LABELING CUSTOMERS USING DISCOVERED KNOWLEDGE CASE STUDY: AUTOMOBILE INSURANCE INDUSTRY

SARAH AZADMANESH [1] and MOHAMMAD J. TAROKH[2]

[1]MSc student in industrial department, K.N.Toosi University,Tehran, Iran

saraazadmanesh@gmail.com

2 Associated professors in industrial department, K.N.Toosi University,
Tehran, Iran

mjtarokh@kntu.ac.ir

1. *Abstract*

In this paper, we used the knowledge discovery in databases and data mining, one of the data-based decision support techniques to help labeling customers in the automobile insurance industry. In most data mining application cases, major tasks including data preparation, data preprocessing, data transformation, data mining, interpretation, application and evaluation, are required. The results of a case study are presented that knowledge discovery of databases and data mining is used to explore decision rules for an automobile insurance company. The decision rules can be used to label the customers as "bad" or "good" for insurance policies.

2. *Keywords*

labeling customers, discovered knowledge, decision tree, ID3

3. Introduction

Customer relationship management (CRM) includes a set of processes and it enables systems to support a business strategy to build long term, profitable relationships with specific customers [11]. Customer data and information technology (IT) tools form the foundation of any successful CRM strategy. Also, the rapid growth of the Internet and its connected technologies has extremely increased the opportunities for marketing and transformed the way that relationships between companies and their customers are managed [14].

From the architecture point of view, the CRM framework can be classified into operational and analytical ([1]; [8]; [21]). Operational CRM apply to the automation of business processes, while analytical CRM refers to the analysis of customer characteristics and behaviors to support the organization's customer management strategies. Analytical CRM could help an organization to better separate and more effectively allocate resources to the most profitable group of customers. Data mining tools are a popular means of analyzing customer data within the analytical CRM framework. Many organizations have collected and stored a wealth of data about their current customers, potential customers, suppliers and business partners. However, the inability to explore valuable information hidden in the data prevents the organizations from transforming these data into valuable and useful knowledge [1]. In order to effectively reveal information in handling data, information technology has introduced the development and application of the techniques of knowledge discovery (KD) and data mining (DM). KD/DM is basically a data-oriented method that merges database management, knowledge representation, and machine

DOI: 10.5121/ijmvsc.2012.3302

learning [3]. It is introduced with both descriptive and predictive ability to discover patterns in historical data that was unknown before, but meaningful and decision-supportable. In response to its applications in business, KD/DM has been widely employed in support of management decisions [6]. Pitta (1998) highlighted KD/DM as an important tool that marketers can use to explore patterns in databases, whereas also emphasizing the one-to-one marketing strategy [16]. Feelders, Daniels, and Holsheimer (2000) presented a fundamental concept for data mining with aiming at its application process [5]. More importantly, the applications in different areas of business shown in literature in the past few years have also witnessed the increased use of KD/DM. With extensive customer data, data mining techniques can provide business intelligence to create new opportunities [14].

Technically, there is much to argue with in order to successfully implement a KD/DM application plan. It is realized that when KD/DM is applied in a field, it systematically pursues six phases: identifying the problem, collecting datasets, preprocessing collected datasets, mining preprocessed datasets, interpreting and using discovered informatics, and also evaluating the discovered knowledge. For identifying the problem, one should specifies the critical question(s) and/or the issues that are associated with the improvement of decision-making capability, like 'who are the potential customers?'

Using data mining tools in CRM is a novel trend in the global economy. A competitive CRM strategy is based on analyzing and understanding customer behaviors and characteristics, for acquiring and retaining potential customers and maximizing customer value. Proper data mining tools, which are good at exploring and identifying practical information and knowledge from customer databases, are one of the best supporting CRM decisions tools [1].

Iranian companies due to the exclusion of services in the hand of governmental section aren't familiar with the importance of customer data and targeting the best customers. They just store their customer data without any expectation and concern. Case company is one of the largest formal governmental companies and the management is looking forward to any tool that helps improving their competitive advantage and keeping their existing position. Insurance companies have limited financial resources and they can't pay all of the recompenses so they should know their customer and their accompanied risk. Labeling customer help them to target the right customer with the lower risk and less behavior anomaly. Labeling customers who bought the car insurance policy can be helpful in advertising the third-party insurance policy too, it's mandatory in Iran. They can use the result in targeting the "good" customer for other insurance policies. We are going to label customers with the help of data mining.

This paper is organized as follows: Section 2 describes the literature review where KD/DM is utilized to explore knowledge in multiple issues and in insurance industry. Discussions are presented in Section 3. Section 4 concludes this research and addresses future research.

4. Literature review

In this section the literature is viewed from two perspectives. One, the published papers that used ID3 decision tree in their applications and the other is the published papers about insurance industry that used data mining techniques and was published from 2000 to 2011.

4.1. Papers that used ID3 decision tree

Prasad et al ([15]) used Machine learning algorithms such as Auto-associative memory neural networks, Bayesian networks, ID3 and C4.5 for diagnosing the Asthma .They presented a comparative study among these algorithms with the use medical expert systems on patient data. They gathered the clinical signs and symptoms of asthma of patients from various resources. Based on the analysis of such data, it was found that the Auto-associative memory neural networks are one of the best among the remaining algorithms in terms of efficiency and accuracy of identifying the proper disease.

Zhilin & Wang ([30]) proposed a fault diagnosis approach based on data mining technologies, in order to improve the maintenance quality and efficiency. By making full use of data stream, they firstly extract fault symptom vectors by processing data stream, and then establish a diagnosis decision tree through the ID3 decision tree algorithm, and finally stored the link rules between faults and the related symptoms into historical fault database as a foundation for the fault diagnosis. This database provides the basis of trend judgments for a future fault.

Mohanty et al ([12]) employed ID3 decision tree (J48) to predict the quality of a web service based on a set of quality attributes. Their experiments were carried out on the QWS dataset. They applied 10-fold cross-validation to test the efficacy of the models. The J48 and TreeNet techniques outperformed all other techniques by yielding an average accuracy of 99.72%. They also performed feature selection and found that web-services relevance function (WSRF) is the most significant attribute in determining the quality of a web service. Later, they performed feature selection without WSRF and found that Reliability, Throughput, Successability, Documentation and Response Time are the most important attributes in that order. Moreover, the set of 'if–then' rules yielded by J48 and CART can be used as an expert system for web-services classification.

In Ture et al ([22]) , they analyzed the simultaneous relationship between risk factors for breast cancer with five decision tree algorithms(classification and regression tree (C&RT), Chi-squared automatic interaction detector (CHAID), quick, unbiased, efficient statistical tree (QUEST), Commercial version 4.5(C4.5) and Interactive Dichotomizer version 3 (ID3)). They compared the relative impacts of each risk factor for breast cancer in the multivariate analysis models.

Ture et al ([23]) determined new diagnostic indexes for the distinguishing of subgroups of breast cancer patients by the decision tree algorithms techniques (C&RT, CHAID, QUEST, ID3, C4.5and C5.0) and analyzing breast cancer patients with Cox regression for disease-free survival (DFS). A retroactive analysis was performed in 381 diagnosed patients with breast cancer. Based on these diagnostic factors, new diagnostic indexes for C&RT, CHAID, QUEST, ID3, C4.5 and C5.0 and Cox regression were obtained. Prognostic indexes showed a good degree of classification, which shows improvement, is possible using standard risk factors. They acquire that C4.5 has a better performance than C&RT, CHAID, QUEST, ID3, C5.0 and Cox regression for determining risk groups using Random Survival Forests (RSF).

Emekci et al ([2]) proposed a new privacy preserving distributed decision tree learning algorithm based on the ID3 algorithm [17] and Shamir's secret sharing [20]. Their suggested algorithm is scalable from computation and communication cost point of view, and hence it can be run even with large number of parties involved. Their goal was implementing constructing decision trees over an arbitrary number of distributed data sources in a privacy preserving manner based on the ID3 algorithm.

4.2. Using DM in insurance industry

Wang et al ([25]) combined quantitative and qualitative analysis, and created a multi-criteria comprehensive evaluation model based on fuzzy mathematic. They proposed it for calculation of risk coefficient for adjusting basic premium rate of Construction Work Safety Liability Insurance (CWSLI).

Lin ([10]) analyzed consumer switching behaviors from consumer psychology and interactions with the society point of view. His results exhibited that when service failures are more severe, higher emotional intelligence consumers exhibit lower intent to switch than those with lower emotional intelligence. Similarly, the more severe the service failures, the more the internal locus-of-control orientation consumers exhibit intention to change their existing utilization than those with external locus-of-control orientation.

Donkers et al ([3]) studied the abilities of a range of models for CLV prediction in the insurance industry. The simplest models can be building at the customer relationship level, i.e. aggregating all services. The more complex models concentrated on the individual services, noticing explicitly to cross buying, but also retention. They discovered that the simple models perform well. Richness of relationship development is expected from more complex models. Unexpectedly, they realized that more complex models do not lead to substantially better CLV predictions.

In [28], Wu et al. used the knowledge discovery in databases and data mining (KDD/DM) to help targeting customers in the insurance industry. A case study of KDD/DM was performed to explore decision rules from company's data base. They suggest that decision rules could be used for investigating the potential customers for an insurance product (existing or new).

Roh at al ([18]) investigated the CRM implementation success factors from three perspectives: efficiency in IS, customer satisfaction in marketing, and firms aggregated profitability. The result of their study discovered multi dimensional metrics for factors that affect profitability through CRM that are reliable. Their measurement model analysis indicated that their suggested metrics have a relatively high degree of validity and reliability.

In ([24]), Verhoef and Donkers introduced a predicting potential value for a current customer model. Their study was the first to focus on the modeling and predicting the potential value of multi-service provider customers. Second, they provide a framework for CRM-managers in multi-service industries, which could be used for predicting customer potential value. This framework considered the data limitations that a company usually has, with using socio-demographic information and transaction information from the customer database only.

5. Application framework

The application framework that we used in this paper is illustrated in Figure 1.It consists of two major components. The first component deals mainly with the task of knowledge discovery and feedback from the application case. The second one is the discussions drawn from the application case, as a reference for any case that is the same as or similar to the current study. The application features are described in Table 1, including the objective of the application, the size of database used, the way to prepare data (e.g. attribute selection) that was mined, the way to granulate attributes, the mining mechanism used the representation of discovered knowledge, and the purpose of the discovered knowledge. Details are described below.

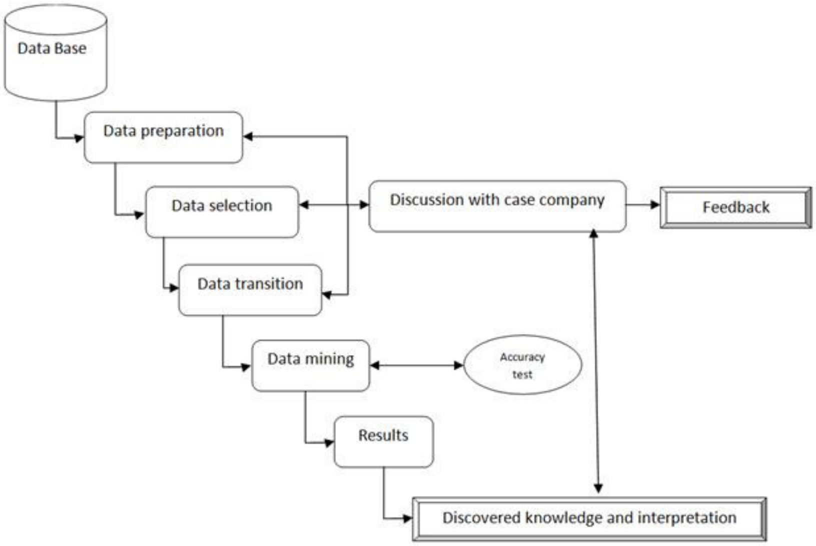


Figure 1. Application framework

Table 1. Features of the company case

Features	Description
Objective	Labeling customers in case company (automobile insurance industry)
Size of database used	19,336
Data preparation	1. Transformation from Persian to English 2. 2 Numeric and 3 text attributes and 1 semantic class selected by a discussion of the case company 3. 4 Concepts in class attribute
Data transition	Listed in Table 2 by a discussion of case company
Mining mechanism used	Induction-based ID3
Representation of discovered	Decision rules via classification
Knowledge	Labeling customers for insurance products

5.1. Case company data and data preparation and selection

The data that were used in this study were selected from a former governmental insurance company in Iran (now a private company). The original data from the case company was in Persian format, so we first transformed these data from Persian to English. The total number of attributes was fifty. It is known that too many attributes involved in data mining will harden the interpretation or cause meaningless discovered information. Therefore, by in-depth discussion with domain experts, we eliminated some of the attributes and finally came to a conclusion of 6 attributes, namely (1) Insured sex (sex of the person who bought an insurance policy) (2) History kind (is there a history about the car or its owner in the data base) (3) Value group (estimation of car value by insurance company granulating in groups like Table 2) (4)Police report (when police came to accident scene how counterparts act, agreed on the damage and the amount of recompense or not agreed, or they didn’t call the police or other conditions) (5)Group of compensation (the amount of money that insurance company paid to insurer after accidents for compensation, granulated in groups like Table 2), and finally Customer labels, type of customer. The attribute of class in the database showed 2 types of Customer Label (see Table 2). Furthermore, because some attributes contained numerical value, Data transition was necessary for the mining mechanism performing the knowledge explosion operation [29]. Generally, the number of granules for an attribute will greatly affect the information granularity definition and finally discovered decision rules [27]. However, a dilemma occurs with this decision since it is very possible that the generated decision trees will be too complex to interpret if the numbers of granules are too large. Specifically, the decision trees that are generated using induction-based technique will be too large to interpret their patterns if there are too many group of granulation for labeling. If too small, a group of granulation may not be able to contain acceptable number of instances, and consequently the discovered results may be meaningless. In spite of full freedom for a granulation method, the number of granules should be carefully chosen to eliminate unnecessary problems [28]. By a discussion with the case company experts, the granulation operation was based on the granulation regulations that are listed in Table 2. For example, the attribute of Value group was granulated into twenty nine levels, from value A, B,C,D to ZC. The defined ranges were 0 to 38500000 Rials(Iran currency) for the value A , 38500000 to 49000000 for value B, 49000000 to 59000000for value C, 59000000to 63500000 for value D, and at last from 1000000000 and more for value ZC, respectively. Note that the class attribute contains 2 types of customer label that are denoted good and bad.

Table 2. Data transition regulation for attributes

<table>
<tr><th>Attribute number</th><th>Names</th><th colspan="7">Data transition regulation</th></tr>
<tr><td>1</td><td>Insured sex</td><td colspan="4">female</td><td colspan="3">male</td></tr>
<tr><td>2</td><td>History kind</td><td colspan="3">New</td><td colspan="2">damaged</td><td colspan="2">No history of damage</td></tr>
<tr><td rowspan="2">3</td><td rowspan="2">Value group</td><td>value A</td><td>Value B</td><td>value C</td><td>value D</td><td>value E</td><td></td><td>value ZC</td></tr>
<tr><td>38500000</td><td>49000000</td><td>59000000</td><td>63500000</td><td>69000000</td><td>…</td><td>1950000000</td></tr>
<tr><td>4</td><td>Police report</td><td colspan="3">Not exist</td><td colspan="2">agreed</td><td>Not agreed</td><td>other</td></tr>
<tr><td rowspan="2">5</td><td rowspan="2">Group of compensation</td><td>Group one</td><td>Group two</td><td>Group three</td><td>Group four</td><td></td><td colspan="2">Group thirty</td></tr>
<tr><td>632000</td><td>737709</td><td>832500</td><td>895000</td><td>…</td><td colspan="2">75500000</td></tr>
<tr><td>6</td><td>Customer Label</td><td colspan="4">Good</td><td colspan="3">Bad</td></tr>
</table>

5.2. Data mining

3.2.1 Mining mechanism used

In this paper, we used the classification method as the mining mechanism. ID3, an important and most widely applied classification method, was used to mine decision rules in the collected database. The ID3 algorithm, introduced by [17], has been wildly used to help measure the information entropy for a set of data under the consideration of multiple classes ([13]; [19]; [26]). ID3, C4.5, and CART adopt a greedy (i.e., non back tracking) approach through the given sets to test each attribute at every tree node and decision trees are constructed in a top-down recursive divide-and-conquer manner. In order to select the attribute that is most useful for classifying a given sets, ID3 uses information gain measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of messages. Let Node N represent or hold the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by

$$\text{Info (D)}= -\sum_{i=1}^{m} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (1)$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Note that, at this point, the information we have is based solely on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

Now, suppose we were to partition the tuples in D on some attribute A having v distinct values, $\{a_1, a_2, \ldots, a_v\}$, as observed from the training data. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, $\{D_1, D_2, \ldots, D_v\}$,where $D_j$contains those tuples in D that have outcome $a_j$of A. These partitions would correspond to the branches grown from node N. Ideally, we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). The amount of information is measured by

$$\text{Info (D)}= -\sum_{i=1}^{m} p_i \, log_2(p_i) \qquad (2)$$

The term acts as the weight of the jth partition. InfoA(D) is the expected information required to classify $\frac{|D_j|}{|D|}$ a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$\text{Gain(A)} = \text{Info(D)} - \text{Info}_A\text{(D)} \qquad (3)$$

In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain(A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum Info_A (D)).[7]

All instances in a granulated database have to consider because they all have to contribute to the generation of decision rules. The discovery process begins partitioning via the value that the attribute work to form a decision tree, and therefore the decision rules can be generated. This process is repeated until same result is obtained to return decision rules.

3.2.2. Learning accuracy test

Table 3. Result accuracy

Class	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Good	0.969	0.035	0.936	0.952	0.997
Bad	0.965	0.031	0.984	0.974	0.997
Weighted Avg.	0.967	0.032	0.967	0.967	0.997

Table 4. Confusion matrix

A	B	Classified as
6451	204	A=Good
441	12240	B= Bad

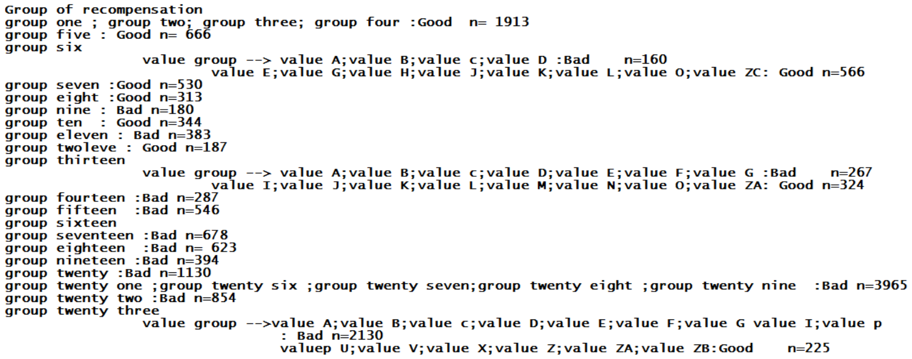
ID3 has shown its prediction accuracy and suitability but we test our instances for accuracy too. As shown in table 3 for class good, True Positive (TP) is 0.969 and False Positive (FP) is 0.035 and F-Measure is 0.952 and precision is 0.936. For Bad class the TP and FP are 0.965, 0.031, respectively and F-Measure is 0.974 and the precision is 0.984. The weighted average precision is 0.967 it's a very good result and its very acceptable. From all 19,336 instances, 18691 (96.6643 %) instances were correctly classified and 645 (3.3357 %) instances were incorrectly classified.

Table 5. Result characteristics

Correctly Classified Instances	18691	96.6643 %
Incorrectly Classified Instances	645	3.3357 %
Mean absolute error	0.0445	
Root mean squared error	0.1491	
Relative absolute error	9.8535 %	
Root relative squared error	31.3905 %	
Total Number of Instances	19336	

3.2.3. Discovered knowledge and interpretation

The study reached the knowledge discovery for the all granulated database cases (19,336 cases). After executing knowledge discovery and data mining application, 3780 rules was discovered. In figure 2 sample of discovered knowledge (tree) with one level of tree (if it is necessary) is shown. The number of *n* implicitly indicates the number of cases in a rule. More precisely, the more the number of *n,* a decision rule has, the more reliable it is because it covers more cases in the transaction database. We think that management will prefer a decision rule if it is supported by a higher number of cases. Since there were 19,366 cases used and 3780 rules discovered, each rule set was supported by 88.8% of accuracy in average. Part of rule sets is illustrated in Fig. 3. These decision rules can be used to serve as an advisor to help in labeling customer based on his/her personnel characteristics.



```
Group of recompensation
group one ; group two; group three; group four :Good  n= 1913
group five : Good n= 666
group six
                value group --> value A;value B;value c;value D :Bad    n=160
                        value E;value G;value H;value J;value K;value L;value O;value ZC: Good n=566
group seven :Good n=530
group eight :Good n=313
group nine : Bad n=180
group ten  : Good n=344
group eleven : Bad n=383
group twoleve : Good n=187
group thirteen
                value group --> value A;value B;value c;value D;value E;value F;value G :Bad    n=267
                        value I;value J;value K;value L;value M;value N;value O;value ZA: Good n=324
group fourteen :Bad n=287
group fifteen  :Bad n=546
group sixteen
group seventeen :Bad n=678
group eighteen  :Bad n= 623
group nineteen :Bad n=394
group twenty :Bad n=1130
group twenty one ;group twenty six ;group twenty seven;group twenty eight ;group twenty nine  :Bad n=3965
group twenty two :Bad n=854
group twenty three
                value group -->value A;value B;value c;value D;value E;value F;value G value I;value p
                                : Bad n=2130
                                valuep U;value V;value X;value Z;value ZA;value ZB:Good    n=225
```

Figure 2. Part of discovered rule set

Noteworthy observations based on the customers' characteristics were also described below.

I. The default label for customers is bad. It is not odd since most of customers were labeled bad before performing the data mining application.

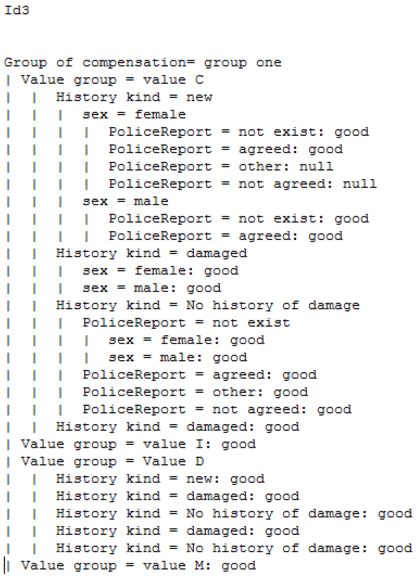
II. The Group of compensation attribute of cases was the most important attribute for classifying customers. In the first place, based on this attribute company can classify its customer. Granulation in this attribute is a very important factor, with no suitable granulation the accuracy of model can be decline dramatically. In this study we first used a C5.0 decision tree for discovering the break point in this attribute and value group (the value of the automobile) and then based on this break point we derived the proper granulation measures. It's obvious that when the amount of compensation increased more customers labeled as bad.

III. After the Group of compensation attribute the value group attribute is the second important attribute for classifying. For this attribute, like the Group of compensation attribute, we used C5.0 decision tree. Suitable granulation of this two attribute has a great effect on the accuracy of model. With increase in value of automobile customers label is change from bad to good, it is sensible that with valuable automobile people drive more carefully.

IV. History and sex attribute of a customer has the less importance in our model. These two attribute have a 3 and 2 possible value, respectively. Most of Men labeled as bad and women have almost equally labeled as bad and good. For history attribute it's like that the default label for all values is bad and just in 'new' value less than half of them labeled as good. This two attribute may have not a significant effect on labeling customer but they are important in classifying them.

V. Police report is the least important attribute for classifying a customer. If the police report wasn't exist, customer labeled good equally as bad, and if both side not agreed with each other most of them labeled as bad, it's not shocking because they couldn't agree on the amount of recompense because in most of these cases the damage is extended.

As shown in figure4, when a car is new and value group is C (third group – one of the lowest group of values) and the amount of recompense is the lowest one the customers mainly labeled as "good", although it's not surprising cause two main attribute for labeling customers is the value group and the group of compensation and in this rule both of them are from the lowest



```
Id3


Group of compensation= group one
| Value group = value C
|   |  History kind = new
|   |  |  sex = female
|   |  |  |  PoliceReport = not exist: good
|   |  |  |  PoliceReport = agreed: good
|   |  |  |  PoliceReport = other: null
|   |  |  |  PoliceReport = not agreed: null
|   |  |  sex = male
|   |  |  |  PoliceReport = not exist: good
|   |  |  |  PoliceReport = agreed: good
|   |  History kind = damaged
|   |  |  sex = female: good
|   |  |  sex = male: good
|   |  History kind = No history of damage
|   |  |  PoliceReport = not exist
|   |  |  |  sex = female: good
|   |  |  |  sex = male: good
|   |  |  PoliceReport = agreed: good
|   |  |  PoliceReport = other: good
|   |  |  PoliceReport = not agreed: good
|   |  History kind = damaged: good
| Value group = value I: good
| Value group = Value D
|   |  History kind = new: good
|   |  History kind = damaged: good
|   |  History kind = No history of damage: good
|   |  History kind = damaged: good
|   |  History kind = No history of damage: good
|| Value group = value M: good
```

Figure 3. Part of discovered knowledge

groups. For other history kinds the label is good too, due the previous discussed logic.

For middle groups of value like G, H when group of compensation is around the middle, they are labeled as good too. As shows in the above rules, one can understand that when the amount of both of group of compensation and value group are near each other, customers are labeled as “good”, and when the value group is higher than the group of compensation the customers labeled as “good” too. Finally when the value group is low and the group of compensation is high the customer label is “bad”.

The importance of labeling customers is for the future when they come for buying another insurance policy for their car, the insurance company has a back ground and can predict the customer label based on his/her demographics and with that in mind can decide whether reinsured the car or not. The other benefit of predicting a customer label is for third-party insurance that it is mandatory in Iran and every vehicle must have one. When a customer is predicted as bad based on his/her car insurance it means the recompense amount was or would be higher than what he/she pays or going to pay to the insurance company, that means she/he had or can have accident or accidents that causes a lot of damage and those accident were serious and can be fatal, so he/she have a high risk and it would be wise to not insure her/his car or insuring the car with a higher tariff.

6. Discussion

As shown in Fig. 1, in this discussions we mainly focus on three things: (1) the experiences of Knowledge Discovery and Data Mining (KD/DM) application in labeling customers for the automobile insurance industry,(2) the usability of KD/DM in supporting decision-making, (3) a lesson from theory to practice. First, the KD/DM technique is basically a data-based pattern used as a management tool in support of decision-making. Although the experience-oriented decision-making has been utilized for a long time, and will continue to have an important role in supporting the management, the decision support mode has incrementally shifted from experience-oriented to information-oriented. At the beginning of the first interview, we experience that the case company seemed unfamiliar with what KD/DM is about and what objective KD/DM can get. In regard to the usability of KD/DM, we have presented the discovered knowledge such as what a rule say, how to explain a rule, what information a rule can get, etc. The usability of the discovered knowledge presented in our research showed that at this stage of KD/DM could be a possible path for improvement of labeling customers. For the measuring the effect of the discovered knowledge at the second stage, it will take time for the case company to collect data and do comparison. For example, does the discovered knowledge significantly help predict the customer label for the next year (or next 2, 3, 4,., years)? The KD/DM techniques, application process and insurance knowledge are all the effective factors that need to be considered carefully. More importantly, the main purpose of the KD/DM technology is not the technology itself, but to help improve management quality, in particular quality of decision-making. This needs continuous involvement of domain researchers and experts with practitioners. An insurance company must be aware of the ability of learning in its data, its potential benefits.

In Iran most of economical sections were exclusively under governmental jurisdiction and now the government gives its jurisdiction to the private sections for economical release. The case company was one of the largest governmental insurance companies in Iran. After this exclusion break down the case company needed new techniques for staying at the top. Because of the governmental exclusion, Iranian companies never had an experience of how they can use their customer data for attracting new ones and retaining the old ones. After the exclusion break down they are eager to use these data for making profit.

7. Conclusion

In this paper, we have described the major tasks necessitated for the KD/DM application, the case application was in an automobile insurance company. Three points obtained from this application case are addressed: (1) KD/DM application experiencing in labeling customers for the automobile insurance industry; (2) the usability of KD/DM techniques in supporting insurance decision-making; (3) a lesson from theory to practice. With fast growth in the volume of information systems collected data, more and more marketers are using data-based decision support tools to enhance the efficiency and effectiveness of their marketing decisions. Using KD/DM technology may need domain knowledge and KD/DM professionals to succeed because it is a highly domain-specific task. Data preparation and transformation is the most time-consuming phase of KD/DM application. From the application viewpoint, the most important part may be the methodology that can both deal with data preparation, transformation and carry out interpretation, utilization and evaluation of knowledge on a domain-by-domain basis.

8. Acknowledgement

The authors thank the editor and anonymous reviewers. The authors would also like to thank the participating insurance company employees for their cooperation.

References

[1] Berson, A, Smith, S, & Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill.

[2] Emekci F , Sahin, O, Agrawal, D, El Abbadi , A.(2007).Privacy preserving decision tree learning over multiple parties. Data & Knowledge Engineering 63 (2007) 348–361

[3] Donkers , B, Verhoef, P, Jong, M. (2007) . Modeling CLV: A test of competing models in the insurance industry. Quant Market Econ (2007) 5:163–190

[4] Fayyad, U, Stolorz, P. (1997). Data mining and KDD: Promise and challenge. Future Generation Computer Systems, 13(2–3), 99–115.

[5] Feelders, A, Daniels, M, Holsheimer, M. (2000). Methodological and practical aspects of data mining. Information & Management, 37, 271–281.

[6] Han, J, Fu, Y. (1999). Mining multiple-level association rules in large databases. IEEE Transactions on Knowledge and Data Engineering, 11(5), 798–805.

[7] Han, J ,Kamber, M .(2006) . Data Mining: Concepts and Techniques. Elsevier.

[8] He, Z., Xu, X., Huang, J, Deng, S. (2004). Mining class outliers: Concepts, algorithms and applications in CRM. Expert Systems with Applications, 27, 681–697.

[9] Hirota, K, Pedrycz, W. (1999). Fuzzy computing for data mining. Proceedings of the IEEE, 87(9), 1575–1600.

[10] Lin, W.(2009).Service failure and consumer switching behaviors: Evidence from the insurance industry. Expert Systems with Applications 37 (2010) 3209–3218.

[11] Ling, R., & Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. Journal of Computer Information Systems, 41, 82–97.

[12] Mohanty R, Ravi V, Patra M. (2010) Web-services classification using intelligent techniques . Expert Systems with Applications 37 5484–5490

[13] Murthy, S. (1998). Automatic construction on decision three from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery, 2, 345–389.

[14] Ngai E., (2005) .Customer relationship management research (1992-2002): An academic literature review and classification. Marketing Intelligence & Planning, Vol. 23 Iss: 6, pp.582 – 605.

[15] Prasad, B. D. C. N., Krishna, P. E. S. N., Sagar,Y.(2011). A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma).Communications in Computer and Information Science, 131,570-576.

[16] Pitta, D. (1998). Marketing on-to-one and its dependence on knowledge discovery in databases. Journal of Consumer Marketing, 15(5), 468–480.

[17] Quinlan, J. (1986). Induction of decision tree. Machine Learning, 1, 81–106.

[18] Roh ,T , Ahn, C, Han I. (2005).The priority factor model for customer relationship management system success. Expert Systems with Applications 28 (2005) 641–654.

[19] Sestito, S, Dillon, T. (1994). Automated knowledge acquisition. New York: Prentice Hall.

[20] Shamir , A .(1979) . How to share a secret, Communications of the ACM 22 (11) 612–613.

[21] Teo, T, Devadoss, P, Pan, S. (2006). Towards a holistic perspective of customer relationship management implementation: A case study of the housing and development board, Singapore. Decision Support Systems, 42, 1613–1627.

[22] Ture, M, Tokatli F, Kurt I(2008 a).Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications 36 2017–2026.

[23] Ture, M , Tokatli F , Omurlu I(2008 b) .The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. Expert Systems with Applications 36 8247–8254.

[24] Verhoef, P, Donkers, B. (2001). Predicting customer potential value an application in the insurance industry. Decision Support Systems, 32, 189–199.

[25] Wang,H.,Ye,G.,Ju,C.(2011). Risk based determination of the prenium rate construction of work safety liability insurance .Computational Risk Management,2, 113-120.

[26] Wu, C.(2003a). Data mining applied to material acquisition budget allocation for libraries: Design and development. Expert Systems with Applications, 25(3), 401–411.

[27] Wu, C. (2003b). On the granulation simplicity for the decision rule discovery in databases: EWI vs. EFI. International Journal of Science and Technology, 14, 28–36.

[28] Wu, C, Kao, S, Su, Y, Wu, C. (2005). Targeting customers via discovery knowledge for the insurance industry. Expert Systems with Applications, 29, 291–299.

[29] Wu, X, Urpani, D. (1999). Induction by attribute elimination. IEEE Transactions on Knowledge and Data Engineering, 11, 805–812.

[30] Zhilin, Z., Wang, P.(2010). Fault Diagnosis of Automobile ECUs with Data Mining Technologies. Applied Mechanics and Materials ,40-41, 156-161.