Mining Investment Venture Rules from Insurance Data Based on Decision Tree

Jinlan Tian, Suqin Zhang, Lin Zhu, and Ben Li

Department of Computer Science and Technology
Tsinghua University., Beijing, 100084, PR China

Abstract. Classification is a basic method of Data Mining. In this paper, we first introduce the basic concept of classifier and how to evaluate the precision of the classifier in this paper. Then we expatiate that how to use the Decision Tree Classifier to search the factors which will bring more venture at the guarantee slip, on the basis of the guarantee slip and compensation information database established by insurance agents. As a result, we gain some useful rules which will be useful to control investment venture.

1 Introduction

Data Mining, which is also called Knowledge Discovery in Databases(KDD), is an advanced process of finding and extracting reliable, novel, effective and comprehensible patterns hidden in a large amount of data. Data Mining technologies have brought significant effects to industries and other domains in the recent years. It is only four or five years from theoretic research to developing Data Mining products abroad. Data Mining technology is more and more often utilized in large companies, business, bank, insurance and telecommunication departments. It just puts up a great power of developing potential.

Insurance is a kind of operation with great venture. Venture evaluation has a significant effect to insurance company. Whether an insurance company could be successful depends on choosing a balance between competitive insurance premium and the venture of insurance. Insurance premium is always confirmed by analyzing and estimating some important factors such as individual health of policy-holders at health-insurance, car style at automobile-insurance, and so on. The situation of insurance market is always changing, so insurance companies should establish insurance premium on the basis of analyzing data of former years. At the present time, professionals of insurance companies adopt only curt analytical methods, analysts make decisions by their experience with a large number of data statistics. These curt methods are very difficult to use and affected by subjective factors.

Data Mining provides a circumstance to analyze insurance investment database. There are many methods of Data Mining which can be applied to venture analysis. We will emphasize on Decision Tree Classifier method in this paper, gain some helpful rule of controlling insurance venture by finding more venturesome area from guarantee slip and compensation information database.

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2658, pp. 20–27, 2003.

2 The Basic Concept of Classifier

Classification is a very important method of Data Mining. Classification is the task of assigning a discrete label value to an unlabeled record. In doing so, records are divided into predefined groups. A classifier is a model that predicts one attribute of a set of data when given other attributes. A training set is needed to construct a classifier. The training set consists of records in the data for which the label has been supplied. An attribute is an inherent characteristic in the dataset. The attribute being predicted is called the label, and the attributes used for prediction are called the descriptive at-tributes. A concrete form of stylebook can be represented as $(v_1, v_2, \cdots, v_n; c)$. The v_i expresses as the value of each field, and the c expresses as a class.

The training set is the base of constructing a classifier. An attribute at the training set is defined as the classification label. The type of label attribute must be discrete, and if the number of the label attribute value is fewer(2 or 3 values is the best), the error-rate is much lower. An algorithm that automatically builds a classifier from a train-ing set is called an inducer. After generating an inducer, unlabeled records in the data-set could be built into such specific classes. Classifier also can predict the value of label attribute. There are several basic classifiers as rendered below.

1) *Decision Tree Classifiers.* A Decision Tree Classifier classifies data from attribute set by predicting the label for each record to make a series of decision. For example, a Decision Tree generated from a training set may predict a man with a family, a car which costs from $15000 to $23000 and two children, will have a good credit. Such Decision Tree classifier could be used to judge the credit degree of a person. MineSet, as a Data Mining tool provided by SGI, generates a Tree Visualization to display the structure of the Decision Tree. Each decision is represent as a node at the tree.

2) *Option Tree Classifiers.* Like Decision Tree classifiers, Option Tree classifiers also assign each record to a class. Instead of picking an attribute to split on for the root node at Decision Tree, Option Tree contain special Option Node, the Option Node may split into several branches. For example, an Option Node in a car-producing-area Option Tree may chooses kilometers per gallon, horsepower, number of cylinder, or weight of a car as the attributes. However, one node just can choose only one at-tribute at most at one time in Decision Tree. We could consider more situations synthetically when using Option Tree. Option Tree is generally more accurate than Decision Tree, but larger.

3) *Evidence Classifiers.* An Evidence Classifier classifies data through checking probability of some specific results of an attribute. For instance, it may estimate a man with a car which costs $15000 to $23000 has a probability of 70% to have a good credit, but the remain 30% person may have unreliable credit. Evidence Classifier predicts the classification result with the maximum probability on the basis of a simple probability model. MineSet Evidence Visualizer displays the result of evidence classification. It gives answers to users' questions such as "if $\cdots$ how about $\cdots$".

3 How to Evaluate the Precision of Classifiers

When a classifier is built, it is useful to know how well you can expect it to perform in the future (what is the classifier's error-rate). Factors affecting classification error-rate include:

1) *The number of records in the training set.* Since the inducer must learn from the training set, the larger the training set, the more reliable the classifier should be; how-ever, the larger the training set, the longer it takes the inducer to build a classifier. The improvement to the error-rate decreases as the size of the training set increases.

2) *The number of attributes.* More attributes mean more combinations for the inducer to compute, making the problem more difficult for the inducer and requiring longer time. Note that sometimes random correlations can lead the inducer astray; consequently, it might build less accurate classifiers (technically, this is known as "over fitting").

3) *The information in the attributes.* Sometimes there is not enough information in the attributes to correctly predict the label with a low error-rate (for example, trying to determine someone's salary based on their eye color). Adding other attributes (such as profession, hours per week, and age) might reduce the error-rate.

4) *The distribution of future unlabeled records.* If future records come from a distribution different from that of the training set, the error-rate probably will be high. For example, if you build a classifier from a training set containing family cars, it might not be useful when attempting to classify records containing many sport cars, because the distribution of attribute values might be very different.

There are two common methods of estimating the error-rate of a classifier as de-scribed below. Both of these assume that future records will be sampled from the same distribution as the training set.

1) *Holdout.* A portion of the records (commonly two-thirds) is used as the training set, while the rest is kept as a test set. The inducer is shown only two-thirds of the data and builds a classifier. The test set is then classified using the induced classifier, and the error-rate or loss on this test set is the estimated error-rate or estimated loss. This method is fast, but since it uses only two-thirds of the data for building the classifier, it does not make efficient use of the data for learning. If all the data were used, it is possible that a more accurate classifier could be built.

2) *Cross-Validation.* The dataset is splitted into k mutually exclusive subsets of approximately equal size. The inducer is trained and tested k times; each time, it is trained on all the data minus a different fold, then tested on that holdout fold. The estimated error-rate is then the average of the errors obtained. Cross-Validation can be repeated multiple times (t). For a t times k-fold cross-validation, $k \times t$ classifiers are built and evaluated. This means the time for cross-validation is $k \times t$ times longer. Increasing the number of repetitions (t) increases the running time and improves the error estimate and the corresponding confidence interval.

Generally, a holdout estimate should be used at the exploratory stage, as well as on dataset over 5,000 records. Cross-validation should be used for the final classifier building phase, as well as on small datasets.

4 Application of Decision Tree Classifier at Insurance Operations

Decision Tree method comes from Concept Learning System (CLS), and then ID3 method emerged as a peak of Decision Tree algorithm. The method has evolved to C4.5 at last which can deal with continuous attributes. Other famous Decision Tree methods include CART and Assistant.

The input of Decision Tree construction is a set of data with class-label, and the result of the construction is a binary tree or a multiple tree. The inner nodes (non-leaf nodes) of the binary tree generally represent as a logical judgment, such as $a_i = v_j$. (a_i is an attribute of some class, v_j is the possible value of the attribute.) The branches of the node are the result of the logical judgments. Each inner node of the multiple tree represents as the attribute of some class, the branches of the node represent all values of the attribute. The number of the branches equal to the number of possible values of the attribute. The labels on Leaf nodes are class for some instance.

Decision Tree starts from the root of the tree, and taking appropriate branches ac-cording to the attribute or question asked about at each branch node. One eventually comes to a leaf node. For example, multiple tree, if all of the data in training set belong to the same class, they will form a leaf node, and the content of the leaf node is the label of that class. Otherwise, the method will choose an attribute with some strategy, dividing the dataset into several subsets according to possible values of the attribute, making the data of each subset have the same attribute value, and then handling each subset in the same way recursively. The binary tree also follows this method except for choosing a reasonable logical judgment.

We will introduce how to utilize MineSet classifiers, the product of SGI, to mine hospitalization insurance data of some city. The hospitalization insurance database consists of individual information table, company information table, periodical (in a month) compensation table and so on. The concrete contents of each table are rendered below:

Table 1. Individual information table

Individual Insurance No.	Name	Sex	Date of Birth	Company No.	Total Salary per Year	Insured Date
3504274308250011	X	male	19430825	0000000663	19411	19970701
3502116405101511	Y	female	19640510	0000000663	15529	19970701
3502115409043551	Z	male	19540904	0000000664	7051	19970901
...	...	...	...	...	...	...

Table 2. Company information table

Company NO.	Company Name	Area Code	Type of Company	Insured Date
0000000330	computer corporation	05	03(enterprise)	19971101
0000000331	tade informatino center	03	03(enterprise)	19970901
0000000352	maternity hospital	01	02(public institution)	19970701
...	...	...	...	...

Table 3. Compensation table in one month

Compensation Bill No.	Compensatory Clerk NO.	Individual Insurance NO.	Compensatory Money	Compensatory Date
424300	01	3526017202011021	17.78	19980101
424190	06	3502056009140011	78.2	19980101
424191	19	3502047201172011	274.5	19980101
...	...	...	...	...

The procedures of Data Mining are discussed below:

1)*Preparing the Data.* We should prepare the data before data mining. For example, we should remove redundant information in the dataset, such as individual name, company name, insured date and so on. We also should make a statistic of compensation times of hospitalization insurance in a period of time. There is an individual compensation information table rendered below after preparing the data.

Table 4. Individual compensation information table

Individual Insurance NO.	Age	Total Salary per Year	Type of Company	Area No.	Compensation Times	If Compensating
3502043808264031	60	7051	03(enterprise)	03	0	0(no)
3502114704291511	51	14287	02(public institution)	01	8	1(yes)
3502042604134011	72	6376	09	01	21	1(yes)
...	...	...	...	...	...	...

2) *Analyzing the Data.* MineSet can build a classifier to predict one particular attribute when given some attributes in a set of data. The attribute being predicted is called the label, and the attributes used for prediction are called the descriptive attributes. MineSet can build a classifier automatically from a training set. The training set is consists of records whose labels are already given on the basis of existent attributes. After the generation, the classifier could be used to classify the records which have no label attribute in the data set. The value of the label can be predicted by the classifier.

Whether policy-holders claim for compensation is the most concerned information when analyzing insurance operation. Towards the dataset mentioned above, we define the attribute "if compensating" as the label attribute. Other in-

formation such as "individual insurance NO." belongs to irrelevant information. The attribute "if compensating" is derived from the attribute "compensation times", so "compensation times" can be removed because of the repetition. The remains of the attributes include "age", "total salary per year", "type of company" and "area code". The training set consists of all of the compensation information of that month.

3) *Data Mining.* We firstly apply "column weightiness" method of MineSet to find the columns which are more effective to label attribute than other columns, so we will avoid subjectiveness based on our experience in this way. The results of "column weightiness" method are three attributes, "age", "total salary per year" and "type of company", which are most effective to label attribute.

Select the "Decision Tree" mining tool, select the mode as "Classifier and Error", and set some options of that mode, then push "go!" button to run the inducer. At last we get a Decision Tree on the insurance dataset. Fig. 1 illustrates the Decision Tree.

4) *Analyzing and Comprehend the Data.* MineSet provides us a binary tree, and it can make a decision at each node according to descriptive attributes. Pointing to a node causes the specific information of the node to be displayed. All possible out-comes are marked on the horizontal lines emanating from each decision node. Each line indicates the value against which the attribute of that mode was tested. Analyzing the specific information of the root node, we can see that there are 6401 records in the training set. The number of customers who had not claimed for compensation is 5377, at the rate of 84.00%. The number of customers who had claimed for compensation is 1024, at the rate of 16.00%.

Note that in this tree the root split on the age of the policy-holders, the age is the most important factor, this result matches our daily experience that older person may not be in a good health condition. However, it is hard to distinguish accurately how old a person can be regarded as an "aged person". MineSet mining tools could give an accurate quantitative conclusion. In our example, we can see that the root node split into two branches by the age of 56. The left branch ($age < 56$) contains 4140 records, and the number of customers in the left branch who had not claimed for compensation is 3742, at the rate of 90.39%. The number of customers who had claimed for compensation is 398, at the rate of 9.61%. The right branch ($age > 56$) contains 2261 records. The number of customers at the right branch who had claimed for compensation is 626, at the rate of 27.69%. The compensation rate increases notable at the right branch. Applying the mining tools to hospitalization insurance dataset, we just gain a rule of the venture of insurance investment that "There is a higher compensation probability when a policy-holder is older than 56." If we apply database query method to such dataset, some condition must be given beforehand, and it will be very difficult and over work loaded by analyzing data statistics artificially.

We can get some other rules about compensation at the right branch of the root node. For example, next factor is "total salary per year". Considering that policy-holders with high salary may pay more money on taking exercises and health care, on the other hand, policy-holders with low salary may pay less. So

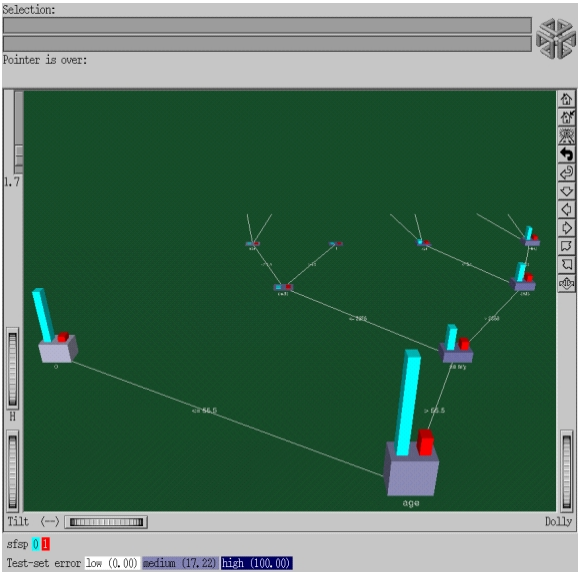

Fig. 1. The Decision Tree on the insurance dataset

it is credible that salary has an obvious influence of compensation situation. The factor "type of company" is another factor on the right branch. We can see from the tree that the compensation probability of the policy-holders who work at enterprise is much lower than that of the policy-holders who work at public institution. Combined with the concrete circumstance of hospitalization insurance domestically, we can explain such result in this way: The payment of fee-for-service is related to the style of company. The policy-holders who work at enterprise will pay more of the total fee, and insurance company will pay lesser. But the policy-holders who work at public institution will pay much less fee of the total and insurance company will pay most of it. Under this circum-stances, the policy-holders who work at enterprise will not go to see the doctor if he or she has a light sickness.

We can predict the compensation probability in the future according to the Decision Tree and detailed information of policy-holders, and then adjust the fee criterion of some kinds of policy-holders on the basis of compensation probability which has been predicted. Just for example, considering a policy-holder at the age of 58, working in enterprise and the total salary of 12000 per year, we follow the binary tree from root to leaf and predict that the compensation probability of that person is 9.84%, lower than the average probability. So the insurance company may decrease the insurance premium of such policy-holders. However, considering a policy-holder at the age of 59, working in public institution and the total salary of 9500 per year, the Decision Tree predicts that the compensation probability of that person is 37.56%, much higher than the average probability.

So the insurance company may increase the insurance premium of such policyholders.

If users want to gain some more detailed rules such as classifying policyholders under 56 years old, MineSet will provide data filtration function. Using such function, you can get the requisite training set by setting "$age < 56$" as the filtrating condition, then apply the Decision Tree method on this training set to get the requisite Decision Tree.

The Option Tree Visualizer's functionality is the same as for Decision Tree except that the Option Tree extends a regular Decision Tree classifier by allowing Option Nodes. An Option Node shows several options that can be chosen at a decision node in the tree. For example, we can choose one of the four branches from the root node. They are "age", "total salary per year", "type of company" and "area node". Instead of using a single attribute at a node in Decision Tree, the option node provides you with several options. However, the time necessary to build an Option Tree under the default setting is much longer than that needed to build a Decision Tree. The Option Tree has two notable advantages:

1) *Higher Comprehensibility.* The option nodes enhance comprehensibility of the factors affecting the class label by showing several choices that can be made. When flying over the tree, you can choose an option that you believe is easier to understand, or better for predictions.

2) *Higher Precision.* The option nodes reduce the risk of making a mistake by averaging the votes made by the options below. Every option leads to a sub tree that can be thought of as an "expert". The option node averages these experts' votes. Such averaging can lead to a better classifier with a lower error rate.

5 Conclusions

In conclusion, the classification method of Data Mining builds Decision Tree or Option Tree based on training sets accumulated in database, and then predicts new data according to the classifier. Classification methods can be applied not only at insurance field, but also at other investment field such as banking and stockjobbing or other trades. It will bring helpful policy supports to managers. Data Mining, as a new technical field, will be applied far and wide in China.

References

1. Heikki Mannnila, Hannu Toivonen and A. Inkeri. Verkamo, "Efficient algorithms for discov-ering association rules," AAAI Workshop on Knowledge Discovery in Databases, pages 181–192, July 1994
2. K.Decker and S.Focardi, "Technology Overview: A Report on Data Mining," ftp://ftp.cscs.ch/pub/CSCS/techreports
3. Tony Xiaohua Hu, "Knowledge Discovery in Databases: An Attribute-Oriented Rough Set Approach," http://www.cs.bham.ac.uk/ anpdm_docs
4. SGI Company, MineSet2.0 Tutorial
5. Gao Wen, "KDD: Knowledge Discovery in Databases," Computer World, vol. 37, 1998