A Data Mining Application: Customer Retention at the Port of Singapore Authority (PSA)

KianSing Ng
Dept of Info Sys and Comp Sci
National University of Singapore
Lower Kent Ridge Rd, Singapore 119260
ngkians1@iscs.nus.edu.sg

Huan Liu
Dept of Info Sys and Comp Sci
National University of Singapore
Lower Kent Ridge Rd, Singapore 119260
liuh@iscs.nus.edu.sg

HweeBong Kwah
Port of Singapore Authority
#04-01 Tanjong Pagar Cpx
7 Keppel Rd, Singapore 089053
hbkwah@hq.psa.com.sg

1. ABSTRACT

"Customer retention" is an important real-world problem in many sales and services related industries today. This work illustrates how we can integrate the various techniques of data-mining, such as decision-tree induction, deviation analysis and multiple concept-level association rules to form an intuitive and novel approach to gauging customer's loyalty and predicting their likelihood of defection. Immediate action taken against these "early-warnings" is often the key to the eventual retention or loss of the customers involved.

1.1 Keywords

Decision-tree induction, deviation analysis, multiple concept-level association rules, customer retention

2. INTRODUCTION

"***Customer Retention***" is an increasingly pressing issue in today's ever-competitive commercial arena. This is especially relevant and important for all sales (e.g. departmental stores, banking, insurance) and services (e.g. Internet / telecommunication service providers) related industries. From an economic point of view, customer retention directly translates to a huge saving in marketing cost, as highlighted by *Coopers & Lybrand's Vince Bowey*:

> "A lot of companies haven't figured out what it costs them to acquire a new customer, it's usually pretty shocking. We estimate that it costs three to five times more money to acquire a new customer than to keep the ones you have."

The problem is especially relevant in the context of PSA – the managing authority of the world's busiest port. With a customer base of only a few hundreds shipping-lines, contributing to an annual turnover of a few hundreds of millions, PSA has a strong interest in retaining each and every one of them. Moreover, the Singapore port is facing keen competition from many upcoming ports in the neighboring regions. This means that the Marketing division faces higher risk of customer defection and potentially escalating marketing costs. Furthermore, the majority of the port activities in Singapore are *entrepot* in nature, i.e. an intermediary center for goods exchange and transshipment. This means that the defection of a customer to another port is likely to influence its associated business partners to also defect, in order to maintain their current transshipment business. Thus, there is a pressing need for an effective method to identify customers' associations and partnerships for the benefit of marketing and customer retention. Finally, contrary to common belief, focusing on customer retention is not a passive policy because existing customers bring in new customers through their business associations and through word of mouth.

The conventional approaches to dealing with this type of problem have a number of shortcomings. Domain experts are required to subjectively determine a set of "*performance indicators*", from which each of the customer's current performance is compared with the previous month's and year's performance. Since each expert has his or her own different list of performance indicators which are tacit, there is no way to tell which is more accurate. This is aggravated by the fact that indicators tend to change over time in the real world due to ***concept drift***. Therefore, there is a risk of monitoring the irrelevant and outdated "*performance indicators*".

Managers and executives rarely realize that the very knowledge that can help them alleviate these problems lies within the wealth of data already at their disposal. In the context of our application, PSA logs in approximately 2,000 vessel-calls and 45,000 TEUs (Twenty Thousand Equivalent Units) of container-transactions into various databases daily. This wealth of information can potentially be mined for the knowledge to gauge the loyalty of her

customers and predict the possible "chain-effects" of a defection. Such knowledge would enable PSA to take steps to retain her valuable customers before it is too late. This work illustrates how the integration of the various techniques of ***data mining*** can give rise to an intuitive and novel approach that provides such knowledge. The application is implemented using C and Gentia[1] on PCs in a client-server network environment.

It should be highlighted that due to the confidentiality of the details involved, all the data attributes are substituted with notations, e.g. S_1, S_2, ...,S_N to represent the different shipping-lines. If possible, we try to cite examples from the context of the port business itself. Otherwise, an analogy from the domain of database marketing is used to illustrate the concept.

3. DECISION TREE INDUCTION FOR CLASSIFICATION

Our first step uses ***C4.5*** [13] to apply ***decision tree induction*** [12] to find "*objective indicators*" that describe the target concept of "customer's loyalty". The training dataset is prepared from the vessel's transactional database residing in Oracle database, which consists of over 60,000 monthly records of more than 30 attributes.

Using an analogy from database marketing, a typical ***classification rule*** obtained from apply decision-tree induction to the training dataset in Table 1 is:

if (**income** > $40K) ∩ (38>**age**>29) ∩ (**education** >= diploma(6))

→ **buyer** = "yes" (85% Confidence)

Id	age	education	gender	income	loan amt	...	buyer
1	35	graduate(7)	male	$50 K	$400K	...	yes
....							
1000	25	diploma(6)	female	$32 K	$80 K	...	no

Table 1: Training data set for classification in database marketing

From the classification rule, we can deduce that the attributes, "income", "age" and "education" are influential to the target class of "potential buyer"; while "gender", "loan amount", "number of children" and other attributes are not. Features appearing in ***classification rules*** with high confidence form a good model for representing the characteristics and influencing factors associated to the target class. These are called "objective indicators".

Such *feature selection* method [5], [10] plays an important role in the application of data mining in a real-world context. This is because data are normally collected as a by-product of other business processes, resulting in the presence of many attributes that are irrelevant or redundant to the target-concept. Hence, raw datasets are usually too large for the user to monitor or for many similarity-based learning algorithms to work effectively or efficiently. Moreover, without the help of an objective evaluating method like decision-tree induction, the set of "subjective indicators" specified by the domain experts are often incomplete and susceptible to ***concept drift***. This is expected because humans cannot perceive "*indicators*" if they are not very significant or if they have complicated correlation with the target concept. Humans are also slow to detect change in trends in a dynamic environment.

As an illustration, the domain experts in our analogy may only perceive "income" and "education", but not "age", as influential to the target concept of "potential buyer". They missed out the potential age group between 29 to 38. To make matter worse, over a period of time, "education" ceases to be one of the influential factors that determine a "potential buyer", whereas "age" gradually becomes one. Domain experts could not perceive the ***concept drift*** and the business continues to use the outdated and irrelevant indicators to plan its strategies.

Hence, the key to solve this dynamic problem is to apply decision-tree induction regularly to the most recent training dataset, so as to derive the updated ***classification rules*** [12] for the segment of "potential buyer". In our context, the "objective indicators" obtained are further cross-validate with the existing domain knowledge. Thus, the eventual set of indicators obtained will be more valid and complete, when used to measure the loyalty of customers and their likelihood of defecting.

4. DEVIATION ANALYSIS AND FORECASTING

The next step involves a statistics-based ***deviation analysis*** that monitors the "*objective indicators*" identified for the signs of defection among the customers. This OLAP-based implementation over the data warehouse, which contains historical transactional data of vessel's performance, container's volume, financial records, etc., seeks to detect and trigger off an exception report of defection should any of the customer shows significant deviations from her expected norms.

Deviation analysis [11] is employed in data mining to discover the sets of data, which deviate significantly from its expected norm. For a time t, deviation σ(t) is given by:

$$\sigma(t) = (A(t) - E(t)) / E(t)$$

where A(t) is the *Actual value* for the indicator,

and *E(t)* is the *Expected* value for the indicator, obtained over a time period from the temporal time-series.

In our context, the expected norm or normative value *E(t)* for each of the 5 to 7 indicators of a customer is first obtained through a ***trend-seasonal forecasting model*** [8] built from the historical time-series, consisting of more than 50,000 records annually per indicator. This forecasting

[1] a multi-dimensional database software with data warehousing and OLAP capabilities

technique significantly improves the accuracy of the predicted normative values because 1) the use of ***seasonal indexes*** (the ratio of the actual value of the time-series to the average for a year) in the model will adjust the forecast according to the annual seasonal pattern in the time-series, and 2) the use of ***exponential smoothing*** method will allow more weights to be assigned to the recent data, thus taking into account the current economic performance. With the normative values of every indicator forecasted for every customer, their respective percentage-deviations are next computed. Those with deviations greater than a certain user-specified threshold are considered to be significant enough for further "interestingness validation" [11]. This is done by comparing the customer's deviation to the average deviation of the aggregated customers operating in the same service-route, as illustrated below.

Deviation Settings :		**Deviation Analysis Result**			
Select	distinct Ship-line, Cons	**Ship-line/Cst**	**c.f.**	**ServiceRoute**	**Interest'g**
From	Indicator I_1	S_1 (-11%)	c.f.	SR_1 (+1%)	√
Where	Deviation > 10 %	S_4 (-11%)	c.f.	SR_5 (-10%)	x
And	Date = "Jan"	C_2 (-15%)	c.f.	SR_3 (-1%)	√

Table 2: Deviation analysis result

In the above analysis, shipping-lines S_1, S_4 and consortium C_2 showed significant deviations of over 10%. These are further compared with the average deviation of the aggregated customers in their respective service-routes. In S_1's case, the general population in SR_1 performs reasonably well (a positive deviation +1%), suggesting that S_1's deviation is *unexpected* and thus *interesting*. In S_4's case, the general population in SR_5 performs equally badly (a deviation of -10%), suggesting that S_4's deviation is expected and thus not interesting. S_4's type of deviation was probably caused by some regional events like the recent economy turmoil in Asia, which affects all the shipping-lines operating in the Asia's service-route. If consistent deviations are also observed across the set of indicators for "deviating" customers like S_1 and C_2, then a monthly exception report will alert the domain experts on these possible "defectors". Domain knowledge and insights are then applied to verify the findings for each of these cases and the suspected potential defectors will be monitored closely for the subsequent months. Persistent deviations would reinforce their likelihood to defect.

Deviation analysis is useful in identifying "trimming patterns", even at the initial stage when the phenomenon is still barely observable. This is more so in the port business, where the sheer size of business commitments, and the establishment of contractual agreements between shipping-lines and the port authority mean that potential defectors will usually take a couple of months to "trim-out" their imports and exports gradually before their eventual withdrawal. Defection predictions help the authority gauge which customer segments are likely to stay put and which are likely to defect. These early-warnings would give PSA ample lead-time to investigate the causes and take rectification actions before defections actually take place. As described in [11], the *interestingness* of a deviation can be related to the estimated benefit achievable through available actions.

Besides performing deviation analysis on the "Customer" concept, similar analysis can also be applied to investigate and identify upcoming or weakening "Market" (continents and countries) and "Service-route" for the purpose of marketing.

5. MULTIPLE CONCEPT-LEVEL ASSOCIATION RULES

With the predictions from *deviation analysis*, we proceed to use a statistical method to mine for ***multiple concept-level association rules*** [2], [4], [9] among the shipping-lines and the consortiums. Knowledge such as, "if Shipping-line S_1 defects, who would follow suit" is important to PSA's *entrepot*-based business. This is because the choice of Singapore as the transshipment point for shipping-lines or consortiums, linking to different service-routes and markets is usually dictated by those major shipping-lines having great influences and massive transshipment dealings. Their defection is likely to cause an "avalanche" of defections from their associated business partners, as well as the breaking up of their respective consortiums. Hence, if PSA can predict the possible "chain-effects" of a defection, she would have a good chance to take relevant actions before a major customer leaves PSA for a new port, thereby influencing its associated business partners to leave too. Despite the importance of this issue, conventional approaches can only give highly subjective predictions obtained through business contacts and information exchanges with other ports and shipping-lines.

In our application, we investigate the transactional associations between the *Inbound Vessels* and the *Outbound Vessels* from the "container transactional database", which contains approximately 120,000 records monthly. The former are vessels that are responsible for bringing in the containers to PSA's port, while the latter are those that ultimately reload and ship out the containers from the Singapore's port. An association rule obtained from using the ***Apriori*** algorithm [2] is of the form:

$$V_1 \rightarrow V_2 \text{ [support, confidence]}$$

Support and ***confidence*** are terminologies used to measure the strength and *interestingness* of an association rule, such that ***support*** $s = P(V_1 \cup V_2)$ and ***confidence*** $c = P(V_2 \mid V_1)$. With these measures, we are able to select the "***strong*** rules of ***high*** confidence" for accurate prediction of associations.

However, these types of vessel-to-vessel associations are overly specific to be of any significance to PSA business. Therefore, we need to mine association rules at multiple concept-level [9]. Generalizing association rules into higher concept-levels is useful in our context because our

problem domain involves background knowledge with some form of natural conceptual hierarchies. For example, we have the concepts "Customer", with the hierarchy: (consortiums ⊃ shipping-lines ⊃ vessels) and "Market", with the hierarchy: (continents ⊃ countries ⊃ ports). For our illustration here, we will only use the concept "Customer". A *consortium* is an association of shipping-lines, formed to gain significance, in order to enjoy a better volume rebate / discount or to qualify for a better charging scheme. Most shipping-lines will therefore form consortiums with their associated business partners. In our context, typical association rules at multiple concept-levels, i.e. *shipping-line* level and *consortium* level are:

Consortium : $C_2 \rightarrow C_4$ [0.5 support, 0.7 confidence]

Shipping-line: $S_1 \rightarrow S_2$ [0.3 support, 0.5 confidence]

The former rule is interpreted as "if consortium C_2 defects, C_4 is also likely (0.7 confidence level) to defect". Likewise, the latter rule is interpreted as "if shipping-line S_1 defects, S_2 may (0.5 confidence level) also defect". Hence, PSA can work out a solution with those valuable customers, like S_1 and C_2 when they show signs of possible defection. This will not only prevent the loss of these valuable customers, the shipping-lines in their respective consortiums, but also safeguard the business with the associated business alliances of S_1 and C_2, i.e. S_2 and C_4.

Therefore, the constant monitoring of associations between the customers can effectively prevent an "avalanche" of unexpected defection caused by the departure of a single customer. Furthermore, the knowledge of associations also allows the marketing department to customize attractive schemes to attract more volume from the customers' alliances identified.

6. CONCLUSION

Currently, a prototype OLAP-based system is helping PSA to identify the fast changing influencing factors for customers' loyalty, monitor for customer segments that show signs of possible defection (deviation) and predict the associated business partners that may follow suit. The success of this application can further demonstrate that the maturity of data mining has reached a point where wider applications to other practical problems are now desirable and feasible. This will hopefully create a chain-effect to motivate the strategic use of data mining in business areas, where conventional approaches fall short.

7. ACKNOWLEDGMENTS

We would like to express our sincere appreciation to PSA for providing the support and assistance in the course of the project. We would also like to thank Roger King and Chris Marshall for their valuable advice.

8. REFERENCES

[1] Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., Swami, A. 1992, An Interval Classifier for Database Mining Applications, Proceedings of the 18th VLDB Conference, British Columbia, Canada.

[2] Agrawal, R., Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In the Proceedings of the 20th VLDB Conference Santago, Chile.

[3] Agrawal, R., Srikant, R. 1995. Mining Sequential Patterns. In the Proceedings of the International Conference on Database Engineering, IEEE, pp.3-14.

[4] Agrawal, R., Srikant, R. 1996. Mining Quantitative Association Rules in Large Relational Tables. In the Proceedings of 1996 ACM-SIGMOD Int. Conference On Management of Data, Montreal, Canada.

[5] Dash, M., Liu, H., 1997. Feature Selection for Classification. Journal of Intelligent Data Analysis, Vol. 1, No. 3. (http://www.elsevier.com/locate/ida).

[6] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U.M., G. Piatetsky-Shapiro, Smyth, P., Uthurusamy, R. (Eds.). Advances in Knowledge Discovery and Data Mining. Cambridge, MA: MIT Press, pp. 1- 34.

[7] Langley, P. 1994. Selection of Relevant Features in Machine Learning. In Proceedings of the AAAI Fall Symposium on relevance. AAAI Press.

[8] Levin, R.I., Rubin, D.S., Stinson, J.P., Gardener, E.S. 1992. Quantitative Approach to Management (8th Edition). Forecasting. McGraw-Hill, pp. 103-138

[9] Han, J., Fu Y. 1995. Discovery of Multiple-Level Association Rules from Large Databases. In the Proceedings of the 21st VLDB Conference, Zurich Switzerland.

[10] Kira, K., Rendell, L.A. 1992. The Feature Selection Problem: Traditional Methods and a New Algorithm. In Proceedings of Ninth National Conference on AI, pp. 129-134

[11] Matheus, C.J., Piatetsky-Shapiro, G., McNeill, D,. 1996. Selecting and Reporting What Is Interesting. In Fayyad, U.M., G. Piatetsky-Shapiro, Smyth, P., Uthurusamy, R. (Eds.). Advances in Knowledge Discovery and Data Mining. Cambridge, MA: MIT Press, pp. 465- 515.

[12] Quinlan, J.R. 1986. Induction of decision trees. Machine Learning, Vol. 1, pp.81-106

[13] Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, California.

[14] Shavlik, J.W., Dietterich, T.G. 1990. Readings in Machine Learning. Morgan Kaufmann.

[15] Thornton, C.J. 1991. Techniques in Computational Learning -- An Introduction. Chapman & Hall, pp.70-85.