CHAID Algorithm을 이용한 제조업에서의 산업재해 데이터 분석

Data Analysis of Industrial Accidents in Manufacturing Industries Using CHIAD Algorithm

임영문*, 황영섭 **
Leem Young Moon, Hwang Young Seob

Abstract

The main objective of this study is to provide feature analysis of industrial accidents in manufacturing industries using CHAID algorithm. In this study, data on 10,536 accidents were analyed to create risk groups, including the risk of disease and accident. The sample for this work chosen from data related to manufacturing industries during three years (2002~2004) in Korea. The resulting classification rules have been incorporated into development of a developed database tool to help quantify associated risks and act as an early warning system to individual industrial accident in manufacturing industires.

Keyword : CHAID, Gains Chart, Decision Tree

1. 서론

최근 들어서 산업재해 예방을 위하여 데이터마이닝 기법이 적용되고 있는데 이러한 노력은 산업재해 분야에서 처음 시도되는 독창적인 연구라고 할 수 있다. 산업재해 통계분석의 커다란 목적은 각 산업별로 주요 위험요인을 도출하여 위험요인별 현실적인 예방 대책을 제시함으로써, 산업재해를 줄이거나 예방하는데 있다고 볼 수 있다. 위험한 화학물질을 취급하는 사업장이나 공공시설에서는 위험도를 정량적 수치로 나타내는 소위 정량적 위험성 평가기법을 적용하여 제도화하거나 사업장 스스로 활용하고 있다. 그러나 일반 제조업에서는 아직까지도 정량적 위험성 평가 기법이 개발되어 있지 않은 실정이다. 따라서

† 본 연구는 한국과학재단 목적기초연구(R01-2003-000-00158-0)지원으로 수행되었음.
* 강릉대학교 산업공학과 교수
** 강릉대학교 산업공학과 박사과정

본 논문에서는 의사결정나무 기법의 한 알고리즘인 CHIAD 알고리즘[1]을 이용하여 정량적 위험성 평가기법이 부재한 제조업에서의 산업재해에 대하여 정량적 평가가 가능한 특성 분석을 제시하고자 한다.

2. 연구내용 및 방법

본 논문에서 사용된 데이터 셋은 2002년부터 2004년까지 산업자원부에서 강원도를
대상으로 3년 동안 집계한 제조업에서의 산업 재해자 통계자료로써, 아래 < 표 1 >에서 볼 수 있는 것과 같이 총 10,536개이다.

< 표 1 > 전체 데이터 셋

	부상자(Injured People)	사망자(Deceased People)
Number	10,313	223

이와 같은 데이터 셋에 대하여 의사결정나무와 CHAID 알고리즘을 적용하여 데이터분석 및 특성을 파악하고자 한다.

2.1 Decision Tree

의사결정나무는 데이터마이닝의 분류작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 기법을 의미한다[3]. 이 기법은 새로운 분류값을 예측하기 위하여 이미 만들어진 분류모형(의사결정나무)이 지시하는 바에 따라 레코드의 속성값을 질문하는 방법을 반복적으로 수행한다. 특히 결정적인 질문을 던지게 되면 다른 속성의 값을 묻지 않고도 레코드의 분류값을 정확하게 물을 수 있다.

2.2 CHAID Algorithm

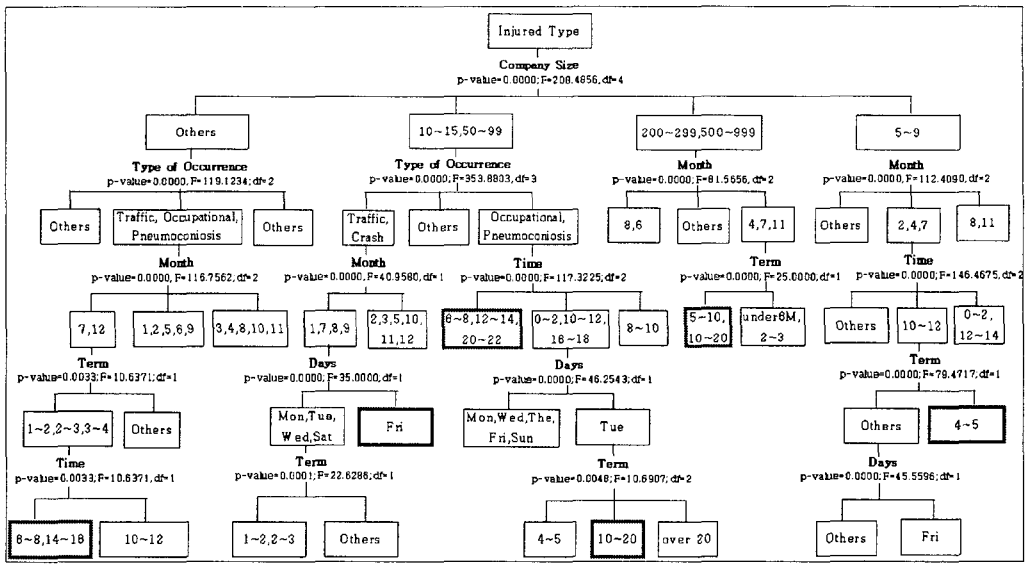
CHAID(Chi-squared Automatic Interaction Detection, Magidson and SPSS INC.(1993), Kass(1980))는 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지 분리(Multiway Split)를 수행하는 알고리즘이다. 여기서 다지 분리란 부모마디에서 자식들이 생성될 때, 2개 이상의 분리가 일어나는 것을 허용함을 의미한다[2][3]. CHAID 알고리즘은 이산형 목표변수에

대해서는 카이제곱 통계량 또는 우도비 카이제곱 통계량(Likelihood Ratio Chi-square Statistic)을 분리기준으로 사용한다. 여기서 목표변수가 순서형 또는 사전 그룹화 된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다.

3. 분석결과

3.1 데이터 분석

앞 절에서 언급한 < 표 1 >의 데이터를 CHAID 알고리즘을 이용하여 수행한 결과 다음 < 그림 1 >과 같은 트리가 형성되었다. 오분류 확률의 감소량을 살펴본 결과 트리를 형성하기 전 2.1166%에서 트리가 형성된 후 0.9507%로써 약 55%의 오분류 확률 감소량을 보였다. 그리고 총 34개의 노드 중 사망 재해자 빈도가 높은 노드는 총 6개의 노드(Node 64, Node 28, Node 57, Node 51, Node 34, 그리고 Node 59)로 나타났다. 사망 재해자 빈도는 모두 100%로 나타났다.



< 그림 1 > 트리 형성 결과

3.2 이익도표 (Gains Chart) 분석

< 표 2 > Gains Chart

	Node-by-Node					*Cumulative*				
Node	Node (n)	Res (n)	Res (%)	Gain (%)	Index (%)	Node (n)	Res (n)	Res (%)	Gain (%)	Index (%)
64	11	11	4.93	100.0000	4724.6637	11	11	4.93	100.0000	4724.6637
28	38	38	17.04	100.0000	4724.6637	49	49	21.97	100.0000	4724.6637
39	16	16	7.17	100.0000	4724.6637	65	65	29.15	100.0000	4724.6637
57	15	15	6.73	100.0000	4724.6637	80	80	32.87	100.0000	4724.6637
31	15	15	6.73	100.0000	4724.6637	95	95	42.60	100.0000	4724.6637
51	12	12	5.38	100.0000	4724.6637	107	107	47.98	100.0000	4724.6637
34	4	4	1.79	100.0000	4724.6637	111	111	49.78	100.0000	4724.6637
59	3	3	1.35	100.0000	4724.6637	114	114	51.12	100.0000	4724.6637
66	18	12	5.38	66.6667	3149.7758	132	126	56.50	95.4546	4509.9062
52	19	12	5.38	63.1579	2983.9981	151	138	61.88	91.3907	4317.9046
...	...	...	...	...	...	...	...	...	...	...
32	10	0	0.00	0.0000	0.0000	10529	223	100.00	2.1180	100.0665
73	7	0	0.00	0.0000	0.0000	10536	223	100.00	2.1166	100.0000

의사결정나무에 의해 생성된 이익도표는 산업재해 관리를 위한 위험분석을 위하여 사용될 수 있다. < 표 2 >에서 볼 수 있는 바와 같이 이익도표는 노드 안에 있는 목표범주에 대하여 최고 비율과 최저 비율을 갖는 노드들에 대한 정보를 보여 준다[4]. 전체 노드 중 사망 재해자에 대하여 가장 큰 영향력을 보이는 노드는 노드 64, 28, 39, 57, 31, 51, 34 그리고 59이다.

Index는 전체 데이터 셋의 사망 데이터 비율과 해당 노드의 사망 데이터 비율이 얼마나 차이를 보이는가를 비교하는 수치이다. 예를 들어, 노드 64의 경우 전체 데이터 셋의 사망 데이터 비율과 비교해 약 47배의 비율을 보인다는 것을 알 수 있다.

또한 이익도표는 의사결정에 매우 유익한 정보를 준다. 물론, 얼마나 많은 세분화 그룹(마디)이 필요할 것인가에 대한 결정이 필요할 수도 있다. 즉, 전체 노드에서 목표 노드의 몇 %를 최종 목표로 설정할 것인가를 정하게 되면, 분석에 필요한 노드들만 세분화하여 분석하고, 예측 할 수 있다는 것이다. 예를 들어, 적어도 90%의 사망 재해율을 평가하고 싶다고 가정하자. 그러면 위의 < 표 2 >의 이익도표에서 Gain (%)가 90%를 넘는 노드 64, 28, 39, 57, 31, 51, 34, 59, 66, 52가 목표 노드가 된다.

전체 데이터 셋에서 분할된 데이터를 바탕으로 모형구축 자료와 모형검증 자료를 비교한 결과 다음 < 표 3 >과 같이 나타났다.

< 표 3 > Training Data Sample과 Testing Data Sample 비교

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training	99.0505	100.0000	53.2710
Testing	98.5579	99.9030	38.7931

4. 특성분석 고찰

QUEST 알고리즘을 이용하여 트리를 형성한 결과 < 그림 1 >에서 볼 수 있듯이 사망 재해자의 빈도가 100%인 노드는 총 8개이고, 그 중 가장 영향력 있는 노드 6개를 조사한 결과 노드 64, 28, 57, 51, 34, 그리고 59였다. 대표적인 예로, 노드 34를 분석해보면, 회사규모가 10~15인, 50인~99인이고, 발생형태가 직업병, 진폐이고, 재해시간이 0시~2시경, 10시~12시경, 16시~18시경이고, 재해요일이 화요일이며, 근속기간이 10년~20년인 그로자의 경우 사망 재해 빈도가 높다는 것을 알 수 있었다.

형성된 트리가 얼마나 타당성을 가지는가를 검증하기 위하여 모형구축 자료와 모형검증 자료 비교를 해 본 결과 < 표 3 >에서 볼 수 있듯이 모형구축 자료의 특성도 확률(정확도:98.4049%, 민감도:98.5460%, 특이도:79.4872%)과 모형검증 자료의 특성도 확률(정확도:98.2008%, 민감도:98.3776%, 특이도:75.6098%)에 차이가 거의 없었으므로 타당하다고 판단할 수 있으며, 형성된 트리를 일반화하기에는 충분하다고 판단되어진다.

5. 결론 및 추후연구

본 논문에서는 의사결정나무 기법의 한 알고리즘인 CHAID를 이용하여 정량적 위험성 평가기법이 부재한 제조업에서의 산업재해에 대한 특성을 분석하였다. 분석 결과 제조업에서 사망 재해자에 영향을 미치는 변수는 총 9개 독립변수 중 6개(회사규모, 발생형태, 재해월, 재해시간, 재해요일, 근속기간)이며, 그 중 가장 큰 영향력을 보이는 변수는 회사규모이다. 그리고 6개 독립변수의 공통된 속성을 분석한 결과, 발생형태의 경우 직업병과 진폐, 재해월의 경우 7월, 재해시간의 경우 6시~8시, 재해요일의 경우 금요일, 근속기간의 경우 4년~5년과 10년~20년인 근로자가 사망재해자 발생 확률이 높은 것을 알 수 있었다. 이러한 트리분석 결과가 얼마나 타당한가를 검증하기 위하여 모형구축 자료와 모형검증 자료를 비교한 결과 본 논문에서 형성된 트리는 충분한 타당성을 보

였다. 또한 교차타당성 검증 결과 역시 구축된 모형과 교차타당성 오분류 확률의 차이가 거의 없으므로 형성된 트리를 일반화하기에 충분한 타당성을 가짐을 보였다.

추후 다양한 업종에 다른 산업재해 데이터를 바탕으로 의사결정나무 알고리즘들, 신경망 알고리즘, 지수회귀분석과 더불어 여러 가지 추론 기법들을 비교하여 가장 효율성이 높고, 분석 결과의 정확성이 높은 알고리즘을 선정하고자 한다.

6. 참고문헌

[1] 송주미, 윤상운, 의사결정나무 분리기준 알고리즘에 관한 연구, 연세대학교 석사학위 논문, 2004, pp.1-19.

[2] 김종현, "우리나라 산업재해 통계를 이용한 제해실태분석과 통계제도의 개선방향", 경일대학교 석사학위논문, pp. 40~60, 1998.

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

[4] 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, pp. 97-105, 2004.