 32-35.