Data Mining for Life Insurance Knowledge Extraction: A Survey

Nanthawadee Sucharittham† Thanaruk Theeramunkong†
Choochart Haruechaisak* Bao Tu Ho** Dam Hieu Chi**

† School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand
* Speech and Audio Technology Laboratory (SPT), National Electronics and Computer Technology Center, Thailand
** School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan
{nanthawadee,bao,dam}@jaist.ac.jp, thanaruk@siit.ut.ac.th, choochart.haruechaisak@nectec.or.th

Abstract

Although life insurance is useful reducing effect of unexpected vital accidents by sharing risk among a group of insurance members. Most people usually do not realize its importance, its benefit, and also do not know which type of life insurance should be acquired due to variety and complexity in insurance types. One of the challenging tasks is to facilitate a new customer to find a suitable insurance type that matches his/her behavior and life-style pattern in order to gain the highest benefits and satisfactions. This paper provides a survey on data mining techniques applied to extract knowledge from life insurance databases. The survey covers data sources, mined knowledge and mining techniques. Finally some suggestions are given as a guideline for research on mining life insurance data.

Keywords: service science, intangible product, customer satisfaction, knowledge extraction

1 Introduction

Recently the service sector becomes larger and gains more impact in our society. However, according to Spohrer et al., the service sector accounts for most of the world's economic activity, but it is the least-studied part of the economy [25]. Some recent promotions of research on service and its system includes IBM's project named Service Sciences, Management and Engineering (SSME), Oracle and IBM's industry consortium called the Service Research and Innovation Initiative and several service companies' service-value networks, such as those in Google, Amazon, Airlines, UPS, Wal-Mart. More details can be found in [11].

As a classic service but complex structures, insurance becomes an interesting target field of study for service science. Currently due to various conditions and needs, a large number of insurance types and policies are developed. With a suitable type and policy, the customer and/or his/her family can sufficiently obtain income when an unexpected inconvenient event, such as accident, sickness or fatal death, takes place. Looking as a whole, good insurance plans can reduce risk of both customers and insurance companies. However, the life insurance is also not so popular in most developing countries including Thailand due to difficulty of insurance terminologies, complexity of insurance policies and intricacy of benefit assessment. Consulting proficiency is required to extract a customer's behavior and life-style from huge data in order to describe effect of each policy to the customer and to determine the level of coverage for an individual people [20].

Recently information technology has been developed to the level of practice. Techniques in knowledge discovery and data mining (KDD) are studied. Especially, purchase and claim patterns mined from insurance databases can be used to suggest current and perspective customers to reduce risk and maximize beneficial outcomes.

In this paper, we survey possible application of data mining techniques on life insurance data, including those related to group insurance, non-life insurance and life insurance. The purpose of this paper is to review past literatures, to discuss on data types and data mining provided for extracting interesting patterns or knowledge from data in insurance industry. The rest of this
paper is organized as follows. Section 2 describes insurance knowledge extraction. Challenging
tasks in mining insurance data are explained in Section 3. In Section 4, a conclusion is made.

2 Insurance Knowledge Extraction

Knowledge utilization provides an insightful understanding for people and organization such as allowing an operation in accordance with their demands, making coordination between the parties, or facilitating the management of information to make better decisions. That means knowledge is valuable however it needs some actions to find the way to make use of knowledge on hand. Knowledge extraction is the process to explore and construct knowledge from various sources of data. This knowledge is extracted into the solid results for easy understanding and we can utilize this useful knowledge to support our purposes.

One of the valuable data sources in a company is its database storing the customer activity records. It is possible to analyze customer behavior from such database, including both structured and unstructured data. For instance, the structured data might consist of demographic information including private information. An unstructured data is narrative information such as opinions from customers and evaluation results of underwriting.

By utilizing knowledge extraction from these structured and unstructured data in the insurance industry to analyze customer behavior, we can understand the way to obtain, use and dispose products. A new customer can consume this knowledge acquired from the current customer behavior if there are the same behavior, life style and conditions.

However, it is not straightforward to figure out which insurance policy can be recommended to a specific customer since even if the most similar customer, he/she may occupy minor different characteristics, life-style and conditions. Challenging in this task is how to collect a large amount of data for analysis and how to analyze such big data.

Customer behavior can be exploited by using knowledge extraction to support in various purposes from insurance activities. The initial criteria which can determine the basic rules for assigning an insurance type to an individual customer have three viewpoints. In this paper, we present how data mining could be used to extract knowledge to reply for three initial criteria as follows.

2.1 Which group or segment of a customer likely to be? This answer shows a position of a customer on current market situation before making new contract from customer segmentation.

2.2 Which product would best suit the customer needs? This answer shows the patterns of product concurrence set which present the customers behavior when they bought the set of products.

2.3 What is the risk rating of customer related to the amount of assurance? This answer shows the prediction for risk level of each customer. Company can also predict the loss when accept this case to be insured.

2.1 Customer segmentation extraction

The important of customer segmentation is to make new customers know their current positions which segment matches up character or life style. Because the customer separations is used for separating the group by determining the similarity of member make more understand the special characteristics of each customer segment. When new customers know themselves or business could comprehend the outstanding characteristics of each customer segment, they can propose the right product or right service to the right segment.

A traditional method to propose new product to new customer is to select by simply target from basic knowledge base or experience of selling-agents or another marketing channel. From this reason, it is necessary to identify the segment of each individual customer utilize some cluster analysis techniques to improve the traditional method.

Generally, insurance company gets demographic data of customer when customer made new contract. They can utilize demographic data of customer such as education, gender, income level, marital status, age, premium, and life stage for extracting customer segmentation. One of data mining technique called “clustering technique” which attempts to group a set of given data into classes, so that members within group have high similarity in comparison to another but these members are very dissimilar to other clus-
ters. Similarities or dissimilarities are assessed based on the attribute values describing the objects where distance measures are used. In generally, the clustering has many techniques such as partitioning method, hierarchical method, density-based method, K-means clustering. However, we give an example how clustering technique works in the area of customer segmentation from K-means clustering.

Customer segmentation can perform extraction using a clustering techniques based on determined attribute values to create new catalogs. The expected result from this analysis target is to find out the position of new comers in the current insurance market. A new customer will know which segmentation matches to his character or his life style by demographic customer data analysis.

Refer to Table 1, there are provided attributes from history demographic customer data such as gender, age, occupation, income. An analysis attribute is life stage attribute. Customers who have attention to make insurance contract should know their position from the results after exercise clustering techniques.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Occupation</th>
<th>Income</th>
<th>Life Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>7</td>
<td>Student</td>
<td>0</td>
<td>Juveniles</td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Soldier</td>
<td>15,000</td>
<td>First Jobbers</td>
</tr>
<tr>
<td>Male</td>
<td>61</td>
<td>No job</td>
<td>15,000</td>
<td>Retirees</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
<td>Student</td>
<td>0</td>
<td>Juveniles</td>
</tr>
<tr>
<td>Female</td>
<td>35</td>
<td>Officer</td>
<td>35,000</td>
<td>Matured</td>
</tr>
<tr>
<td>Female</td>
<td>55</td>
<td>Teacher</td>
<td>65,000</td>
<td>Matured</td>
</tr>
</tbody>
</table>

Table 1: Sample history demographic customer data

The data from the determined attributes can perform a clustering to separate the cluster of customer. In brief, the steps of K-means consist of do partition objects into non-empty subsets, compute mean point of the clusters of the current partition, assign each cluster center to the mean of its assigned item and then repeat until convergence. Since clusters are not predefined, a domain expert is often required to interpret the meaning of the created clusters. The supposed results after took experiment from K-means clustering consists of a first cluster is male, 32, soldier, 16,500, first jobbers and a second cluster is male, 35, officer, 25,000, matures in terms of gender, age, occupation, income and life stage. Next, we put a test data as male, age is 30, occupation is officer, and income is 20,000. This case the cluster of this test data is the cluster one and life stage should be first jobbers. It might be different from find out result from basic analysis which may classify this case to be matured.

Many researchers often utilized clustering techniques in their tasks. Various papers used K-means algorithms in their experiments such as A.B. Devale and Dr. R.V.Kulkarni used K-means clustering technique for classify the group of customer from demographic information such as age, occupation, income and education level [2]. This paper introduced the basic concept of each data mining concept for knowledge discovery in insurance business. In clustering, they used K-Means in order to give an example in case of company does not have any predefined any labels. They presented based on the outcome of the grouping they will target marketing and advertising campaigns to the different group for particular types of policy. Shu-Hsien Liao et al. utilized the concept of data mining in real work to determine the level of customers’ needs in life insurance products by extracting specific knowledge patterns and rules from consumers and their demand chains [27]. This paper also used K-means clustering to classify the customer group from demographic data which is quite more different features they not only used basic information as name, gender, birth date but also included life stage needs, residential house level, life partner, and transportation mode. Depth details of features can distinguish classification of customer groups more clearly.

Many previous works classified customer groups in different purposes from clustering method. Not only clustering method but also other methods can utilize for customer segmentation. For instance, the association rules are widely utilized especially market baskets theory for finding existing customer patterns such as Yu Yan and Haiying Xie recommended using association analysis in customer behavior model to improve the value of customer in customer relationship management task (CRM) of insurance industry [29]. For example, the recommendations of new insurance service to old customer in order to upgrade sales volume and improve return rate of sale-investment. Moreover, they also gave some recommendations about an important level of informatization in pattern construction. Yongqiang Chen & Leifang Hu employed data mining knowledge into customer relationship management (CRM) in insurance business to enhance the analysis of customer value model and market classification model [28]. In part of
customers values model constructed from purpose of lifetime value conception from formula (customers’ lifetime value = potential value + current value) by adapt knowledge of economic into data mining process. In sum, the final model got potential value of insurance equal to customers has potential to purchase policy plus customers’ recommendation value plus the other value. Market classification models were constructed by decision tree from huge number of customer data. Using data mining can distinguish different sorts of customer to be many customer classification as customer satisfaction, customer loyalty, customer declining, individualized service, etc.

2.2 Occurrence product patterns extraction

Finding out the suitable of life insurance product for each individual people is a difficult task because there are various kinds of insurance product which is not only basic product type but also rider concurrence product. They are very difficult to understand the useful of the different purpose of each product and difficult to know which group of product should be hold simultaneously in order to support the highest benefits.

At present, selling channels or agents still offer a new product to a customer from their experience without considering the highest benefit of customer because it is not straightforward to understand clearly for the relationship of product benefits between basic and rider so they cannot recommend the right product to the right person correctly. Normally, they offer base on their idea and sometimes base on their commission of each product at that time.

Normally, life insurance product has two product types. First type called basic product which means one main policy have only basic product for each insured. Second type called rider product which is supplement product into basic product. Each basic product has different purpose. Some of basic product has purpose to build up cash value and others is a protection purpose in short term. The example of basic product in different purposes such as [17]

- Term life insurance is for a limited time such as 10, 15, or 20 years.
- Universal life insurance is flexible and allows to increase premium over time
- Whole life insurance is permanent insurance and provides lifetime coverage.

The premiums are fixed and builds cash value. The value functions as a saving account that may be tax deferred.

Rider product is the additional benefits that can be added into a basic product. Rider product is increase a level of coverage and can be blended, for an additional cost, according to future coverage needs [19]. Here is some examples of common riders in life insurance.

- Waiver of Premium (WP) is a rider attached to the basic policy to maintain coverage in the event of total and permanent disability; subsequent premiums are waived.
- Hospital & Surgical Benefit (H&S) rider provides reimbursement for medical expenses in case of hospitalization in a licensed hospital as an in-patient (IPD)
- Accident Death Benefit (ADB) is one-time payment after insured is diagnosed with a terminal illness that will considerable shorten lifespan.

Each insured hold the different kind of riders within one policy which are under only one basic product. That means it has different kind of riders that are often sold together. In order to solve this problem, it is necessary to search the pattern of product group which suit for each individual customer. Some techniques from data mining such as association rules can help extracting the hidden product concurrence group from previous history purchasing of customer data. The target results are finding all the association patterns from various features of history customer data that present the customers behavior when they bought products as a subset of frequent item sets, most of the time also bought the remaining items in the same frequent item set.

Many researchers used association rules mining technique such as apriori algorithm and market basket analysis which are techniques for generating rules from a large data set based on the co-occurrences among the items. The results from this analysis is to learn life insurance item sets of basic and rider product concurrence in order to understand what customer is demanding and these patterns can be used for recommend product sets. The concept of association rules is below.

Association rules mining results in the implication of the form X => Y (If X then Y), where X and Y are the itemsets

**Step 1**: finding the candidate itemsets which pass
the minimum support; Support (X => Y) = P(X and Y).

Step 2: generating rules from itemsets which pass the minimum confidence (accuracy)
Confidence (X => Y) = P(Y|X) = P(X and Y)/P(X) = support (X => Y)/support(X)

Table 2: Sample product occurrences set

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Item sets</th>
<th>Rider</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Term WP H&amp;S1 H&amp;S2</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>UL WP ADB</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>WL WP H&amp;S2</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>UL WP H&amp;S1 ADB</td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>WL WP H&amp;S1 H&amp;S2</td>
<td></td>
</tr>
</tbody>
</table>

Remark: UL = universal life, WL = whole life, WP = waiver of premium, H&S1/2 = hospital & surgical benefit type 1/2, ADB = accident death benefit.

Table 2 is a sample product occurrence data set which consists of basic and rider product in these insurance policies. We presents this example from produces the association rule generation with 3-itemsets rules under condition minimum support = 30% and minimum confidence = 80%. The results found that when people bought one basic product 100% bought one waiver of premium (WP) and at least one type of hospital & surgical benefit (H&S). In terms of rider occurrences, 100% when bought accident death benefits (ADB), people bought waiver of premium (WP) and hospital & surgical benefit (H&S).

Association rules mining plays an important rules to find out hidden pattern from basic product sell concurrence with rider product. Besides, many researchers also use a statistical concept to analysis this point such as K-nearest neighbor, pearson correlation. A.B. Deval and Dr. R.V. Kulkarni discussed how data mining method perform the sophisticated classification and correlation between policy designing and policy selection [2]. They presented association rules for determined the frequent item sets based on a predefined support like a riders that are often sold together with basic products. They used market basket analysis to determine capitalized on the association between different policies that are sold for different purposes. In part of classification algorithms, A.B. Deval et al. presented classification method of K-nearest neighbors which was used for classifying product selection in each age range and occupation, providing big pictures how to benefit from K-nearest neighbors by do segmentation based on income, paid premium, premium mode and net of assurance price. They showed many examples to classify the product type in relationship of each features above such as “Life security”, “Tax benefit”, “Investment” [2].

Nan-Chen Hsieh & Kuo-Chung Chu used K-means classified customers into five clusters and then applied apriori algorithm as association analysis for each product group in these customer clusters [18]. This experiment investigated what functionalities best fit the customer’s needs in life insurance product by extracting the specific knowledge patterns and rules from customer behavior and their demand chain. Generating new product needs not only customer data but also risk rate, ridership and ridership were modeled based on limited data. The main purpose is to share the risk and reduce premiums after adding in expense and profit, a final premium is determined more accuracy.

In order to improve the performance of product recommendation, we can utilize the benefit by integrating knowledge from customer model to product model to improve this process. Knowledge Discovery in Databases (KDD) process in data mining can help merging customer data from different platforms. A.B. Devale and Dr. R.V. Kulkarni presented the statistical principles can be enhanced the capabilities of data mining method to improve directly into insurance procedures [2]. For example, Pearson correlation of statistical methods to specific correlation coefficient in positive and negative correlation was used for finding out a correlation between policy designing and policy selection.

2.3 Risk assessment

Risk assessment is one of the core functions of insurance business. It estimates a likelihood and impact of risk in order to represent the different level of impact with this risk. Insurance industries estimate and set the risk rating for predict the loss from accidents, health claims, or disaster rates in order to support the underwriting process which set up risk rating of customer by placing customer in a predetermined risk class with ambiguous names such as “super preferred”, “preferred”, “standard” and “substandard”. This rating is estimated from customer health and physical characteristics (such as age, gender, height, weight, alcohol, drug, tobacco use or extreme sports) and this rating will impact directly to
premium rate which customer must pay to make contract with insurance company. Traditional method made use of actuarial statistics within the difficulty to predict base on stochastic in nature. It is necessary to construct risk prediction model from some techniques. Data mining and various fields of machine learning, statistics, and optimization techniques are utilized.

In this paper, we give some examples of previous works from various kinds of techniques. The specific analytical techniques from data mining such as market basket analysis, clustering technique, and statistical method can help improve the fraud claim process. IBM T.J.Watson research center [4, 5] also developed method called “ProbE” (probabilistic estimation) predictive-modeling algorithm. This method uses C++ kernel rule based models of insurance risk, where each rule represents a risk group. The purpose of this method is for underwriting profitability analysis which mines property and casualty insurance policy and claims data to construct predictive models for insurance risk. In actuarial science, accuracy is measured in terms of statistical confidence intervals that are estimated risk parameters can deviate from true values. In this experiment, there are concerning point about avoiding over-fitting, it occurs when the best model relative to the training data tends to perform significantly worse when applied to new data.

Jianbing Xiahou and Yang Mu presented the comparable result between traditional approach and data mining approach [15]. Traditional method is statistical using generalized linear models (GLMs). It is an extension of general linear function \( Y = X^T \beta + \epsilon \) which consists of three factors: random factor means \( Y \) follows a normal distribution, \( \mu = EY \); system factor, namely \( \eta = X^T \beta \) (\( X^T \) is transport matrix of \( X \)); connection factor, namely \( h=m \). In practice, \( Y \) does not follow a normal distribution. Researchers also comment in the general statistic method in this experiment. This task needs to process data several times to fitting in practice since the section of risk variables and classification of variables are not very exact. That means the cost and complexity are huge because input matrix is huge. Moreover, the models they get by fitting may not be convergent. They also compared results of experiment to select the risk in classification using decision tree algorithm to rate making into risk analysis. They expected decision tree classifier can solve the problem of multitudinous information classification. However the results form this experiment still did not satisfy since they found risk variables which make an important impact on rate making are too few in their models. Besides, the risky variables provided by insurance agents are not ranked reasonably. Yu Yan and Haiying Xie explained the useful of data mining in task of dividing credit customers into several grades, predicting the customer risk, making credit rating and avoiding service risk [29]. They also explained in part of classification and prediction to construct forecasting model of future tendency of data using decision tree, neural network, and Bayesian network.

3 Data mining challenging for knowledge extraction in insurance business

An efficiency of knowledge extraction is depend on a quality of data which need all preprocessing processes in KDD especially cleansing data to handle missing data, reducing noise and accounting for time series. However, we still face with some obstacles of high volume and complexity of data and now many researchers are interested in solving problem with data mining hybrid techniques to get high performance results.

3.1 High volume and high complexity in the different kinds of data

There are the different kinds of data types used in data mining process such as structural and unstructured data type. A structured data is readily searchable by simple, straightforward search engine algorithms; whereas an unstructured data is essentially the opposite. All data types are collected in different repositories such as relational databases, time-series database, data streams or multimedia database. We need to integrate all data platforms into the same platform before extract knowledge by data mining technique. Most of historical data in company is the structured data however the data has huge volume of data stream. The unstructured data also plays important roles now because some data in company is narrative platform and there are plenty of free texts such as customer comments are retrieved from various websites. The unstructured data needs some necessary steps to
transform unstructured data to be a structured data before perform data mining by some techniques such as natural language processing (NLP) and human language technology (HLT).

Pennock et al. performed a semantic classification of reviewing messages in unstructured data using natural language processing by N-gram and be continued data mining processes to compare the results between support vector machine (SVM) and Naïve Bayes from C|Net text corpus [6]. Hu & Liu extracted the co-occurring sets of terms from unstructured data then they did feature generation from these occurrence sets and used association rule mining to find all frequent items set. In this context, an item set is a set of words or phrases that occurs together. They used apriori algorithm in step of find all frequent item sets by satisfy a user-specified minimum support however they did not construct any sequent patterns. In sum, the process of unstructured data utilization needs more complicated methods such as NLP, IR to preparing data before extract knowledge from data mining algorithms [3, 12, 13].

3.2 Hybrid techniques in data mining research

In order to make data mining method more powerful, many researchers presented novel methods by consolidated data mining algorithms mixed together or consolidated with other knowledge such as artificial intelligence, information retrieval (IR), NLP. For example, Nan-Chen Hieh & Kuo-Chung Chu presented in the topic of enhancing customer behavior analysis by data mining techniques [18]. This paper showed two-stage frame work of consumer behavior analysis, and the key feature is a cascade involving self-organizing map (SOM), a neural network and a decision tree inducer. This hybrid framework used clustering techniques to pre-process input samples into homogeneous clusters, and decision tree techniques to build customer profile. This paper did the clustering customer behavior from general demographic information such as repayment cycle, number of purchases, card age month, block code, sex and credit line. They gave some suggestions about the clustering algorithm affects the performance of the clustering result. The quality of input samples leads to misclassification.

Shu-Hsien Liao et al. investigated what functionalities best fit the consumers’ needs in life insurance products by extracting specific knowledge patterns and rules from consumers and their demand chains [27]. This paper used apriori clustering algorithm to illustrate the marketing segmentations and demand chain on life insurance in Taiwan. They also did segmenting of potential insurance buyers into similar groups using K-means and association rule based on buyer’s previous purchased.

4 Conclusion

We investigated the knowledge extraction from data mining techniques used for life insurance. Main purpose is to provide more understanding in the current customer behaviors and to utilize them for recommending a suitable policy to new customers in order to gain their highest benefits and satisfactions. Company also has an efficient customer relationship management and gain higher persistency rate. Various data mining methods were used for knowledge extraction. Each method has difference strengthen points to utilize them. In this paper, we gave some examples of data mining methods to give some ideas to utilize them.

Furthermore, we found that many researchers are interested in hybrid techniques to make knowledge construction more efficiency and the correctness of knowledge extraction from data mining need to overcome the obstacle from high volume and high complexity from the different kind of data such as structured and unstructured data which generated from diverse medium.

Acknowledgment

This research was conducted under a grant in the SIIT-JAIST-NECTEC Dual Doctoral Degree Program.

References