QUEST 알고리즘을 이용한 제조업에서의 산업재해 특성 분석

- Feature Analysis of Industrial Accidents in Manufacturing Business Using QUEST Algorithm -

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

Abstract

So far, there is no technique of quantitative evaluation on danger related to industrial accidents. Therefore, 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 feature of industrial accidents, a retrospective analysis was performed in 10,536 subjects (10,313 injured people, 223 deaths). The sample for this work chosen from data related to manufacturing businesses during three years (2002~2004) in Korea. The analysis results were very informative since those enable us to know the most important variables such as occurrence type, company size, and occurrence time which can affect injured people. Also, it is found that classification using QUEST algorithm which was performed in this study is very reliable.

Keyword : Gains Chart, QUEST, Cross-Validation, Training Data, Testing Data.

1. 서 론

정부는 산업재해 취약 부분에 행정역량을 집중하는 등 '산업재해 예방 제 2차 5개년 계획'을 수립하고 2005년부터 본격 추진키로 하였다. 2004년 노동부에 따르면 한국의

† 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.
* 강릉대학교 산업공학과 교수
** 강릉대학교 산업공학과 박사과정
2006년 2월 접수 2006년 3월 수정본 접수 2006년 4월 게재 확정

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 나타났다. 산업재해와 관련된 기존 연구 대부분은 통계자료의 재해 구성 비율 분석과 같은 빈도분석[2][3][11]에만 의존하였고, 최근에서야 산업재해 예방을 위하여 데이터마이닝 기법이 적용되고 있는데 이러한 노력은 산업재해 분야에서 처음 시도되는 독창적인 연구라고 할 수 있다. 기존 연구의 경우 매우 많은 분석 시간을 필요로 하며, 산업재해와 관련된 재해 예측이나 예방에 있어서 중요한 변수가 어떤 것이며, 중요하지 않은 변수가 어떤 것인지를 알 수 없었다.

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 논문에서는 의사결정나무 기법의 한 알고리즘인 QUEST[1][7][8]를 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 산업 재해자 통계자료이다. 아래 < 표 1 >에서 보는 바와 같이 총 10,536개의 데이터 중 부상자(Injured People) 데이터는 10,313개이고, 사망자(Deceased People) 데이터는 233개이다.

< 표 1 > 전체 데이터 셋

	부상자(Injured People)	사망자(Deceased People)
Number	10,313	233

이와 같은 데이터들에 대하여 의사결정나무와 QUEST 알고리즘을 적용하여 데이터 분석 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[4][5][6]. 이 기법은 새로운 분류값을 예측하기 위하여 이미 만들어진 분류모형(의사결정나무)이 지시하는 바에 따라 레코드의 속성값을 질문하는 방법을 반복적으로 수행한다[9]. 특히 결정적인 질문을 던지게 되면 다른 속성의 값을 묻지 않고도 레코드의 분류값을 정확하게 물을 수 있다. 따라서 의사결정나무에서 가장 중요한 핵심은 레코드를 분류하고 예측할 수 있는 나무모형을 얼마나 잘 만드느냐는 것이 된다.

2.2 QUEST Algorithm

QUEST(Quick Unbiased Efficient Statistical Tree)[1][7][8]는 C4.5와 마찬가지로 명목형 목표변수에 대해서만 분석을 수행할 수 있다. 분리방법은 예측변수의 측도에 따라서 서로 다른 분리기준을 사용하여 이지분리를 하는데 분리변수의 선택과 선택된 분리변수에서 분리점의 선택으로 나누어 실행된다. 분리변수의 선택은 예측변수가 순서형 또는 연속형인 경우에는 ANOVA F-검정 또는 Levene의 검정을 사용하며, 예측변수가 명목형인 경우에는 Pearson의 카이제곱 검정을 사용하여 가장 작은 유의확률에 대응되는 변수를 분리변수로 선택한다. QUEST에서 변수선택 알고리즘을 간단히 요약하면 다음과 같다[1].

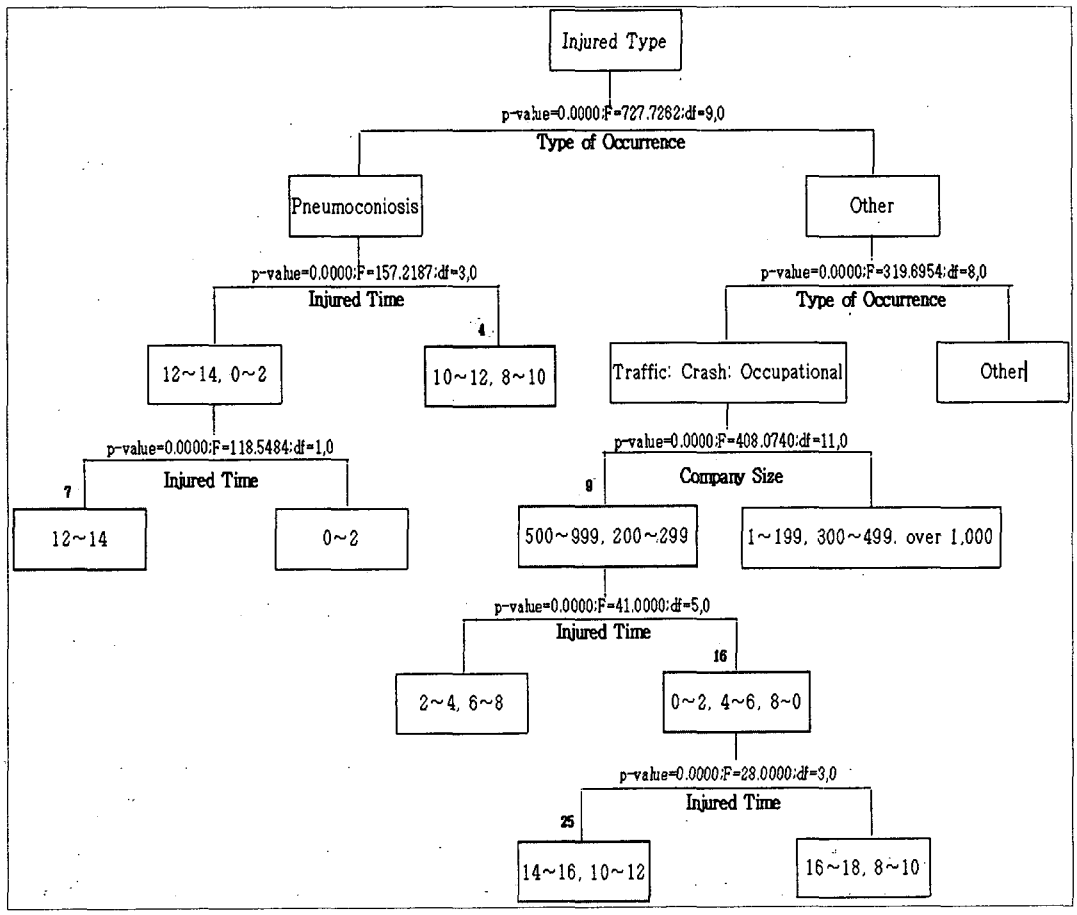
① 순서형 예측변수에 대해서는, ANOVA F-검정의 유의확률을 계산한다.
② 범주형 예측변수에 대해서는, 예측변수와 목표변수의 분할표에서 카이제곱 검정의 유의확률을 계산한다.
③ 1, 2 단계에서 가장 작은 유의확률이 Bonferroni 수정 임계값보다 작으면 그에 대응되는 변수를 분리변수로 선택한다.
④ 그렇지 않으면, 순서형 예측변수에 대하여 Levene F-검정의 유의확률을 계산하고 Bonferroni 수정 임계값과 비교한다. 유의확률이 임계값보다 작으면 대응되는 변수를 선택하고, 아니면 1, 2 단계에서 구한 가장 작은 유의확률에 대응되는 변수를 분리변수로 선택한다.

목표변수의 범주가 3개 이상인 경우에는 CART의 Twoing 기준에서와 유사하게 2-평균 군집분석(Two-means clustering)을 수행하여 두 개의 그룹을 만든 후 분석을 수행한다. 또한 각 예측변수의 최적분리를 찾기 위하여 2차 판별분석(Quadratic discriminant analysis)을 수행하고, 목표변수를 가장 잘 분류하는 예측변수의 최적분리를 이용하여 자식마디를 형성한다. QUEST는 관측치의 수가 많거나 복잡한 자료에 대해서는 효율적이지만, 이 방법 역시 모든 관점에서 다른 알고리즘보다 항상 좋은 결과를 주는 것은 아니다.

3. 분석결과

3.1 데이터 분석

앞 절에서 언급한 < 표 1 >의 데이터를 QUEST 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 1.6420%로써 약 25%의 오분류 확률 감소량을 보였다. 그리고 총 34개의 노드 중 사망 재해자 빈도가 높은 노드는 총 5개의 노드(Node 4, Node 7, Node 9, Node 16, 그리고 Node 25)로 나타났다. 각각 사망 재해자 빈도는 Node 4가 100%, Node 7이 100%, Node 9가 56.10%, Node 16이 67.86% 그리고 Node 25가 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 (%)
25	11	11	4.93	100.00	4,724.66	11	11	4.93	100.00	4,727.66
7	20	20	8.97	100.00	4,724.66	31	31	13.90	100.00	4,724.66
4	19	19	8.52	100.00	4,724.66	50	50	22.42	100.00	4,724.66
26	17	8	3.59	47.06	2,223.37	67	58	26.01	86.57	4,090.01
21	15	5	2.24	33.33	1,574.89	82	63	28.25	76.83	3,629.92
27	117	36	16.14	30.77	1,453.74	199	99	44.39	49.75	2,350.46
...	...	...	...	...	...	...	...	...	...	...
24	116	0	0.00	0.00	0.00	10,452	223	100.00	2.13	100.80
12	84	0	0.00	0.00	0.00	10,536	223	100.00	2.12	100.00

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[8]. 전체 노드 중 사망 재해자에 대하여 가장 큰 영향력을 보이는 노드는 노드 25, 노드 7, 그리고 노드 4이다. 노드 25에서는 총 11개의 데이터 중 응답이 11개로써 100%의 평균 응답률을 보이고, 노드 7과 노드 4에서도 100%의 평균 응답률을 보였다.

Index는 전체 데이터 셋의 사망 데이터 비율과 해당 노드의 사망 데이터 비율이 얼마나 차이를 보이는가를 비교하는 수치이다. 노드 25, 노드 7, 그리고 노드 4의 경우 전체 데이터 셋의 사망 데이터 비율과 비교해 약 47배의 비율을 보인다는 것을 알 수 있다. 또한 노드 26의 경우는 약 40배의 비율을 보이는 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 70%의 사망 재해율을 평가하고 싶다고 가정하자. 그러면 위의 < 표 2 >의 이익도표에서 Gain (%)가 70%를 넘는 노드 21, 26, 4, 7, 25가 목표 노드가 된다.

3.3 모형구축 자료와 모형검증 자료

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. 즉, 하나의 자료는 의사결정나무를 구축하는데 사용하고 이 때 얻어진 결과를 나머지 다른 자료에 적용하여 타당성을 평가해 보는 것이다. 이 경우에 모형을 구축하는데 사용되는 자료를 모형구축 자료(Training Data)라 하고, 얻어진 결과에 적용해 보는 자료를 모형검증 자료(Testing Data)라고 한다. 만약 모형구축 자료에 의해서 얻어진 결과가 모형검증 자료에서도 뛰어난 성능을 보이는 경우에는 모형을 일반화시키게 되는 것이다.

본 논문에서는 모형구축 자료와 모형검증 자료 타당성 평가를 위하여 각각의 비율을 50% : 50%로 설정을 하였으며, 그 결과 모형구축 자료 표본은 5,266개로, 모형검증 자료는 표본은 5,270개로 분할 된 것을 알 수 있다. 모형구축 자료 표본에서 부상은 5,159개, 사망이 107개이고, 모형검증 자료 표본은 부상이 5,154개, 사망이 116개로 분할되었다. 분할된 데이터를 바탕으로 모형구축 자료와 모형검증 자료를 비교한 결과 다음 < 표 3 >과 같이 나타났다. 여기서 정확도(Accuracy)와 민감도(Sensitivity)의 경우 모형구축 자료와 모형검증 자료에서 약 0.2%의 미미한 차이를 보였다. 또한 특이도(Specificity)에서도 약 3.8%의 적은 차이를 보이기 때문에 QUEST 알고리즘을 이용한 산업 재해 분석은 타당한 것을 알 수 있다.

< 표 3 > Training Data Sample과 Testing Data Sample 비교

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

3.4 교차타당성

교차타당성(Cross-Validation)[12] 평가는 어느 하나의 집단을 제외한 다음 나머지 자료를 가지고 모형을 구축하여 최종적으로 k개의 모형을 구축하게 되는 것이다. 모형구축이 끝나면 k개의 위험의 평균을 계산하고 이를 교차타당성 평가에 의한 위험 추정치로 사용하는 것이다. 이와 같은 방식의 교차타당성 평가에 의한 위험 추정치로 사용

하는 것이다. 이와 같은 방식의 교차타당성 평가를 k-Fold 교차타당성 평가라고 한다. 본 논문에서는 교차타당성 평가를 위하여 10-Fold 교차타당성 평가를 실시하였다.

교차타당성 평가 결과 다음의 < 표 4 >에서 보는 바와 같이 구축 모형의 오분류 확률과 10-Fold 교차타당성 오분류 확률의 차이가 거의 없으므로 형성된 트리는 일반화하기에 충분하다고 평가되어 진다.

< 표 4 > Cross-Validation 결과

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

4. 특성분석 고찰

QUEST 알고리즘을 이용하여 트리를 형성한 결과 < 그림 1 >에서 볼 수 있듯이 사망 재해자의 빈도가 높은 경우를 확인할 수 있었다. Node 25의 경우 사망 재해자 빈도율이 100%로써, 재해 발생 형태가 교통사고, 충돌, 직업병이고, 회사 규모가 200인~299인, 500인~999인이며, 재해 시간이 0시~2시경, 4시~6시경, 8~밤 12시까지로 나타났다. Node 7의 경우 사망재해자 빈도율이 100%로써, 발생 형태가 진폐이고 재해 시간이 12시에서 14시 사이인 것을 알 수 있다. 또한 Node 4의 경우 사망 재해자 빈도율이 100%로써, 발생형태가 진폐이고, 재해 시간이 8시~10시경, 10시~12시경으로 나타났다.

형성된 트리가 얼마나 타당성을 가지는가를 검증하기 위하여 모형구축 자료와 모형검증 자료 비교를 해 본 결과 < 표 3 >에서 볼 수 있듯이 모형구축의 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. 또한 교차타당성 검증 결과 < 표 4 >에서 볼 수 있듯이 구축된 모형의 오분류 확률과 교차타당성 오분류 확률의 차이가 0.02%로 매우 미미한 차이를 보였다. 따라서 교차타당성 검증 결과 형성된 트리는 일반화하기에 충분하다는 결론을 내릴 수 있다.

5. 결론 및 추후연구

본 논문에서는 의사결정나무 기법의 한 알고리즘인 QUEST를 이용하여 정량적 위험성 평가기법이 부재한 제조업에서의 산업재해에 대한 특성을 분석하였다. 분석 결과 제조업에서 사망 재해자에 가장 큰 영향을 미치는 변수의 발생형태로는 교통사고와

충돌 그리고 직업병이 주원인이었다. 그리고 발생형태가 교통사고와 충돌 그리고 직업병이고, 회사 규모(200인~299인, 500인~99인)가 결정적인 변수로 작용을 한다는 것을 알 수 있었다. 그리고 발생형태가 교통사고와 충돌 그리고 직업병이고, 회사 규모가 200인~299인, 500인~99인일 경우 재해 시간이 10시~12시경, 14~16시경일때 사망재해자의 발생 빈도율이 매우 높은 것을 알 수 있었다. 이러한 트리분석 결과가 얼마나 타당한가를 검증하기 위하여 모형구축 자료와 모형검증 자료를 비교한 결과 본 논문에서 형성된 트리는 충분한 타당성을 보였다. 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, 결과의 정확성이 높은 알고리즘을 선정하고자 한다.

6. 참 고 문 헌

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

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

[3] 노동부, 산업재해현황분석, 2004.

[4] Cezary Z. Janikow, "Fuzzy Decision Trees : Issues and Methods", IEEE Transactoins on Systems, Man, and Cybernetics-Part B : Cybernetics, Vol. 28, No. 1, Feb, 1998.

[5] 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.

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

[7] F. Berzal, iet al., "On the quest of easy-to-understand splitting rulws", Data and Knowledge Engineering", Vol. 44, pp. 31-48, 2003

[8] 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.

[9] 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.

[11] R. Godin, R. Missaoui, "An incremental concept formation approach for learning from databases", Theoret. Comput. Sci. 133, pp. 387~419, 1994.
[12] Weiss, S. M., Kulikowski C. A., *Computer Systems That learn*, Morgan Kaufmann, San Mateo, CA, 1991.

저 자 소 개

임 영 문 : 연세대학교에서 학사, 석사학위를 취득하였고, 미국 텍사스주립대학교 산업시 스템공학과에서 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 바사과정에 재학 중이며 관심분야는 Ubiquitous System, 알고리즘 분석 및 활용 등이다.