Decision Tree Approach for Factor Analysis of Industrial Accidents
-산업재해의 요인분석을 위한 의사결정나무-

임 영 문 *
Leem Young Moon
황 영 섭 **
Hwang Young Seob

Abstract

의사결정나무 알고리즘은 데이터마이닝 기법중 하나인데 관심이 되는 데이터들에 대하여 분류 및 예측을 가능하게 해준다. 이 기법은 데이터 형태의 특성을 분석할 수 있고 산업재해 형태의 차이점을 찾아내는데 사용될 수 있다. 본 연구에서는 산업재해 데이터의 특성을 파악하고자 C4.5 알고리즘을 사용하였다. 본 연구에서 분석을 위하여 사용된 데이터는 강원도에서 발생한 2년 동안의 산업재해 관련 데이터로서 연구에 적용된 데이터의 수는 19,909개로 구성되어 있다. 본 연구의 목적을 위하여 한 개의 목표변수와 여덟 개의 독립변수가 산업재해 형태에 따라 세분화 되었다. 분석 후 데이터는 222개의 전체 나무가지와 151개의 줄기가지로 분류되었다. 또한 본 연구에서는 재해자들의 위험도 관리와 감소를 위하여 이익도표를 제공하였다.

Keywords: Decision Tree, C4.5 Algorithm, Industrial Accidents, Gains Chart

† 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

* Professor, Dept. of Industrial and Systems Engineering, Kangnung National University

** GTA, Dept. of Industrial and Systems Engineering, Kangnung National University

2006년 5월접수; 2006년 8월 수정본 접수; 2006년 8월 게재확정

1. Introduction

In Korea, the occurrence ratio of industrial accidents has increased since 1994. Related reports have indicated that this is due to inaccurate analysis of industrial accidents. In other words, the policy for accident prevention has a potentially serious problem of structure which cannot eliminate the basic hazardous factors. Many researches have been focused on analyzing data related to industrial accidents. Most research has used frequency analysis, correlation analysis and factor analysis for this purpose. However, those methodologies can only give us insufficient information for prevention of industrial accidents. The main objective of this paper is to provide a gains chart for efficient management of industrial accidents. This paper uses C4.5 algorithm, one of the data mining techniques [2] for the feature analysis.

2. Dataset and Analysis Method

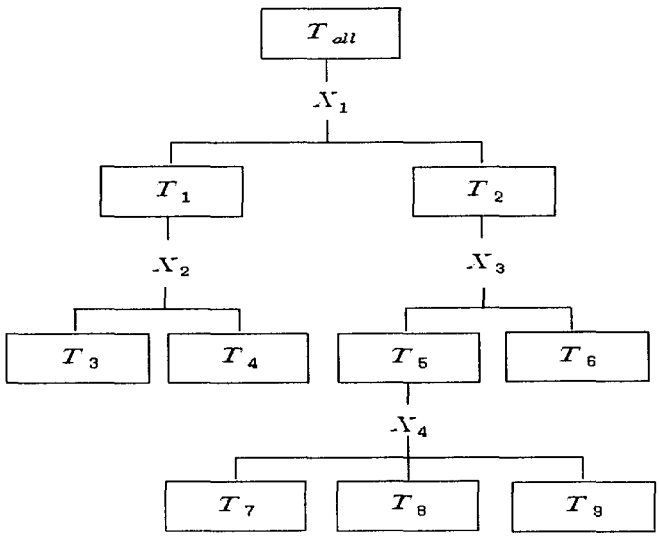
2.1 Dataset

The data used in this paper are from Ministry of Commerce, Industry and Energy of the Korean Government (2003.1.1~2004.12.31) and are related to occurrence of accidents according to the types of business in Kangwon territory.

The business type includes manufacturing, construction, mining, transportation, communication, forestry, finance, insurance, utility, fishery, agriculture, etc. The raw data used in this study have 32 variables such as accident date, type of injured people, type of occurrence, type of business, company size, days of medical treatment, days of hospitalization, age, gender, days of continuous service, and days of labor loss, etc. The dataset consists of 19,909 features obtained by observation during 2 years. For the purpose of this paper, one target value and eight independent variables are detailed by type of industrial accidents. After data cleansing process, nine variables (type of occurrence, type of occupation, type of business, company size, age, gender, days of continuous service, occurrence time and month) were finally used to analyze for the study. The others of dataset consists of various data including receiving an electric shock, accident in a mining field, impossible data for classification, collapse or failure of scaffolds death from drowning, etc.

2.2 Analyzing method

The decision tree is very useful for clustering and classification of data [4] and is very helpful for classification of examples with limited number of class [5][9]. Clustering divides the population into segments by similar characteristics. Tree structure is used to decide whether an object belongs to some class. In general, this process for decision tree starts with small subsets which are selected randomly and then continues to whole data for effectiveness.



<Figure 1> Structure of a decision tree from data set T.

(T_{all} is a data set, $X_1 \sim X_4$ is a decision rule, and $T_1 \sim T_9$ is a divided data set.)

When generating the decision tree, a data set is divided into more detailed observation by following decision rules until one subset corresponds to a certain class. This is very similar to displaying a hard disk directory structure, which is shown in <Figure 1>. Each node has an attribute name and arc from node indicates values which attributes can have. This paper used C4.5 algorithm to construct trees and Enterprise Miner of SAS [11] as a tool for C4.5 algorithm.

2.3 C4.5 algorithm

C4.5 algorithm is an algorithm which is modified and developed by J. Ross Quinlan [10]. ID3 (Interactive Dichotomizer 3, 1986), initial version of C4.5

algorithm has been applied to the field of machine learning [6][7][8]. CART (Classification and Regression Tree) has a tree structure with binary split [1] whereas C4.5 algorithm makes a tree structure with binary split for continuous variables but has a tree structure with multi-way split branches for nominal variables. A process of divide and conquer is the first step to make a decision tree. This process repeats procedure to divide training set until every subset contains one class. In C4.5 algorithm, the concept of information is used. Information theory measures information in bits (symbolized by H). A bit is the amount of information required to decide between two equally likely alternatives. When the probabilities of the various alternatives are equal, the amount of information H in bits is equal to the logarithm, to the base 2, of the number N of such alternatives, or $H = \log_2 N$. With only two alternatives, the information, in bits, is equal to 1.0 (because $\log_2 2 = 1$). When the alternatives are not equally likely, the information conveyed by an event is determined by the following formula:

$$h_i = \log_2 \frac{1}{p_i} \tag{1}$$

Where h_i is the information (in bits) associated with event i, and p_i is the probability of that information of just one such event. The average information H_{av} is computed as follows:

$$H_{av} = \sum_{i=1}^{N} p_i (\log_2 \frac{1}{p_{i=1}}) \tag{2}$$

There are some necessary equations to execute C4.5 algorithm. When we choose one case from S, set of case, the probability that the case belongs to C_j is

$$\frac{freq(C_j, S)}{|S|} \tag{3}$$

Where, $|S|$ is the number of all cases which S contains and freq (C_j, S) is the number of case which belongs to C_j in set S. Now, information carried by the case can be calculated by equation 4.

$$-\log_2 (\frac{freq(C_j, S)}{|S|}) \tag{4}$$

Expected information can be calculated by equation 5 and expected information after T is partitioned to n by X, can be obtained by equation 6.

$$\text{info}(S) = -\sum_{j}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2\left(\frac{freq(C_j, S)}{|S|}\right) \quad (5)$$

$$\text{info}_x(T) = \sum_{i}^{n} \frac{|T_i|}{T} \times \text{info}(T_j) \quad (6)$$

Equation 7 means information obtained from partition by X.

$$Gain(X) = \text{info}(T) - \text{info}_x(T) \quad (7)$$

Split Info means amount of information. It is calculated by equation 8 after T is partitioned into subsets.

$$\text{Split info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{T} \times \log_2\left(\frac{|T_i|}{|T|}\right) \quad (8)$$

Now, we can calculate gain ratio by equation 9.

$$\text{Gain Ratio}(X) = \frac{gain(X)}{splitinfo(X)} \quad (9)$$

The following is main procedure to use above equations: Find split point that can optimize gain ratio according to variables (step 1). Iterate step 1 and compare optimized gain ratio for each variable (step 2). Select the largest variable among them (step 3).

3. Results and Discussion

3.1 Results of tree analysis

The data used in this study were collected for two years but the number of data (19,909) is not sufficient to make specific analysis according to the job type and business type. So this paper tried to find general characteristics on type of injured people. There are 222 total tree nodes and 151 leaf nodes after grouping according to the analysis result. Because the number of classes is large, the number of tree nodes became large. As can be known from the result of tree analysis, accuracy and error rate of the tree were listed in <Table 1>.

<Table 1> Accuracy and error rate of tree

	Accuracy (%)	Error Rate (%)
Training Data	69.21	30.79
Testing Data	68.53	31.47

<Table 1> indicates that the accuracy of the tree is approximately 70% and the accuracy of validation data is similar. Likewise, error rates of training data and validation data are also similar.

3.2 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[3]. 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 192, contains 34 deceased people cases out of 34 subjects, or a deceased people rate of 100%. The third node (node 136) contains 66 deceased people cases out of 67 subjects, or a deceased people rate of 99%.

The Index score shows how the proportion of deceased people for this particular node compares to the overall proportion of deceased people [3]. For node 192, the index score is about 344.8640%, indicating that the proportion of respondents for this node is about 3.44 times the deceased people rate for the overall sample. Also for node 136, the index score is about 341.4154%, indicating that the proportion of respondents for this node is about 3.41 times the deceased people rate for the overall sample.

The gains chart also provides valuable information about which segments to target and which to avoid. The cumulative statistics demonstrate how well we do at finding deceased people cases by taking the best segments of the sample. In this example, suppose we want an estimated deceased people rate of at least 95%.

<Table 2> Gains chart for deceased people by C4.5 Algorithm

	Node-by-Node					*Cumulative*			
Node	Node (n)	Res (n)	Res (%)	Gain (%)	Index (%)	Node (n)	Res (n)	Gain (%)	Index (%)
192	34	34	0.59	100	344.86	34	34	100.00	344.86
195	37	37	0.64	100	344.86	71	71	100.00	344.86
136	67	66	1.15	99	341.42	138	137	99.51	343.19
194	78	77	1.34	99	341.42	216	215	99.33	342.55
197	65	64	1.10	98	337.97	281	278	99.02	341.49
193	27	26	0.45	96	331.07	308	304	98.76	340.58
191	56	52	0.90	93	320.72	364	356	97.87	337.52
111	230	207	3.59	90	310.38	594	563	94.82	327.01
115	214	193	3.34	90	310.38	808	756	93.55	322.61
220	74	65	1.13	88	303.48	882	821	93.05	321.00
114	403	334	5.79	83	286.24	1285	1155	89.98	310.10
119	99	81	1.41	82	282.79	1384	1237	89.35	308.14
...	...	...	...	...	...	...	...	...	...
142	31	0	0	0	0	19883	5695	28.68	100.16
183	26	0	0	0	0	19909	5695	25.64	100.00

To achieve this, we would target the first seven nodes (192, 195, 136, 194, 197, 193, 191). This segment-specific information can be used for planning a deceased people management program.

3.4 Discussion

After generation of trees, we can find that occurrence types of accidents in a mining field were pneumoconiosis, ejection, entanglement, collision, fall or flying object from scaffolds, and traffic accident, etc. In case of pneumoconiosis, 98% of data which belongs to pneumoconiosis was from workers in mining and the number of injured people was very large compared to other injured people. In case of falling, it was common from the men who worked more than ten years in mining. It might be explained by worker's carelessness or perhaps lack of agility due to age. In case of ejection and entanglement, it happened from the men who worked less than ten to twenty years and collision frequently happened to the men who worked less than ten to twenty years. Fall or flying object from scaffolds happened most frequently to the men who worked less than four to five years or more than ten years. Traffic accidents frequently happened in July and August. In case of forestry, common accidents were fall, and flying object and they frequently happened from the men who worked less than six months. Common time period

which accidents happened was between 6 and 10 o'clock, 16 and 18 o'clock. In case of finance and insurance, common accident was falling down and it frequently happened from the men who worked less than four to five years and mainly occurred in May and October. In case of business on transportation, common accidents were falling down, collision, traffic accident and illness. Falling down frequently happened from the men who worked less than four to five years and mainly occurred in January. Collision frequently happened from the men who worked more than twenty years and traffic accident mainly occurred in November. Illness happened from the men who worked for three to five years. In case of manufacturing industry, common accidents were very various such as collision, fall or flying object, ejection, entanglement, abnormal motion, descent, and traffic accident, etc. Falling down frequently happened from the men who worked less than three to four years and ejection and entanglement frequently happened from the men who worked less than one to five years and more than twenty years. Also, accidents mainly happened in January, February and April when days of continuous service are less than six months. Collision frequently happened from the men who worked less than two to four years and ten to twenty years. Abnormal motion and fall or flying object frequently happened from the men who worked more than ten years, less than one to two years, respectively. Illness happened from the men who worked more than one year. Falling down had two types of case. Accidents mainly occurred in January, March, July, August and September when days of continuous service are less than four years to five years. Accidents mainly occurred in February, June, April, and December when days of continuous service are less than one year. In case of construction business, there were all kinds of accidents and the frequency of accidents was very high. Especially, in case of construction business, almost all of accident types happened from unskilled laborers who worked less than six months. This means that there were many accidents due to lack of management and training on unskilled laborers who worked less than six months.

4. Conclusion

In this study, error rate was 30.79% when industrial accidents were analyzed according to type of business. Also we found that error rate became smaller as tree is divided repeatedly. Summarized results of analysis are as follows.

(1) Variables that affect classification of mining are occurrence type, days of continuous service, month of disaster occurrence. In case of pneumoconiosis, 98% of data which belongs to pneumoconiosis was from workers in mining and the number of injured people was very large compared to other injured people. In case of ejection and entanglement, collision, fall and flying object, they frequently happened from the men who worked more than ten years. This might be explained by worker's carelessness, lack of agility in an aging worker, excessive confidence, or ignorance of work process.

(2) In case of business on transportation, manufacturing and mining, the ratio of accidents was high in skilled workers who worked more than three years. This means that there are some important factors to prevent accidents; more considerate management from manager and better understanding of safety and danger by skilled workers. In case of business on transportation like mining, trees were divided by occurrence type of accidents, days of continuous service and months that accidents were happened.

(3) On the contrary, in case of forestry and construction business, more injured people were occurred from unskilled laborers that worked less than six months. In mining and business on transportation, carelessness appears to be the most important factor to generate accidents except the days of continuous service. This indicates that managers should keep considerate management on both of skilled and unskilled workers and should improve work environment and training in order to enhance worker's safety. Variables affected partition of construction business are occurrence type of accidents, days of continuous service, months and time that accidents were happened, and age. In case of forestry, trees were divided by occurrence type of accidents, days of continuous service months and time that accidents were happened.

The gains chart provided in this paper can be used for planning a deceased people management program. In order to analyze data more efficiently, it is necessary to get systematic large dataset and to select a proper algorithm. Future research will focus on comparing accuracy and error rate by various algorithms (CART, C4.5, CHAID, QUEST, logistic regression and neural network).

5. References

[1] Breiman, L. Friedman, J.H, Olshen, R.A. and Stone, C.J. Classification and Regression Trees, Champman and Hall, New York. 1984.

[2] Fayyard, U.M., Piatetsky-Shapiro, G., Smyth, P., From Datamining in Knowledge Discovery. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, pp. 1 -24. 1996.

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

[4] Holder, L.B., Intermediate Decision Trees. Proceedings of the 14th International Conference on Artificial Intelligence, Montreal, Canada. Morgan Kaufmann, San Francisco, USA, 1995. pp. 1056-1062.

[5] Kamber, M., Winstone, L., Gong, W., Cheng, S., Han, J,. Generalization and Decision Tree Induction: Efficient Classification in Data Mining. Proceedings of the International Workshop Issue of Data Engineering (RIDE' 97) Birmingham, UK.1997, pp. 111 -120.

[6] Kubat, M., Holte, R.C., Matwin, S.. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Mach. Learn. 30(2-3), 1998, pp. 195- 215.

[7] McQueen, R.J., Garner, S.R., Nevill-Manning, C.G., Witten, I.H., Applying Machine Learning to Agricultural Data. Comput. Electron. Agric. 12 (4), 1995. pp. 275- 293.

[8] Mitchell, T.M., Machine Learning. McGraw Hill, New York. 1997.

[9] Paul R. Harper, David J. Winsleet Classification trees: A Possible Method for Maternity Risk Grouping, European Journal of Operational Research. 2004.

[10] Quinlan, J.R., C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, USA. 1993.

[11] SAS Inst. Inc., SAS/STAT User's Guide, Version 8, 1st ed. SAS Inst., Inc, Cary, NC, USA. 2002.

저 자 소 개

임 영 문 : 연세대학교에서 학사, 석사학위를 취득하였고, 미국 텍사스주립대학교 산업시스템공학과에서 공학박사를 취득하였으며, 미국 ARRI (Automation and Robotics Research Institute) 연구소에서 선임연구원 및 연구교수를 거쳐 현재는 강릉대학교 산업공학과 부교수로 재직 중이다.

황 영 섭 : 현재 강릉대학교 산업공학과 대학원 박사과정에 재학 중이며 관심분야는 Ubiquitous System, 알고리즘 분석 및 활용 등이다.

저 자 소 개

임 영 문 : 서울시 서초구 서초4동 아크로비스타 C동 910호

황 영 섭 : 강원도 삼척시 원덕읍 월천 1리 6반 269