Vol. 5, No. 1, pp. 37-41, June 2006

INTERNATIONAL
JOURNAL OF SAFETY
www.kosos.or.kr

Feature Analysis on Industrial Accidents of Manufacturing Businesses Using QUEST Algorithm

Young Moon Leem*, K.J. Rogers[1] and Young Seob Hwang

Industrial Systems Engineering, Kangnung National University, Kangwon-Do, 210-702 Korea
[1]*Industrial & Manufacturing Systems Engineering, The University of Texas at Arlington, TX 76019 USA*

(Received March 4, 2006; Accepted June 12, 2006)

Abstract : The major objective of the statistical analysis about industrial accidents is to determine the safety factors so that it is possible to prevent or decrease the number of future accidents by educating those who work in a given industrial field in safety management. So far, however, there exists no quantitative method for evaluating danger related to industrial accidents. Therefore, as a method for developing quantitative evaluation technique, this study presents feature analysis of industrial accidents in manufacturing field using QUEST algorithm. In order to analyze features of industrial accidents, a retrospective analysis was performed on 10,536 subjects (10,313 injured people, 223 deaths). The sample for this work was chosen from data related to manufacturing businesses during a three-year period (2002~2004) in Korea. This study used AnswerTree of SPSS and the analysis results enabled us to determine the most important variables that can affect injured people such as the occurrence type, the company size, and the time of occurrence. Also, it was found that the classification system adopted in the present study using QUEST algorithm is quite reliable.

Key words : cross-validation, gains chart, quest, testing data, training data

1. Introduction

According to the report of Ministry of Labor in 2004, rate of industrial accidents in Korea had reduced less than 1% in 1995 after the rate was 4~5% in the 1960s and 2~3% in the 1980s. Finally in 1998, the rate of industrial accidents recorded initially 0.68%. However, since 1998, it has increased from that rate to a recorded level of 0.9% in 2004. Many previous research studies have been focused on the analysis of industrial accidents in order to reduce them. However, most previous research works only provide managerial and educational policies using frequency analysis and comparative analysis based on data from past industrial accidents. This approach is insufficient for effective prevention and analysis of industrial accidents. In order to prevent and analyze industrial accidents, this study performs QUEST algorithm. Most previous studies needed a large amount of time for analysis and it was difficult to find important variables and necessary factors for forecasting and prevention of problems related to industrial accidents.

The major objective of the statistical analysis about industrial accidents is to find out what is the dangerous factor in its own industrial field so that it is possible to prevent or decrease the number of the possible accidents by educating those who work in the field in safety management. However, so far, there is no technique of quantitative evaluation on danger related to industrial accidents. As an endeavor for obtaining technique of quantitative evaluation, this study presents feature analysis of industrial accidents in manufacturing field using QUEST algorithm. In order to analyze data using QUEST algorithm, this study used AnswerTree of SPSS. AnswerTree is a very popular tool, which is easy for data mining and result analysis.

2. Theoretical Background

2.1 Decision Tree

Among data mining techniques, a decision tree is one of the most frequently used methods for knowledge discovery. A decision tree is used to discover rules and

*Corresponding author: ymleem@kangnung.ac.kr relationships by systematically breaking down and subdividing the information contained in data [3,4]. As decision tree is a powerful tool for classification and prediction by finding out the patterns or relationships between data, it is one of the most frequently used data mining methods [2]. In decision tree, there are two main types of trees, which are differentiated according to measurement level of variables. When target variables are discrete type, they make a classification tree, and if they are continuous types, they build a regression tree [1,6]. Nevertheless, all these trees have the same structure.

A decision tree is a non-linear discrimination method, which uses a set of independent variables to split a sample into progressively smaller subgroups. The procedure is iterative at each branch in the tree; it selects the independent variable that has the strongest association with the dependent variable according to a specific criterion [11]. A decision tree features its easy understanding and a simple top-down tree structure where decisions are made at each node. The nodes at the bottom of the resulting tree provide the final outcome, either of a discrete or continuous value.

Among the most popular algorithms (CART, C4.5, CHAID and QUEST) for decision tree, this study uses QUEST algorithm for data analysis and classification.

2.2 QUEST Algorithm

QUEST (Quick Unbiased Efficient Statistical Tree) is a binary-split decision tree algorithm for classification and data mining [10]. QUEST can be used with univariate or linear combination splits. A unique feature is that its attribute selection method has negligible bias. If all the attributes are uninformative with respect to the class attribute, then each has approximately the same change of being selected to split a node.

QUEST is a famous algorithm for least squares fitting of the attitude quaternion of a satellite to vector measurements. QUEST is also a single-point attitude determination algorithm, which yields a closed-form solution of the quaternion and therefore experiences no divergence problems. The divergence problems are sometimes encountered in the use of extended Kalman filter approach [7].

The following indicates procedures for the variable selection algorithm in QUEST.

1) Calculate p-value of ANOVA F-value about continuous predictor variable.

2) Calculate p-value of chi-square test in contingency table of predictor variable and target variable about categorical predictor variable.

3) If smallest p-value in 1 and 2 step is smaller than critical value of modified Bonferroni, smallest variable is selected split variable.

4) If smallest p-value in 1 and 2 step is bigger than critical value of modified Bonferroni, calculate p-value of Levene F- value about continuous predictor variable, and compare with critical value of modified Bonferroni.

3. Results of Data Analysis

3.1 Data

The data used in this study are from Ministry of Commerce, Industry and Energy of the Korean Government (2002.1~2004.12) and are related to occurrence of accidents in Kangwon territory. The sample for this work was chosen from 10,536 data. Table 1 shows the number of occurrence type on industrial accidents in manufacturing field.

The raw data have 32 variables such as type of injured people, type of occurrence, company size, gender, age and days of continuous service, etc. From raw data, this study selected one target variable and eight independent variables which have influence on injured people.

3.2 Tree Analysis

Figure 1 shows the results from an AnswerTree based on the data discussed above. The decision tree totally has 34 nodes. The misclassification rates at the root node and the final node are 2.1166% and 1.6420%, respectively. Therefore, reduction of miscl- assification rate is about 25%. The results of AnswerTree present 5 nodes (node 4, node 7, node 9, node 16 and node 25)

Table 1. The number of injured people according to type of occurrence

Type of occurrence	Type of injured people		
	Injured	Deceased	Total
Entanglement	3,101	28	3,129
Traffic accident	385	18	403
Fall or flying object from scaffolds	897	3	900
Abnormal motion	949	0	949
Slipping	1,435	4	1,439
Cutting	497	0	497
Occupational disease	1,053	82	1,135
Pneumoconiosis	165	50	215
Crash	921	33	954
Collision	910	5	915
Total	10,313	223	10,536

Total	10,313	223	10,536

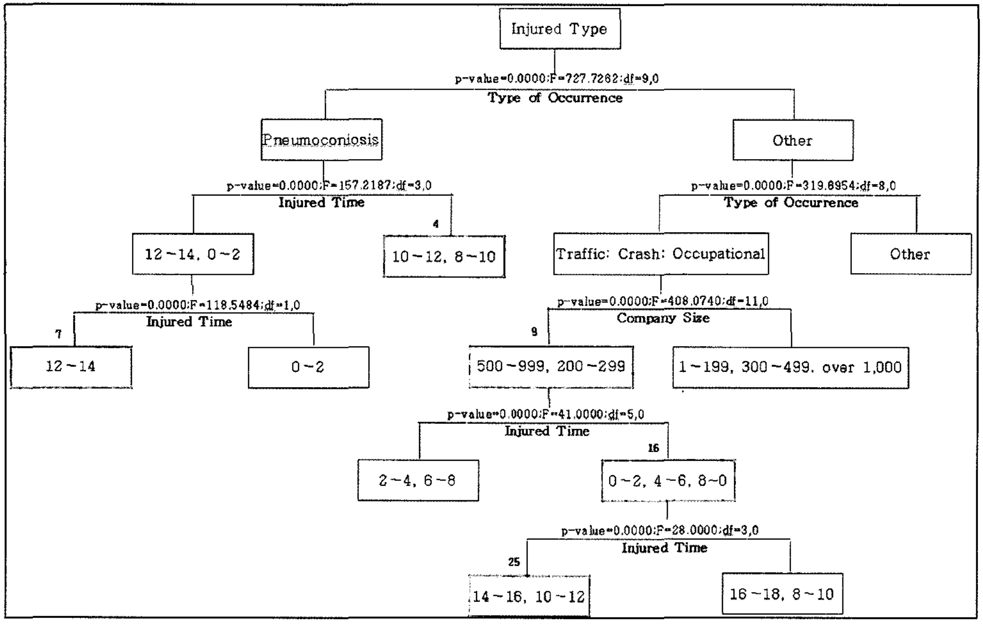
Injured Type
p-value=0.0000;F=727.7262;df=9,0
Type of Occurrence
Pneumoconiosis
Other
p-value=0.0000;F=157.2187;df=3,0
Injured Time
4
12~14, 0~2
10~12, 8~10
p-value=0.0000;F=319.6954;df=8,0
Type of Occurrence
Traffic: Crash: Occupational
Other
p-value=0.0000;F=118.5484;df=1,0
Injured Time
7
12~14
0~2
p-value=0.0000;F=408.0740;df=11,0
Company Size
9
500~999, 200~299
1~199, 300~499, over 1,000
p-value=0.0000;F=41.0000;df=5,0
Injured Time
16
2~4, 6~8
0~2, 4~6, 8~0
p-value=0.0000;F=28.0000;df=3,0
Injured Time
25
14~16, 10~12
16~18, 8~10

Fig. 1. Result of answertree

which frequency of deceased people is highest among overall 34 nodes. The percentage of deceased people of nodes shows that node 4 is 100%, node 7 is 100%, node 9 is 56.10%, node 16 is 67.86%, and node 25 is 100%, respectively.

3.3 Gains Chart Analysis

The gains chart produced by the decision tree can be used for a risk analysis for industrial accidents management. As can be seen in table 2, the gains chart shows which nodes have the highest and lowest proportions of a target category within the node [5]. There are two parts to the gains chart (node-by-node statistics and cumulative statistics). In the gains chart, nodes were sorted by the number of cases in the target category for each node. The first node in the table 2, node 25, contains 11 deceased people cases out of 11 subjects, or 100% deceased people rate. The second node (node 7) contains 20 deceased people cases out of 20 subjects, or 100% deceased people rate. The fourth node (node 26) contains 8 deceased people cases out of 17 subjects, or 47.06% deceased people rate.

The Index score shows how the proportion of deceased people for this particular node compares to the overall proportion of deceased people. For node 25, the index score is about 4,724.66%, indicating that the proportion

Table 2. Gains chart for deceased people by QUEST algorithm

	Node-by-Node				Cumulative			
Node	Node (n)	Res (n)	Gain (%)	Index (%)	Node (%)	Res (%)	Gain (%)	Index (%)
25	11	11	100.00	4,724.66	0.10	4.93	100.00	4,724.66
7	20	20	100.00	4,724.66	0.29	13.90	100.00	4724.66
4	19	19	100.00	4,724.66	0.47	22.42	100.00	4727.66
26	17	8	47.06	2,223.37	0.64	26.01	86.57	4090.01
21	15	5	33.33	1,574.89	0.78	28.25	76.83	3629.92
27	117	36	30.77	1,453.74	0.89	44.39	49.75	2350.46
...	...	...	...	...	...	...	...	...
24	116	0	0.00	0.00	99.20	100.00	2.13	100.80
12	84	0	0.00	0.00	100.00	100.00	2.13	100.00

of respondents for this node is about 47 times the deceased people rate for the overall sample. The cumulative statistics demonstrate how well we do at finding deceased people cases by taking the best segments of the sample. If we take only best node (node 25), we reach 4.93% (respondent percentage) of deceased people cases by targeting only 0.1% (node percentage) of the sample. If we include the next best node as well (node 7), then we get 13.90% of the deceased people cases from only 0.29% of the sample. Including node 4 increases those values to 22.42% of the deceased people cases from 0.47% of the sample. At this stage, we are at the crossover point described above, where we start to see diminishing returns.

The gains chart also provides valuable information about which segments to target and which to avoid. This study bases the decision on the number of prospects this study wants, the desired deceased people rate for the target sample, or the desired proportion of all potential deceased people cases this study wants to contact. In this example, suppose we want an estimated deceased people rate of at least 70%. To achieve this, we would target the first five nodes (25, 7, 4, 26, and 21) with a gain percentage greater than 70%. This segment-specific information can be used for planning a deceased people management program.

3.4 Comparison of Training Data Sample and Testing Data Sample

One model matrix cannot be guaranteed to induce identical results from diverse data, although the matrix is constructed in an exact way from one datum. A tree structure, therefore, should be generalized only after it is confirmed that the structure from one datum is applicable to other data. One datum, so called training data, is used for constructing a decision tree, and the resulted tree is tested with the remaining data, which is called testing data, for the validity [12]. In this study, the ratio of training sample versus testing sample among partition data was controlled as 50% : 50% for a validation test of data division. In the results, training sample are 5,266 data and testing sample are 5,270 data.

A comparison of the accuracy, sensitivity, and specificity for the two samples (training data sample and testing data sample) is shown in Table 3. In short, accuracy means ability on proper classification of tree and sensitivity means ability of declaration it is true when it is true. Also, specificity means ability of declaration it is wrong when it is wrong. As can be seen in table 3, accuracy shows difference of about 0.2% and sensitivity shows difference of about 0.2%. Also, specificity shows difference of about 3.8%. Generally, it is known that classification on data is valid and reliable when the difference value between training data sample and testing data sample is less than 10%. In this study, the differences of three comparisons are less than 4%. This means that classification which was performed in this study is very reliable.

3.5 Cross-Validation

In machine learning methods, such as the decision tree, the classification accuracy, is often predicted by stratified 10-flod cross-validation [8,9,13]. The whole dataset is split into 10 parts, 9 parts of the dataset being dedicated to the training and 1 for the test. The training set is used to learn the algorithm and generate the tree, and the test set is used to estimate the classification parameters of the classifier described above. This procedure is repeated 10 times so that every part of the dataset is used for both training and testing (of course one at each time). Afterwards, the overall accuracy parameters were calculated as means from the evaluation of the individual cross-validation subset. 10 sample folds were selected for cross validation in this study. Table 4 shows the results of 10-fold cross validation. As can be seen in table 4, the difference values between re-substitution and cross-validation were about 0.02%. This means that classification which was performed in this study is very valid.

4. Discussion

The conducted tree using QUEST algorithm is shown in figure 1. Figure 1 indicates that the most salient variable is the type of occurrence.

In case of node 25, the proportion of deceased people is 100%. The types of occurrence in node 25 are traffic accidents, collision, and occupational disease, the size of company is 200~299 peoples and 500~599 peoples, the time which accidents occurred is between 0 and 2

Table 3. Comparison of training data sample and testing data sample

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training	98.4049	98.5460	79.4872
Testing	98.2008	98.3776	75.6098

Table 4. The result of Cross-Validation Test

	Re-Substitution (%)	Cross-Validation (%)
Risk Estimate	1.6419	1.6230

o'clock, between 4 and 6 o'clock, and between 8 and 21 o'clock.

In case of node 7, the proportion of deceased people is 100%. The type of occurrence is pneumoconiosis, the time which accidents occurred is between 12 and 14 o'clock.

Also, in case of node 4, the proportion of deceased people is 100%. The type of occurrence is pneumoconiosis, the time which accidents occurred is between 8 and 16 o'clock and between 10 and 12 o'clock.

In order to test validity of conducted tree, this study compared training data sample and testing data sample. As can be seen in table 3, the differences of training data sample (accuracy : 98.4049%, sensitivity : 98.5460 %, specificity : 79.4872%) and testing data sample (accuracy : 98.2008%, sensitivity : 98.3776%, specificity : 75.6098%) are very small. Therefore, the conducted tree is enough valid. And, as can be seen in table 4, the difference of re-substitution and cross-validation was about 0.02%. Therefore, it means that classification which was performed in this study is very reliable.

5. Conclusion and Future Research

This study presents a feature analysis of industrial accidents in manufacturing field in which there currently exists no technique of quantitative evaluation on danger using QUEST algorithm. According to the analysis results, it is found that the most important variables leading to fatality are the types of occurrence, collision, and occupational disease in manufacturing businesses. Also we found that the type of occurrence (which are traffic accidents, collision, and occupational disease) and the size of the company (divided into 200~299 employees and 500~999 employees) are important factors. Also, when the time which accidents occurred is between 10 and 12 o'clock and between 14 and 16 o'clock, it is found that the rate of occurrence leading to fatality was very high.

After comparing the training data sample and the testing data sample, we know that the conducted tree was valid. Also, the classification adopted in this study is quite reliable since the misclassification rate was very low.

For future research, we plan to select an algorithm with high performance among various algorithms (neural network, LR, CART, C4.5, and CHAID, etc.) based on data of industrial accidents in various businesses.

Acknowledgement

This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government

References

[1] J. Bala, "Using learning to facilitate the evolution of features for recognizing visual concepts", Evoltionary Computation, Vol. 4, pp. 297-312, 1996.

[2] M. J. Berry, G. S, Linoff, Mastering data mining: The art and science of customer relationship management, New York: John Wiley & Sons, 2000.

[3] Y. L. Chen, C. L. Hsu, S. C. Chou, "Constructing a multi-valued and multi-labeled decision tree", Expert Systems with Applications, 25(2), pp. 199-209, 2003.

[4] P. A. Chou, "Optimal partitioning for classification and regression trees". IEEE Transactions on Pattern Analysis and machine Intelligence, 12, 340-354, 1991.

[5] S. H. Ho, S. H. Jee, J. E. Lee, J. S. Park, "Analysis on risk factors for cervical cancer using induction technique", Expert Systems with Applications, 27, pp. 97-105, 2004.

[6] K. J. Hunt, "Classification by induction: application to modeling and control of non-linear dynamical systems", Intelligent Systems Engineering, 24, pp. 231-245, 1993.

[7] Jinlu Kuang, Soonhie Tan, "GPS-based attitude determination of gyrostat satellite by quaternion estimation algorithms", Acta Astronautica, Vol. 51, No.11, pp. 743-759, 2002.

[8] R. Kohavi, Wrappers for Performance Enhancement and Oblivious Decision Graphs, Ph.D. Thesis, University of Stanford, USA, 1995a.

[9] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada, Morgan Kaufmann, San Francisco, CA, USA, 1995b.

[10] W. Y. Loh, Y. S. Shih, "Split selection methods for classification trees", Statistica Sinica, 7, pp. 815-840, 1997.

[11] J. A. Michael, S. L. Gordon, "Data mining technique: for marketing, sales and customer support", New York: Wiley, 1997.

[12] S. Y. Shon,, Tae Hee Moon, "Decision Tree Based on Data Envelopment Analysis for Effective Technology Commercialization", Expert Systems with Applications, 26, pp. 279-284, 2004.

[13] S. M. Weiss, C. A. Kulikowski, Computer systems that learn, Morgan Kaufmann, San Mateo, CA, USA, 1991.