The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 11 || Pages || 45-61 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Data Mining Applied To Construct Risk Factors For Building Claim on Fire Insurance

Chun-Ling Ho

--ABSTRACT---

Recently, risk evaluators in Taiwan non-life insurance companies generally refers to the methods of evaluating the risk of fire insurance that have been developed abroad to carry out professional risk evaluation for huge asset targets or large-scale factories. As for middle and small commercial building projects, wrong underwriting judgment caused by human factors can be foreseen. In many claim factors related to fire, how to objectively evaluate the extent of the risk and establish predicting factors is investigated in this study.This study makes use of the classification of data mining and predicting technology to analyze and extract the useful information for decision-making according to client data by referring to past expert experience. The important factors used in the rate regulations approved by the Ministry of Finance, R.O.C. as well as the important and related items in the risk evaluation system by referring to most insurance companies are finally added to local data analysis as the input items of back-propagation neural network(BPN). After many experiments of parameter sensitivity are carried out, the overall accuracy of training sample can achieve 85.50%. The overall accuracy of test sample can achieve 82.42%. Therefore, one can infer this network can have generalization and confidence. The network parameters can be used for the reference module of the prediction of building claim. The purpose of this study is to establish the items suitable for risk evaluation of local fire insurance via adjusting the weight ratio in each item and test training as well as provide risk evaluators or underwriters can establish an integration and long-term decision-making mode of underwriting, so that the operation of insurance company can make more flexible and improve efficiency.

KEYWORD: *data mining, building fire insurance, neural network, risk evaluation.*

Date of Submission: 26 September 2014 ⟷ Date of Accepted: 25 November 2014

I. INTRODUCTION

Insurers undertake each kind of risk transferred from enterprises. If they can't systematically and efficiently analyze these risk factors to rashly accept insurance or only take account of business to ignore insurance consideration relationship, the confronted risks will be larger, unable to give the stable power to society but causing larger burden to society instead. Owing to fire insurance with the properties of commodity heterogeneity and accumulative danger, risk evaluators or fire insurance underwriters in Taiwan non-life insurance companies recently have generally referred to the methods of evaluating the risk of fire insurance that have been developed abroad. However, these evaluation methods almost evaluate the risk of huge asset targets or large-scale factories. For middle and small fire insurance targets, most non-life insurance companies in Taiwan only deliver their investigators to check on site, and standardized investigation reports are used to evaluate by underwriters, where most of data forms lack complete and decisive data for the reference. As for underwriters how to accept insurance really lack of data persuasion, it is fully decided by experienced underwriters to accept insurance according to the standardized underwriting principle. However, the cases of large fire insurance are processed with coinsurance and reinsurance. The company shares the low self-keeping ratio. On the contrary, the cases of middle and small fire insurance share the higher self-keeping ratio, which is a key to directly influencing the underwriting profit. How to decide insurance acceptance or conditions is judged with professional knowledge by senior and experienced underwriters. Each company could have difficulty to train professional underwriters, and professional knowledge is gone due to leaving office or retirement or wrong underwriting judgment caused by human factors.

For the consideration of factors above-mentioned, the purpose of this study hopes not only can refer to expert experiences but also can analyze and extract the useful information from client data when evaluating the fire risk and making underwriting decision. Therefore, we propose applying data mining for analysis, which the important related factors used in the rate regulations approved by the Ministry of Finance, R.O.C. as well as the important and related items in the risk evaluation system by referring to most insurance companies are finally added to local data analysis as the input items of back-propagation neural(BPN)network. Via adjusting the weight ratio in each item, we expect to establish the items suitable for risk evaluation of local fire insurance and provide risk evaluators or underwriters can establish an integration and long-term decision-making mode of underwriting, so that the operation of insurance company can make more flexible and improve efficiency.

There are many factors to evaluate the risk of fire insurance. If the important factors or variables can searched via related laws or reference theory and actual verifications, they can be more beneficial to construct the network model and training so that the obtained accuracy can be closer to the target value. In other words, the higher factors that influence the risk correlation are involved in the neural networks system, the established models provide more functions to predict. Therefore, this study investigates the factors that currently evaluate the fire insurance risk from three aspects. First of all, we introduce the important factors evaluated in the rate regulations of current fire insurance; secondly, the claim possibility caused by a fire is found via related references as well as measures and countermeasures are provided to prevent a fire; finally, data mining references are looked back to establish the predicting model of neural network.

II. RELATED WORKS

Current analysis of building fire insurance : The rate regulations suitable for current fire insurance are established by experts and scholars authorized by the non-life insurance association, who collect statistic data related to fire insurance market, analyze the cost that the risk of insurance could cause indemnification and payment, consider the expense, make use of actuarial principle of insurance and formulate the rate system in compliance with full, appropriate and fair principle. Each non-life insurance company only accepts any residence and fire insurances inside the border of R.O.C. Residence fire, earthquake and commerce fire insurances shall obey these regulations. Each non-life insurance company must select one of the following policies for insurance target to issue according to operation properties: 1) basic policy of fire and earthquake: applicable objects are the operation properties of residence; 2) commerce fire policy: applicable objects are the operation properties of office, company and store, warehouse, public place and factory. In the coverage of fire insurance, the target contents can be roughly classified as residence, office, company and store, public place, factory and warehouse, where each kind of factory insurance shares the most parts. The classification of fire insurance risk becomes the important basis of loss reduction. The so-called appropriate risk classification makes an insurant to pay the expense of insurance in accordance with the expected cost. If the classification is improper, the difference of expected loss for insurants in the same classification will be extremely excessive.

Recently, non-life insurers have keen competition on commerce fire insurance, so that many companies could run business in the high risk environment. If a claim is unfortunately made, the ability of insurer to pay off could be influenced. After implementing rate liberalization in the future, the market will change more rigorously. Thus, the quality of risk control will become a key to directly influencing the stable operation of insurance company. Confronting the challenge and competition to rate liberalization and how to get balance between fire insurance and risk evaluation must become important issues for non-life insurance companies to think over in the future.

Data Mining : Data mining has been successfully applied in the wide fields, such as production, manufacture, health care, finance, marketing and error detection. Hall (1996) proposed IBM had used Fuzzy model and statistics to analyze the cases of insurance deceit and claim for health-case industry. Australia health-care insurance organizations make use of neural network and statistics to find the cases of deceit and medical resource abuse, which the execution shows the good results. Data mining plays a very important role for the exploration of plentiful data. Berry (1997) thought data mining is to mine the meaningful characteristics or rules, which automatically or semi-automatically explore and analyze data. Kleissner (1998) thought data mining is one kind of new and recycling process to make decision, support and analyze, which can discover knowledge of hidden value from data to provide enterprise professionals for the reference. Therefore, the maximum purpose of data mining is to find the effective information in compliance with complicated inquiry conditions in a lot of data via precise analysis and computation, statistic screening and layer filtering. Problems among application fields dependent on data mining methods along with different tools can obtain the different results.

Data mining is the presentation of knowledge mining. The KDD (knowledge discovery in databases) processes defined by Fayyad et al. (1996) include (1) data selection, (2) pre-processing, (3) data simplification and transforming, (4) data mining, (5) interpreting and evaluating. Han and Kamber (2001) proposed the interactive flow related to the knowledge discovery including a series of data or information transferred to knowledge discovery. The KDD processes mainly consist of the following steps: (1) data cleaning is to eliminate noise and out-of-group value; (2) data integration is to integrate many kinds of data; (3) data selection is to retrieve the data related to analysis and mission from database; (4) data transformation is to transfer the data or unify as appropriate data mining; (5) data mining is to find data characteristics and adopt intelligent methods; (6) data evaluation is to confirm the weight characteristics based on knowledge; (7) knowledge presentation is to visualize the knowledge presentation. In addition, Roiger et al. also proposed (2003) four steps in data mining are (1) goal identification: clearly describe the problems to be solved; (2) data preprocessing: process data noise and clear missing information; (3)data mining: establish supervision or non-supervision learning model; (4) interpretation and evaluation: decide the learning model can be accepted.

For the types of data mining, Kristin (1999) proposed four models: Classification is defined and established according to the properties of analysis object; Estimation: obtain the unknown value with certain property according to existing data; Affinity: decide those related objects to be put together; Clustering: separate more homogeneous clusters from heterogeneous matrix. The classification is the most common type of data mining, and the type analyzed in this study as well. It is always used to handle problem screening and prediction. If some of historical data classified are used to study the characteristics, the prediction will be made to those not classified or new ones in accordance with these characteristics. This kind of data mining includes decision tree, discriminate analysis, Bayesian classifiers, neural networks, memory-based reasoning and fuzzy theory.

The method of neural networks is a common application in data mining, which attempts to learn and reduces the wrong mode to approach to the target value. The reason why neural networks can be widely applied in each field is that linear and non-linear problems have good ability to learn. Neural networks can be supervised learning or unsupervised learning, which is completed via input examples and module weighed training. For the applications of neural networks on insurance, Kitchens (2000) recorded and analyzed 174,000 auto insurance policies provided by an international company, including application form, driver's vehicle record, and loss activity previously done by client and policy premium and loss. Using variables in a policy finds a model of neural networks to predict personal policy loss. Jablonowski (1998) also explained a good risk management plan is started from accident analysis, including the possibility to occur an accident and result evaluation. That study used the method of neural networks to train the results of risk evaluation proposed by experts. It was found neural networks provide good results for risk evaluation automation. Moreover, computerization can traditionally strengthen the problems by paper work to analyze questionnaires and problems related to special risk index. Hence, the advantages of neural networks compared to other classification methods include high accuracy, which can construct nonlinear model; high data tolerance, allowing discreteness, continuation, classification and sequence as input variable. Furthermore, the less modeling results will be influenced by the missing value.

III. RESEARCH METHOD

Case database usually holds a lot of hidden information, which can be used for the reference of intelligent commerce decision-making after analyzed by appropriate methods and tools. Therefore, the classification and prediction are two common kinds of data analysis output mode. This study will apply data mining to construct the predicting model of fire claim by referring to references, which the study architecture is as shown in Fig. 1.

Data Pre-processing : Prior to data modeling and data mining, the quantitative data will be pre-processed, which the processes include data cleaning, data integration and data transformation. The data obtained in this study is based on classification and claim data from past fire insurance as the cleaning source. The cleaning purpose is to remove repeated and less important factors, and investigate objective factors via industrial references as the basis of data integration, where the classification data of fire insurance is based on commerce fire insurance.

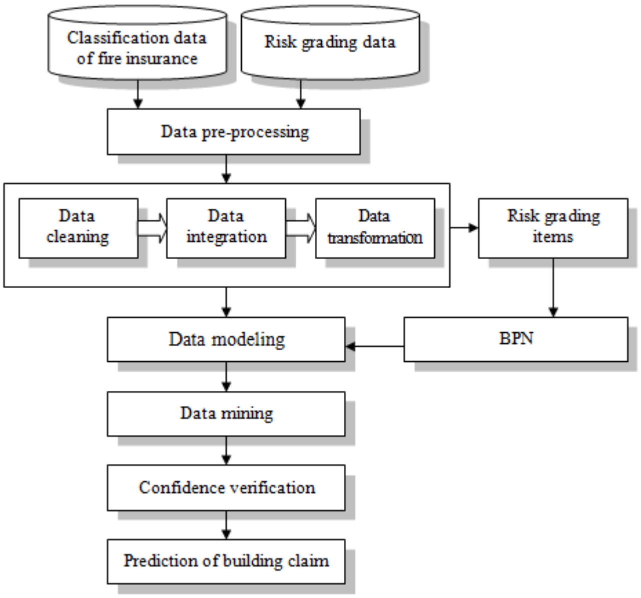

Fig. 1 The framework of Research

There are 689 classifications according to operation properties as a basis of the rate. The rate has 79 grades, excessively classified and complicated. Thus, inappropriate operation properties must be merged or deleted. According to the target, the contents can be roughly classified as residence, office, company and store, public place, factory and warehouse. For the part of risk grading data, the data sources of industrial reference originate from foreign insurance companies, such as Münchener Rück or Munich Re Group, Hanover Re Group, ACE Group, Allianz Group and Marsh Ltd., to compare and sort according to the grade of fire risk and grading system. In order to make data modeling practicable, data transformation will be performed and reduced in accordance with expert interviews, modeling conditions and data mining tools to enhance analysis accuracy. The results obtained in data pre-processing in this phase are as shown in Table 1. According to the common of risk grading item in each company, the appropriate risk grading items are selected for this study object - middle and small fire insurance places as the input items to analyze neural networks of data mining.

Table 1 Risk grading items

1	Building structure
2	Operation property
3	fire-fighting equipment
4	Automatic fire-extinguishing equipment
5	Electric equipment
6	Guard and security system
7	Fire prevention education training
8	Smoking control

Data Modeling : Data mining used in this study is the most widely BPN in neural networks, which the basic principle makes use of Gradient Steepest Descent Method to minimize error function. The collected data are divided as two sets, where a set of training example is provided to neural networks to learn internal mapping rules between input and output variables, namely modifying the weighed value of network; the other set of test example is used for the test classification. In order to find the optimal network model, the claim factors and the possibility of predicted claim are tested and verified for different network setting parameters, including transfer function, training algorithm and network architecture. 16 input variables are used in this study, where the numbering of variable "Operation Property" is as shown in Table 2.

Table 2 Variables in input layer

No.	Variable Name	Numbering
1	*Building grade	【1：A1】【2：A2】【3：B】【4：C】【5：D】
2	Building age (indoor wiring age)	【1：within 3 years】【2：in 3~10 years】【3：in 10~15 years】【4：above 15 years】
3	Distance among buildings	【1：one building】【2：in 10 m】 【3：in 10 m ~15 m】 【4：in 15 m ~20 m】 【5：above 20 m】
4	Operation property	【1: factory】【2: storehouse】【3: public place】【4: textile mill】【5: store】【6: public warehouse】【7: market】【8: office building】【9: parking lot or gas station】【10: open-air storage tank】【11: electrical engine room】
5	Warehouse management	【1：no cargo】【2：good(neat)】【3：normal】 【4：not good(disorderly)】
6	Special danger work	【1：Painting】【2：Galvanization】【3：Dust】 【4：No special danger work】
7	Boiler installation	【1：with boiler installation】【2：without boiler installation】
8	Transformer position	【1：inside the building and only for it】 【2：inside the building and non-wall isolation】 【3：inside the building and wall isolation】 【4：outside the building and open-air】 【5：no transformer】
9	Employee	【1：in 10 persons】【2：in 11-50 persons】 【3：in 51~100 persons】【4：in 101-300 persons】 【5：above 301 persons】
10	Ordinary fire practice and complete prevention plan	【1：well】【2：normal】【3：not well】
11	Fire-fighting equipment	【1：basic equipment】 【2：basic equipment and partial regions with automatic equipment】【3：basic equipment and over 80% regions with automatic equipment】

12	System security and around-the-clock guard	【1：without any one】【2：only have system security】 【3：only have around-the-clock guard】 【4：with both of them】
13	Operating time of equipment	【1：no equipment】【2：in 8 hours per day】 【3：over 8 hours and engine off per day】 【4：operating all day and engine off once every week】【5：operating all day and engine off only if maintenance 】
14	Smoking control	【1：no limit】【2：limit and with smoking area】 【3：no smoking】
15	Position to put waste and handling situation	【1：no waste】【2：good 】【3：normal】【4：not good】
16	Conditions to manage and use inflammables	【1：not use inflammables 】【2：well】【3：normal】 【4：not good】

Note: * Building grade is classified five ranks (A1, A2, B, C, D) based on building's materials from strong (A1) to weak (D).

For variables in output layer, the numbering is decided by: the number of claim is set as 1, and that of normal object is set as 0, as shown in Table 3. In order to avoid divergence, this study will divide the output value of network prediction as 2 sets. If the claim probability lies between 10% and 50%, it is predicted as the non-claim set. If the claim probability lies between 50% and 90%, it is predicted as the claim set. After using neural networks to train, the range of network output will lie between 0 and 1 so that can predict the target of claim probability in each policy. This study establishes the predicted accuracy rate of neural networks = predicted correct number /total number of sample.

Table 3 Variables in output layer

Variable	Code
Claim	【0：claim】【1：non-claim】

Data Mining : When applying BPN, the input vectors must be normalized. Hence, this study adopts minimum - maximum normalization so that the value data will lie within the specific range when transforming. That is normalized data to [0.1, 0.9]. The study samples are the actual data of fire insurants in an insurance company, which covered the claim and non-claim samples during 1999 to 2005. 70% of samples are used as training ones for network. The rest of 30% are used for test ones.During network training, the neuron number (m), learning rate (η) and momentum (α) at the hide layer are investigated for the influence of the classification accuracy rate. This study selects optimal parameters via the sensitivity analysis, where $m =\{4,5,K,20\}$, $\eta = \{0.1,0.2,K,1\}$ and $\alpha = \{0.1,0.2,K,1\}$. In different m, η and α, the program will send back the corresponding classification accuracy rate after training, which means optimal parameters that achieve the maximum classification accuracy rate can be found in 17 × 10 × 10 = 1700 combinations. All programs can be programmed by MATLAB 7 software to construct BPN.

Claim prediction of BPN : This network learning adopts surveillance to train the binding value of network by back-propagation algorithm. For the number of hide layer, we consider the network size, training time and accuracy. The network architecture in this study is classified as three layers (I-H-O): Input Layer (I), Hide Layer (H) and Output Layer (O). The values recommended by Matlab are randomly adopted to set initial weight, which the range lies between [-5, 5]. The initial values of bias in input and hide layer lie between [-10, 10], and the number of learning cycle is set as 2000. The standards to stop training are based on Epoch = 2000 or MSE = 1e-7. If one of conditions meets, weight training will be stopped. The total number of data used in this study is 1690, where 1184 training samples include 564 claim ones and 620 non-claim ones. There are 506 test samples, including 241 claim ones and 265 non-claim ones. For setting network parameters, the value will influence the speed of training convergence and classification accuracy. Therefore, transfer function selects to use S function. The learning rate (briefly called as η) is set as η = 0.1, η = 0.2, …, η = 1, and momentum (briefly called as α) is set as α = 0.1, α = 0.2 , …, α = 1 to observe the influence of different momentum on the classification accuracy rate. In this study, the different neuron number (m), learning rate (η) and momentum (α) at the hide layer are used to train neural networks, where m ={4,5,K,20}, η = {0.1,0.2,K,1} and α = {0.1,0.2,K,1}. In different m, η and α, the program will send back classification accuracy rate of training samples in total of 1700 tests.

Optimal Parameter Selection : The neuron number as variable is used to investigate and achieve the optimal classification accuracy rate. The learning rate and momentum are set as one combination, which the results of the classification rate obtained from each combination are as shown in Fig. 2. The overall classification accuracy rate lies between 0.842 and 0.855, only changing 0.013. Therefore, this study selects the neuron number at the hide layer equal to 7 and 10 to achieve the maximum overall accuracy rate of 85.5% as standard. In order to select the optimal neuron number, there are 63 and 60 miss detected when the neuron number is equal to 7 and 10, as shown in Fig. 3. Therefore, summarizing the observations above-mentioned, this study will adopt the neuron number = 10, the learning rate = 0.6, momentum = 0.6 to carry out example training.

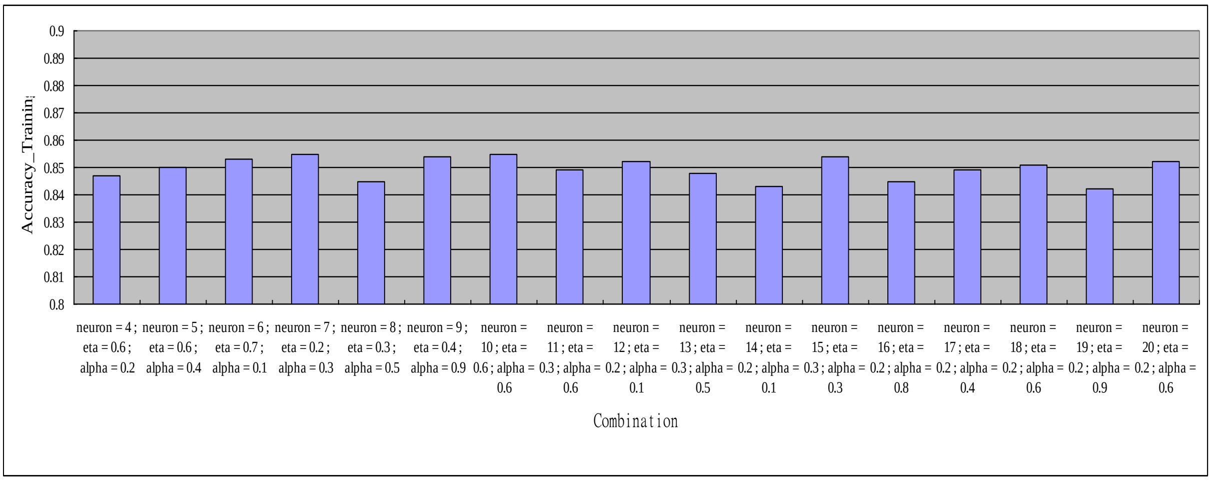

Fig. 2 Classification accuracy rate of every combination

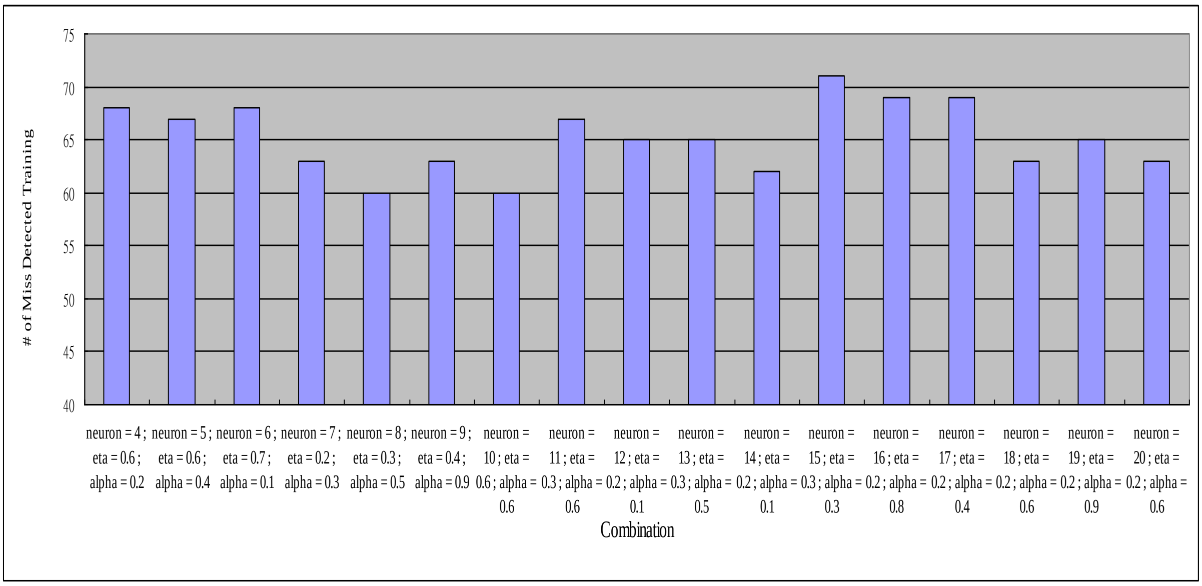

Fig. 3 Miss detected number of every combination

Confidence Verification : Network parameters in test sample adopt the neuron number = 10, learning rate = 0.6 and momentum = 0.6 at the hide layer to carry out training, as shown in Table 4 ~ Table 6. For confusion matrix obtained in each kind of neuron number and the corresponding optimal learning rate and momentum, the overall accuracy rate, the inaccuracy rate and the error rate of miss detected (those actually claiming are classified as non-claim ones.) can have 82.42%, 17.58 % and 4.74 % respectively. We can infer that the results of neural networks should have their confidence, as shown in Table 7.

Table 4 Classification accuracy rate of training samples

Num_neuron = 10						Training Data					
						Momentum					
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	0.1	0.810	0.769	0.782	0.784	0.821	0.817	0.784	0.794	0.815	0.779
	0.2	0.798	0.773	0.823	0.826	0.846	0.793	0.812	0.824	0.834	0.809
	0.3	0.825	0.818	0.820	0.827	0.828	0.785	0.830	0.817	0.812	0.810
	0.4	0.824	0.522	0.712	0.524	0.808	0.753	0.827	0.776	0.722	0.812
Learning Rate	0.5	0.821	0.773	0.633	0.791	0.665	0.817	0.765	0.709	0.828	0.801
	0.6	0.788	0.824	0.639	0.524	0.641	0.855	0.795	0.754	0.731	0.796
	0.7	0.642	0.718	0.686	0.660	0.520	0.657	0.827	0.523	0.667	0.523
	0.8	0.771	0.524	0.650	0.653	0.660	0.524	0.671	0.524	0.508	0.512
	0.9	0.709	0.524	0.524	0.531	0.649	0.624	0.729	0.633	0.582	0.508
	1	0.476	0.476	0.476	0.476	0.524	0.524	0.476	0.476	0.476	0.524

Table 5 Weight of test samples before training

						Initial Weight					
						Hidden					
		H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_10
	X_1	0.812	1.799	1.542	1.018	-0.620	0.109	-1.996	1.495	-0.712	-3.105
	X_2	-3.202	2.247	0.571	-2.890	-0.760	-2.392	-1.202	-0.564	-2.624	0.335
	X_3	-0.040	1.271	3.103	-1.071	-3.479	-2.903	-3.379	3.006	2.591	0.075
	X_4	0.677	-2.157	-1.036	3.334	-3.454	-1.717	-3.914	-3.105	2.936	1.798
	X_5	-1.838	-2.859	-1.511	2.769	1.481	-0.784	-0.053	2.331	-3.706	-2.729
	X_6	-2.235	0.867	-0.169	2.161	2.644	-0.708	-0.265	1.044	-2.137	2.843
I	X_7	-1.826	-2.765	-3.222	2.452	-1.355	-1.438	-2.273	0.013	-2.001	-0.975
N	X_8	-1.756	2.283	-3.208	2.123	-2.359	2.568	-2.299	-2.207	-0.446	2.065
P	X_9	-2.489	0.089	-1.760	-1.186	-1.971	-1.252	0.199	-1.536	-2.049	-1.724
U	X_10	3.364	-1.879	1.104	-0.519	-0.717	-2.252	1.045	3.198	1.625	-1.453
T	X_11	1.236	-2.723	-3.131	1.586	0.813	-3.302	2.572	1.672	-0.062	3.021
	X_12	-0.697	-3.152	-1.934	-0.122	3.219	2.372	-0.419	-0.897	0.714	-2.643
	X_13	2.661	-0.381	-1.265	-0.832	-2.733	1.648	-0.775	-1.132	-1.199	-1.496
	X_14	1.969	-1.438	1.330	-1.131	-0.128	-3.268	-1.758	-2.489	-0.259	-1.801
	X_15	0.574	1.214	2.256	-2.789	0.516	1.252	-2.463	-1.871	-3.043	2.213
	X_16	-2.893	2.105	-1.623	-2.622	0.767	0.563	-2.231	-2.355	1.414	-0.182
	biases	-0.392	0.225	2.680	-2.218	4.428	6.111	8.527	3.495	1.963	-1.354
		Output									
	H_1	0.322									
	H_2	-0.349									
	H_3	-0.723									
	H_4	0.696									
H	H_5	-0.715									
I	H_6	0.509									
D	H_7	-0.170									
D	H_8	0.681									
E	H_9	0.237									
N	H_10	-0.431									
	biases	0.729									

Table 6 Weight of test samples after training

						Hidden					
		H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_10
	X_1	0.718	1.822	1.776	0.466	-0.961	-0.281	-2.082	1.142	-0.785	-3.204
	X_2	-3.205	2.203	1.131	-1.892	-1.122	-2.696	-1.297	-0.920	-2.717	0.344
	X_3	0.034	1.256	3.748	-0.212	-3.757	-3.333	-3.461	2.718	2.542	0.136
	X_4	0.705	-2.039	-0.180	3.717	-3.921	-2.347	-4.048	-3.629	2.886	1.693
	X_5	-1.853	-2.748	-2.614	2.161	1.096	-1.236	-0.163	1.910	-3.763	-2.826
	X_6	-2.267	0.976	-0.306	-0.270	2.285	-1.250	-0.332	0.814	-2.202	2.752
	X_7	-2.006	-2.704	-1.488	2.805	-1.797	-1.838	-2.322	-0.247	-1.972	-1.141
Input	X_8	-1.773	2.296	-3.054	2.153	-2.673	2.309	-2.363	-2.513	-0.461	1.955
	X_9	-2.562	0.174	-1.456	-0.569	-2.389	-1.680	0.099	-1.820	-2.140	-1.821
	X_10	3.246	-1.847	-0.311	-1.534	-1.129	-2.843	0.929	2.873	1.466	-1.635
	X_11	1.149	-2.689	-2.308	1.476	0.601	-3.508	2.526	1.517	-0.106	3.022
	X_12	-0.489	-3.215	-1.326	1.205	2.819	1.909	-0.516	-1.347	0.678	-2.624
	X_13	2.633	-0.223	-1.720	-1.785	-3.122	1.151	-0.877	-1.494	-1.282	-1.691
	X_14	1.848	-1.508	2.029	-1.113	-0.527	-3.607	-1.872	-2.956	-0.358	-1.790
	X_15	0.534	1.321	1.587	-3.059	0.085	0.800	-2.572	-2.245	-3.129	2.054
	X_16	-2.991	2.162	-2.179	-3.019	0.387	0.243	-2.326	-2.602	1.294	-0.332
	biases	-0.404	0.179	3.370	-2.707	3.619	5.240	8.314	2.689	1.755	-1.493
		Output									
	H_1	0.074									
	H_2	0.088									
	H_3	-1.011									
	H_4	-0.936									
Hidden	H_5	0.015									
	H_6	-0.354									
	H_7	0.099									
	H_8	-0.238									
	H_9	0.250									
	H_10	-0.273									
	biases	1.119									

Table 7 Confusion Matrix of test samples

		Actual situation	
		non-claim	claim
Prediction result	non-claim	39.53 %	4.74 %
	claim	12.84 %	42.89 %

Relative importance of network input to output : Garson (1991), Goh (1995), Olden and Jackson (2002), Gevrey et al. (2003) thought the weighed value saved in a network can be further combined after BPN completely learned. This study adopts the relative importance (briefly called as RI) proposed by Garson to analyze each input parameter corresponding to that of an output parameter, which the calculation method is shown in equation (1):

$$RI_k = \frac{\sum_{j}^{m}\left[\frac{|w_{ji}|}{\sum_{l}^{16}|w_{ji}|}|v_{kj}|\right]}{\sum_{i}^{16}\sum_{j}^{m}\left[\frac{|w_{ji}|}{\sum_{l}^{16}|w_{ji}|}|v_{kj}|\right]} \qquad \text{equation (1)}$$

In this study, seen from Table 8, the factors to influence the claim probability are according to importance priority: operation property, transformer position, distance among buildings, warehouse management, fire-fighting equipment, boiler installation, conditions to Manage and Use Inflammables, ordinary fire practice and complete prevention plan, building age, position to put waste and handling situation, smoking control, operating time of equipment, system security and around-the-clock guard, employee, special danger work and building grade, as shown in Fig. 4.

Table 8 Relative importance of network input

Code.	Input	RI
1	Operation property	0.087550
2	Transformer position	0.075105
3	Distance among buildings	0.073727
4	Warehouse management	0.071400
5	Fire-fighting equipment	0.065405
6	Boiler installation	0.064497
7	Conditions to manage and use inflammables	0.061868
8	Ordinary fire practice and complete prevention plan	0.061345
9	Building age (indoor wiring age)	0.061270
10	Position to put waste and handling situation	0.061183
11	Smoking control	0.060551

12	Operating time of equipment	0.056040
13	System security and around-the-clock guard	0.055733
14	Employee	0.051162
15	Special danger work	0.046798
16	Building grade	0.046366

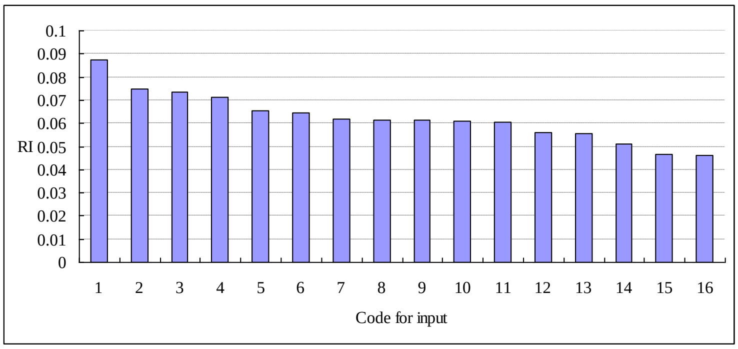

Fig. 4 Relative importance of network input

Sensitivity Analysis of Network Input Unit to Output Unit : If the relationship of the input and output unit approximates to a monotone function when using reverse network, sensitivity of network input to output unit can be analyzed from linked weight. The larger positive sensitivity indicates the larger positive correlation. The larger negative sensitivity indicates the larger negative correlation. In order to understand sensitivity of network input to output unit, we can use sensitivity formula below to calculate. It is assumed i X represents the ith processing unit input at the input layer; j H represents the jth processing unit output at the hide layer; k Y represents the kth processing unit output at the output layer, j net represents an integrated function of the jth processing unit at the hide layer; k net represents an integrated function of the kth processing unit at the output layer. i X to k Y is as shown in equation (2) via sensitivity of the jth processing unit at the hide layer, simplified as equation (3):

$$S_{ki} = \frac{\partial Y_k}{\partial X_i} = \sum_j \frac{\partial Y_k}{\partial net_k} \frac{\partial net_k}{\partial H_j} \frac{\partial H_j}{\partial net_j} \frac{\partial net_j}{\partial X_i} = \sum_j f'(net_k) v_{kj} f'(net_j) w_{ji} \qquad \text{equation (2)}$$

Provided $f'(net_k)$ and $f'(net_j)$ are constant and ignored,

$$S_{ki} = \sum_j^m v_{kj} w_{ji} \qquad \text{equation (3)}$$

After this study adopts sensitivity equation above-mentioned to calculate the linked weight of neural networks, sensitivity can be obtained for the input unit via the processing unit at the hide layer to sum that via the processing unit at all hide layers. The correlation of 16 input units to output result is shown and sorted in Table 9.

Table 9 Sensitivity analysis of input

Input	Sensitivity	Correlation
Operation property	-1.927390	Negative
Transformer position	-0.002230	Negative
Distance among buildings	-2.742610	Negative
Warehouse management	0.051809	Positive
Fire-fighting equipment	1.088305	Positive
Boiler installation	-1.237820	Negative
Conditions to manage and use inflammables	5.723718	Positive
Ordinary fire practice and complete prevention plan	3.040656	Positive
Building age (indoor wiring age)	0.840133	Positive
Position to put waste and handling situation	0.071168	Positive
Smoking control	1.185006	Positive
Operating time of equipment	3.539993	Positive
System security and around-the-clock guard	0.416512	Positive
Employee	2.792739	Positive
Special danger work	-0.571590	Negative
Building grade	-1.733960	Negative

According to Table 9, it can be realized variables more obviously to influence claim probability are conditions to manage and use inflammables, the operating time of equipment, ordinary fire practice and complete prevention plan, employee, distance among buildings, operation property and so on. A variable of "Conditions to Manage and Use Inflammables" to investigate the influence of the claim probability, which display the positive correlation, means the fire risk will depend on the conditions to manage and use inflammables. If the management conditions get worse, the risk to occur a fire will be higher. From Table 10, the conditions to manage and use inflammables are symbolized as acceptable and bad sample, which the claim rate is over 20 %.

Table 10 Claim ratio of conditions to manage and use inflammables

Operation property	Non- Claim	Claim	Total	Claim rate
	(1)	(2)	(1) + (2)	(2) / (3)
【General factory】	501	106	607	0.174629
【Warehouse, stack and outdoor equipment in general factory】	41	5	46	0.108696
【Public place】	47	2	49	0.040816
【Textile factory】	45	1	46	0.021739
【Company, store and shed】	107	1	108	0.009259
【Public	34	0	34	0

warehouse】				
【Market and store】	5	0	5	0
【Government agency, office, folk museum, cram school and so on】	91	0	91	0
【Parking lot, passenger and freight station and gas station】	5	0	5	0
【Outdoor tank, stack and equipment】	1	0	1	0
【Empty house, electrical room and cogeneration equipment】	8	0	8	0

In the variables of the operating time of equipment, they display the positive correlation with the claim probability to mean the fire risk will increase with the operating time of equipment. From Table 11, the claim ratio will be higher if the operating time of equipment is longer.

Table 11 Claim ratio of operating time of equipment

Operating time of equipment	Non-Claim	Claim	Total	Claim rate
	(1)	(2)	(1) + (2)	(2) / (3)
【1：no equipment】	251	3	254	0.011811
【2：in 8 hours per day】	184	6	190	0.031579
【3：over 8 hours and engine off per day】	247	22	269	0.081784
【4：operating all day and engine off once every week】	152	68	220	0.309091
【5：operating all day and engine off only if maintenance】	51	16	67	0.238806

In the variables of ordinary fire practice and complete prevention plan, one can find they display the positive correlation with the claim probability, representing the fire risk will depend on ordinary fire practice and complete prevention plan. If ordinary fire practice and complete prevention plan gets worse, the risk to occur a fire will be higher. From Table 12, ordinary fire practice and complete prevention plan is symbolized as acceptable and bad sample, which the claim ratio is over 10 %.

Table 12 Claim ratio of Ordinary fire practice and complete prevention plan

Ordinary fire practice and complete prevention plan	Non- Claim	Claim	Total	Claim rate
	(1)	(2)	(1) + (2)	(2) / (3)
【1：well】	220	5	225	0.022222
【2：normal】	523	89	612	0.145425
【3：not well】	142	21	163	0.128834

Prediction of building claim : Through many experiments of parameter sensitivity, BPN constructed by the neuron number = 10, the learning rate = 0.6 and momentum = 0.6 can train the maximum overall accuracy rate and the lower weight value relative to the miss detected number as well as the overall accuracy rate of training sample achieves 85.50%; the overall accuracy rate of test sample achieves 82.42 %. Therefore, one can infer this network can have generalization and confidence. The network parameters can be used for the reference module of the prediction of building claim. In sorting the relative importance of input variables, the top three variables are the operation property, transformer position and distance among buildings, relating to their status; the top three variables to more obviously influence the claim probability from the sensitivity analysis include conditions to manage and use inflammables, the operating time of equipment as well as ordinary fire practice and complete prevention plan, relating to human management. However, according to the current regulations of fire insurance rate, the operation properties of a place indeed share the most portions in deciding the rate of fire insurance. According to the actual claim experience, one can realize place management also leads an important place to match with the study results. For the prediction analysis of claim risk for building, the study finds the high-danger operating conditions that can be predicted exist in high-risk buildings and places exist such as improper conditions to use objects and management.

IV. CONCLUSION

This study pays attention to middle and small commerce fire insurance as study object. By way of the analysis of data mining, the important and common items in the risk grading system are used as the input items of BPN and analyzed via past expert experiences and client data to further establish the claim factors of commerce fire prediction as well as evaluate the fire risk and underwriting decision-making.Neural networks using data mining select the neuron number at the hide layer equal to 7 and 10 to achieve the maximum accuracy rate of 85.5% as standard. According to the miss detected number, the weight trained by the neuron number = 10, the learning rate = 0.6 and momentum = 0.6 is employed as test model parameter. The research results can find the correlation of each variable and claim probability includes the positive and negative correlation, where the positive correlation represents the conditions to manage and use inflammables, ordinary fire practice and complete prevention plan, building age, position to put waste and handling situation, smoking control, operating time of equipment, system security and around-the-clock guard, employee, warehouse management, fire-fighting equipment; the negative correlation represents the operation property, transformer position, distance among buildings, boiler installation, special danger work and building grade. From viewpoint of risk management and prediction, the higher safety grade of building conditions (such as operation property, transformer position, and distance among buildings, fire-fighting equipment, boiler installation and so on) plus the fulfillment of full management system will reduce the probability of fire claim.

REFERENCE

[1] Berry, M., and Linoff, G(1997), "Data Mining Techniques for Marketing, Salesand Customer Support", John Wiley and Sons, New York.

[2] C. Kleissner(1998), "Data mining for the enterprise", Proc. of the Thirty-First Hawaii International Conference, Vol. 7, pp. 295-304,1998.

[3] Fayyad, U., Piatetsky, S., G., Smith, P., and Uthurusamy, Reds. (1996), "Advances in Knowledge Discovery and Data Mining".

[4] Garson G. D., (1991), "Interpreting Neural-network Connection Weights," AI Expert, 6(4), pp.46-51.

[5] Gevrey M., Dimopoulosb I. and Leka S., (2003), "Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models," Ecological Modelling, 160(3), pp.249-264.

[6] Goh A. T. C., (1995), "Back-propagation Neural Networks for Modeling Complex Systems," Artificial Intelligence in Engineering, 9(3), PP.143-151.

[7] Hall, C. (1995), "The devil's in the details: Techniques, tools, and applications for database mining and knowledge discovery—Part Ⅱ.", Intelligent Software Strategies, Vol. XI, no.9, p.1-16.

[8] Han, J., and Kamber, M. (2001), "Data mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann, 2006.

[9] Kononenko, I. (1993), "Inductive and Bayesian Learning in Medical Diagnosis", Applied Artificial Intellifence, Vol. 7, p.317-337.

[10] Kristin, R. N., and I. P. Matkovsky (1999), "Using Data Mining Techniques for Fraud Detection.", SAS Institute Inc. and Federal Data Corporation.

[11] M. Berry and G. Linoff (1997), "Data Mining Techniques for marketing, sales, and Customer Support," New York. Wiley Computer Publishing.

[12] Olden J. D. and Jackson D. A. (2002), "Illuminating the "Black Box": a Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks" , Ecological Modelling, 154(1-2), pp.135-150.

[13] Roiger, Richard J., and Geatz, Michael W. (2003), "Data Mining: a Tutorial-Based Primer", USA.