CHAID 알고리즘을 이용한 산업재해 특성분석

- A Feature Analysis of Industrial Accidents Using CHAID Algorithm -

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

Abstract

The main 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 fields for safety tools. However, so far, there is no technique of quantitative evaluation on danger. Almost all previous researches as to industrial accidents have only relied on the frequency analysis such as the analysis of the constituent ratio on accidents. As an application of data mining technique, this paper presents analysis on the efficiency of the CHAID algorithm to classify types of industrial accidents data and thereby identifies potential weak points in accident risk grouping.

Keywords : Industrial Accidents, Decision Tree, AnswerTree, CHAID

1, 서론

2004년 노동부에 따르면 한국의 산업 재해율은 60년대 4~5%에서 80년대 2~3%로 줄어들다가 95년 최초로 1% 미만으로 진입하였다. 하지만 98년 산업 재해율이 0.68%로 사상 최저를 기록한 이후 증가세를 보이기 시작해 지난해 산업 재해율은 98년 이후 최고수준인 0.9%로 나타났고, 일본보다 2.5배나 높은 수준으로 나타나는 등 여전히 선진국에 비교하여 높은 산업 재해율

† 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.
* 강릉대학교 산업시스템공학과 교수
** 강릉대학교 산업시스템공학과 석사과정
2005년 11월 접수; 2005년 12월 수정본 접수; 2005년 12월 게재 확정

을 나타내고 있다. 이러한 결과는 기존 연구 대부분이 산업재해와 관련된 통계자료의 재해 구성 비율 분석[1][2]과 같은 빈도 분석에만 의존한 결과라고 판단되어진다. 이러한 접근 방법은 매우 많은 분석 시간을 필요로 하며, 산업재해와 관련된 재해 예측이나 예방에 있어서 중요한 변수가 어떤 것이며, 중요하지 않은 변수가 어떤 것인지를 알 수 없다. 위험한 화학물질을 취급하는 사업장이나 공공시설에서는 위험도를 정량적 수치로 나타내는 소위 정량적 위험성 평가기법을 적용하여 제도화하거나 사업장 스스로 활용하고 있다. 그러나 일반 제조업이나 건설업 등에서는 아직까지도 정량적 위험성 평가 기법이 개발되어 있지 않은 실정이다. 따라서 본 연구에서는 데이터의 효율적이고 체계적인 분석을 위하여 데이터마이닝 기법 중에서 의사결정나무를 사용하여 재해 관련 데이터들을 CHAID 알고리즘을 이용하여 분석하였으며, 데이터 가공을 위하여 SPSS의 AnswerTree[4]를 사용하였다.

2. CHAID 알고리즘

CHAID(Chi-squared Automatic Interaction Detection, Magidson and SPSS INC.(1993), Kass(1980))[5][6]는 카이제곱-검정(이산형 목표변수)[3] 또는 F-검정(연속형 목표변수) 이용하여 분리기준을 정하여 다지 분리(Multiway Split)를 수행하는 알고리즘이다[7].

2.1 이산형 목표변수에 대한 분리기준

CHAID는 목표변수가 이산형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량(Likelihood Ratio Chi-square Statistic)을 분리기준으로 사용하고, 목표변수가 순서형 또는 사전 그룹화 된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다. 카이 제곱 통계량은 관측도수(Frequency, f_{ij})로 이루어진 r × c 분할표(Contingency Table)로부터 계산된다. 분할표의 구조는 아래 < 표 1 >과 같다.

< 표 1 >. 관측도수로 이루어진 r × c 분할표

목표변수 / 예측변수	범주 1	범주 2	…	범주 C	합 계
범주 1	f_{11}	f_{12}	…	f_{1c}	$f_{1.}$
범주 2	f_{21}	f_{22}	…	f_{2c}	$f_{2.}$
…	…	…	…	…	…
범주 r	f_{r1}	f_{r2}	…	f_{rc}	$f_{r.}$
합 계	$f_{.1}$	$f_{.2}$	…	$f_{.c}$	$f_{..}$

< 표 1 >의 분할표로부터, Pearson의 카이제곱 통계량은 아래 식(1)과 같이 정의 되고,

$$\chi^2=\sum_{i,j}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \quad \cdots (1)$$

우도비 카이제곱 통계량은 아래 식(2)로 정의 된다.

$$\chi^2=2\sum_{i,j}f_{ij}\times\log_e\frac{f_{ij}}{e_{ij}} \quad \cdots (2)$$

이때 두 통계량의 자유도(Degree Of Freedom)는 (r-1)(c-1)로서 동일하다. 여기서 e_{ij}는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(Expected Frequency)를 말하며, 아래 주어진 식(3)과 같이 계산된다.

$$e_{ij}=\frac{f_{i.}\times f_{.j}}{f_{..}} \quad \cdots (3)$$

카이제곱 통계량이 자유도에 비하여 매우 작다는 것은, 예측변수의 각 범주에 따른 목표변수의 분포가 서로 동일하다는 것을 의미하며, 따라서 예측변수가 목표변수의 분류에 영향을 주지 않는다고 결론 내릴 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 p-값으로 표현될 수 있는데, 카이제곱 통계량 값이 자유도에 비해서 작으면 p-값은 커지게 된다. 결국, 분리기준을 카이제곱 통계량 값으로 한다는 것은, p-값이 가장 작은 예측변수와 그 때의 최적 분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

2.2 연속형 목표변수에 대한 분리기준

CHAID는 목표변수가 연속형인 경우에는 두 개 이상의 그룹에 대해서 평균치 차를 검정하는 분산분석표(ANOVA Table: Analysis of Variance Table)의 F 통계량을 분리기준으로 이용한다. y_{ij}를 i번째 예측변수의 범주에 속하는 j번째 관측치의 목표변수의 값이라고 하고, $\overline{A}_i$를 i번째 범주의 평균, $\overline{y}$를 전체 평균이라고 할 때, 분산분석표는 아래 < 표 2 >와 같이 만들어진다. 여기서 r은 예측변수의 범주 수, n_i는 i번째 범주의 관측개체 수, n은 전체 관측치수를 말한다. 이렇게 계산된 F 통계량은 자유도 (r-1, n-r)인 F-분포를 따르는 것으로 알려져 있다. F 통계량이 자유도에 비해서 매우 작다는 것은 예측변수의 각 범주에 따른 목표변수의 평균치 차가 존재하지 않는다는 것을 의미하며, 따라서 예측변수가 목표변수의 예측에 영향을 주지 않는다고 결론 내릴 수 있다. 카이제곱 통계량과 마찬가지로 자유도에 대한 F 통계량이 크고 작음은 p-값으로 표현될 수 있는데 F 통계량이 자유도에 비해서 작으면 p-값은 커지게 된다. CHAID에서는 이와 같이 계산된 F 통계량의 p-값을 기준으로 명목형 목표변수인 경우와 유사하게 병합과 분리를 계속하여, p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디가 형성된다.

< 표 2 > 분산 분석표

요인	자유도	평방합	평균평방	분산비
예측변수	r-1	$SST=\sum_{i=1}^{r} n_i(\overline{y_i}-\overline{y})^2$	MST = SST/(r-1)	$F=\frac{MST}{MSE}$
오차	n-r	$SSE=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2$	MSE = SSE/(n-r)	
전체	n-1	$SST=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$		

3. 데이터 셋

본 연구에서 사용된 데이터 셋은 2002년부터 2004년까지 산업자원부에서 강원도를 대상으로 집계한 업종별 산업 재해자 통계자료이다. 이 데이터들의 특성은 업종에 따른 예측변수는 사업장명, 재해자명, 재해일자, 재해자 구분, 발생형태, 규모, 진료일수, 입원일수, 통원일수, 재가 일수, 공사규모, 연령, 성별, 요양기간, 근속기간, 재해월, 재해시간 그리고 근로손실일수로 총 18개이다. 하지만 여기서 개인 신상보호를 위한 사업장명과 재해자명을 제외하고, 또한 결측치를 다수 포함하고 있는 예측변수를 제외하면 발생형태, 규모, 연령, 성별, 근속기간, 재해월, 재해요일 그리고 재해시간으로 함축된다. 아래 < 표 3 >에서 볼 수 있듯이 재해자 형태와 관련된 데이터는 총 67,278개로서, 부상은 60,249개이고, 사망은 7,029개이다.

< 표 3 > 업종에 따른 재해자 분포

업 종	재해자 형태		합 계
	부 상 자	사 망 자	
건설업	18,975	599	19,574
광업	12,903	5,459	18,362
금융보험업	450	24	474
기타산업	10,880	427	11,307
제조업	10,313	223	10,536
농업	269	3	272
어업	131	7	138
운수·보관업	2,946	203	3,149
임업	3,249	70	3,319
전기·상수도업	133	14	147
합 계	60,249	7,029	67,278

4. 분석 결과

AnswerTree를 실행하기 위해서는 분리 기준을 정하여야 한다. 본 연구에서는 분리 기준을 정함에 있어서 업종의 빈도수가 너무 낮아서 제외된 4가지 업종 중에서 가장 높은 빈도수를 갖는 업종의 수를 최소 부모 노드의 경우의 수(금융보험업; 474)로 정하였으며, 제외된 4가지 업종 중에서 가장 낮은 빈도수를 갖는 업종의 수를 최소 자식 노드의 경우의 수(어업; 138)로 정하였다. 또한 최대 트리의 깊이는 5로 정하였다. 그 이유는 트리가 너무 깊어지면 분석이 어려워지며, 많은 시간을 필요로 하기 때문이다. 목표변수는 재해자 구분이었으며, 예측변수는 성별, 연령, 규모, 재해월, 재해시간, 재해요일, 근속기간, 발생형태, 업종 총 9개였다.

4.1 오분류 확률 분석

오분류 확률을 평가해 본 결과 표 4와 같이 나타났다. 최초 부모 마디의 오분류 확률인 10.5378%에서 7.4132%로 오분류 확률이 3%p 정도의 감소량으로써, 1/3 가량 줄어들어 매우 높은 감소량을 보였다.

< 표 4 > 오분류 확률

Misclassification Rate (%)	
Root Node	Final Node
10.5378	7.4132

4.2 모형구축 자료(Training Data Set)와 모형검증 자료(Testing Data Set) 비교 분석

하나의 자료에 대해서 적절한 방법을 적용하여 정확하게 모형을 구축하였다고 할지라도, 이러한 결과가 다른 자료에서도 동일한 결과를 얻을 수 있음을 보장해 주는 것은 아니다. 따라서 하나의 자료로부터 구축된 나무구조가 다른 자료에 대해서도 잘 적용되는가를 확인한 후 그 나무구조를 그대로 일반화하여야 할 것이다[4]. 본 연구에서는 모형구축 자료와 모형검증 자료 비교분석을 위하여 데이터 분할을 각각 50%로 할당하였다. 분할된 데이터는 < 표 5 >와 같고, < 표 6 >에서 보는 바와 같이 모형구축 자료와 특성도와 모형 검증 자료의 특성도 확률의 차이가 거의 없으므로, CHAID 알고리즘으로 형성된 트리는 충분한 타당성을 보이며 일반화하기에 충분하다. 아래 < 표 5 >는 타당성 검증을 위하여 분할된 데이터 셋이다.

< 표 5 > 분할 데이터

	Training Data Set		Testing Data Set	
	%	n	%	n
부 상	89.37	29,794	89.56	29,472
사 망	10.63	3,544	10.44	3,347
합 계	100.00	33,338	100.00	32,909

분할된 데이터 셋으로 AnswerTee를 실행한 결과 아래의 < 표 6 >과 같은 오분류 행렬이 생성되었고, 모형구축 자료와 모형검증 자료의 특성도를 산출한 결과는 < 표 7 >과 같다.

아래 < 표 7 >에서 보는 바와 같이, 정확도(Accuracy)는 "트리가 얼마나 잘 분리되었는가에 대한 능력"으로써, 모형구축 자료와 모형검증 자료 모두 높은 확률을 보이며, 두 확률 값의 차이가 거의 없는 것을 알 수 있다. 민감도(Sensitivity)는 "맞는 것을 맞다라고 선언하는 능력"으로써, 모형검증 자료와 모형구축 자료 모두 높은 확률을 보이며, 역시 두 확률 값의 차이가 거의 없다. 특이도(Specificity)는 "아닌 것을 아니라고 선언하는 능력"으로써, 모형구축 자료와 모형검증 자료 모두 비교적 높은 확률을 보이며, 두 확률 값의 차이가 거의 없으므로, CHAID를 이용하여 형성된 트리는 충분한 타당성을 보인다.

< 표 6 > 모형구축 자료와 모형검증 자료의 오분류 행렬

Misclassification Matrix					Misclassification Matrix				
		Actual Category					Actual Category		
		사망	부상	Total			사망	부상	Total
Predicted Category	사망	1,710	275	1,985	Predicted Category	사망	1,583	258	1,841
	부상	1,834	29,519	31,353		부상	1,854	29,214	31,068
	Total	3,544	29,794	33,338		Total	3,437	29,472	32,909
		Training Sample					Testing Sample		
Risk Estimate		0.063261			Risk Estimate		0.064177		

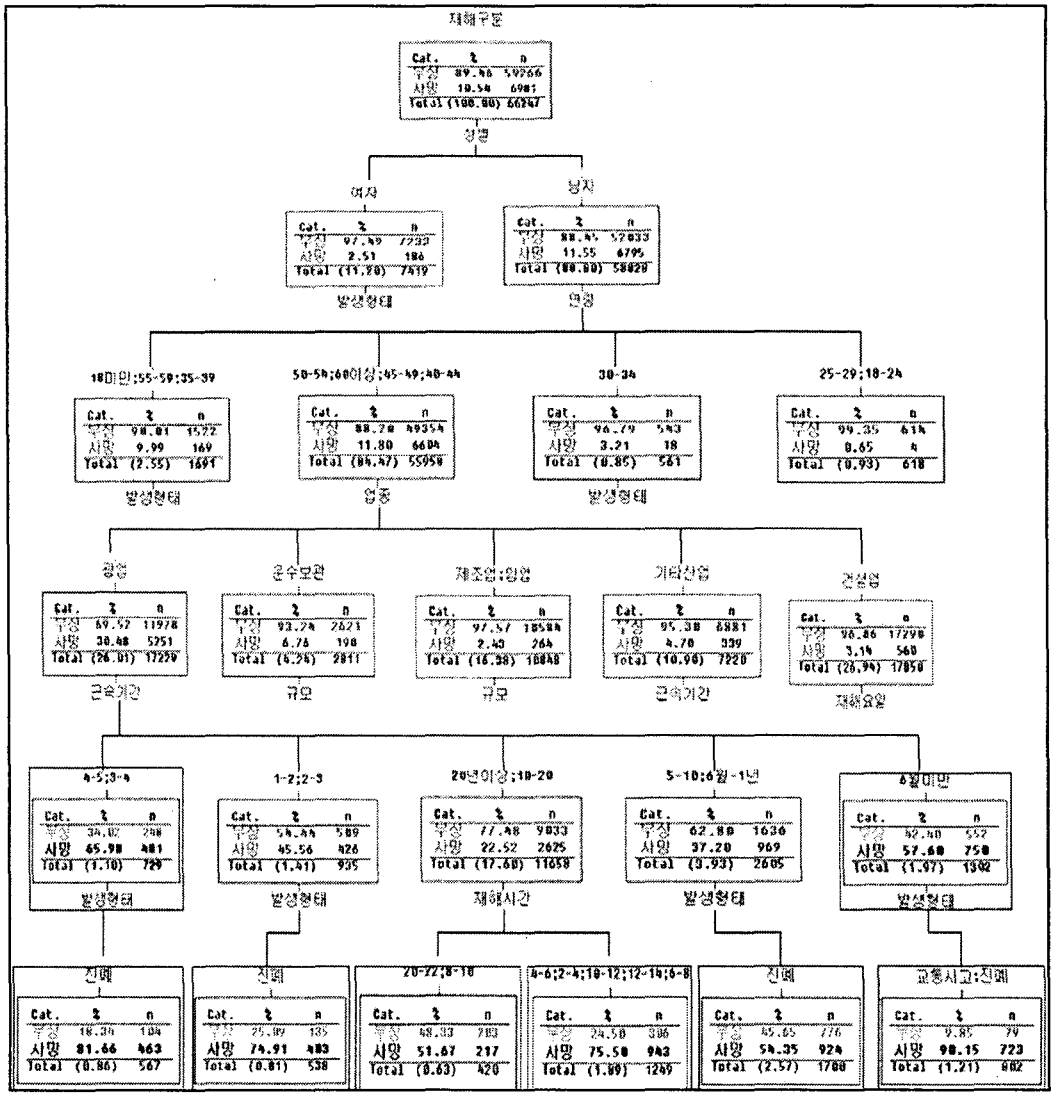
< 표 7 > 모형구축 자료와 모형검증 자료의 특성도

Training Data Set (%)			Testing Data Set (%)		
Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
93.67389	94.15048	86.14610	93.58230	94.03244	85.98588

4.3 트리 분석

트리 형성 결과 5 Depth로 총 97개의 노드로 구성되어 있으며, 끝 노드는 66개로 나타났다. 아래 < 그림 1 >은 총 노드에서 사망의 빈도가 높은 노드들을 정리한 것이다. 총 10개 업종 중에서 사망 재해자 빈도가 가장 높았던 업종은 광업이었으며, 다른 업종들은 사망 재해자의 빈도가 높지는 않았다. 이러한 결과는 강원도라는 지역적 특성 때문에 광업 근로자가 많은 이유라고 볼 수 있다. < 그림 1 >에서 볼 수 있는 것과 같이, 재해자 구분에 있어서 가장 큰 영향을 미치는 예측변수는 성별로 남자의 경우가 사망 재해자가 높다. 성별이 남자이고 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상인 경우 총 55,958명중 사망 재해자의 빈도는 11.8%로 높은 것을 알 수 있다. 성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하는 근로자의 사망 재해자 빈도(30.48%)가 다른 업종에 비해 높은 것을 알 수 있다. 그리고 성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하고

< 그림 1 > 형성된 트리의 부분 줄기



근속기간이 3년~4년, 4년~5년인 근로자는 진폐로 인한 사망 발생 빈도가 81.66%로 매우 높게 나타났다. 또한 성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하고, 근속기간이 6개월 미만인 근로자는 진폐와 교통사고로 인한 사망 재해자 빈도는 90.15%로 가장 높은 사망 빈도율을 보였다.

성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하고, 근속기간이 1년~2년, 2년~3년일 경우 진폐의 발생률이 높아지고, 사망자는 74.91%의 빈도를 보였고, 성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하고, 근속기간이 10년~20년, 20년 이상인 경우는 재해시간에 따라 그 사망 재해자 빈도율이 51.67%(재해시간이 08시~10시경, 20시~22시경일 경우)와 75.50%(재

해시간이 02시~04시경, 04시~06시경, 06시~08시경, 10시~12시경, 12시~14시경일 경우)로 나타났다.

성별이 남자이고, 연령이 40세~44세, 45세~49세, 50~54세, 60세 이상이며, 광업에 종사하고, 근속기간이 6월~1년 미만, 5년~10년인 근로자의 경우는 진폐로 인하여 사망한 빈도가 54.35%로써, 높은 빈도율을 보였다.

5. 결론 및 추후 연구 사항

AnswerTree 실행 결과, 사망 재해자와 높은 상관성을 가지는 변수는 총 9개 중에서 성별, 연령, 업종, 근속기간, 발생형태, 재해시간이었다. 그리고 CHAID 알고리즘을 적용하여 트리를 형성한 결과 오분류 확률의 감소량은 30%로 높은 감소량을 보였다. 또한 모형구축 자료와 모형검증 자료의 비교 결과 충분한 타당성을 보였다. 트리 분석 결과 업종 중에서 광업에서 사망 재해자의 빈도가 가장 높았으며, 근속기간에 따라 6개월 미만일 경우는 교통사고와 진폐로 인한 사망 재해자가 많았고, 3년에서 5년일 경우는 진폐로 인한 사망 재해자가 많은 것으로 나타났다. 본 연구에서 분석된 데이터들은 강원도라는 특정 지역에서 조사된 재해자 데이터였기에 지역적 특성에 의해 광업에서의 사망 재해자의 빈도가 높게 나타났지만, 전국에서 조사된 재해자 데이터를 사용하여 분석을 하였을 경우에는 차이를 보일 것으로 판단된다. 추후 전국에서 조사된 방대한 재해자 데이터들을 중심으로 대표 업종별 산업재해 예방에 적합한 최적 알고리즘 선정에 대하여 연구하고자 한다.

6. 참 고 문 헌

[1] 김종현, 우리나라 산업재해 통계를 이용한 재해실태분석과 통계제도의 개선 방향, 경일대학교 석사학위논문, pp.40~60, 1998.

[2] 노동부, 산업재해현황분석, 2004.

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

[4] 최종후, 한상태, 강현철, 김은석, (AnswerTree를 이용한) 데이터마이닝 의사결정나무 분석, 고려 정보 산업, pp. 17~74, 1998.

[5] Kass, G., "An exploratory technique for investigating large quantities of categorical data", Applied Statistics. 29:2, pp. 119~129, 1980.

[6] Magidson, J. and SPSS Inc., *SPSS for Windows CHAID Release 6.0*, Chicago, IL : SPSS Inc., 1993.

[7] R. Srikant, R. Agrawal, "Mining generalized association rules, Future Generation Comput. Systems" 13, pp. 161~180, 1997.

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 재학 중이며 관심분야는 Ubiquitous System, 알고리즘 분석 및 활용 등이다.