# 산업재해 예방을 위한 최적 알고리즘 선정

# Selection of an Optimal Algorithm for Prevention of Industrial Accidents

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

## Abstract

산업재해 통계분석의 커다란 목적은 각 산업별로 주 위험요인을 도출하고 이에 따른 안전교육의 실시 또는 안전장치 등을 보완함으로써 산업재해를 줄이거나 예방하는데 있다고 볼 수 있다. 그러나 일반 제조업이나 건설업 등에서는 아직까지도 정량적 위험성 평가 기법이 개발되어 있지 않은 실정이다. 따라서 효율적인 위험성 평가 기법의 개발이 필요하다. 본 연구에서는 데이터마이닝 기법을 이용한 산업재해 예방을 위한 최적 알고리즘 선정 방법을 제시한다.

Keyword : 데이터마이닝, AnswerTree, CART, CHAID, QUEST

## 1. 서론

위험한 화학물질을 취급하는 사업장이나 공공시설에서는 위험도를 정량적 수치로 나타내는 소위 정량적 위험성 평가기법을 적용하여 제도화하거나 사업장 스스로 활용하고 있으나, 일반 제조업이나 건설업 등에서는 아직까지도 정량적 위험성 평가 기법이 개발되어 있지 않는 실정이다. 산업재해와 관련한 기존 연구 대부분이 통계자료의 재해 구성 비율 분석과 같은 빈도 분석에만 의존하였고[3][4], 최근에서야 산업재해 예방을 위해 데이터마이닝 기법이 적용되고 있는데 이러한 노력은 산업재해 분야에서 처음 시도되는 독창적인 연구라고 할 수 있다. 그러나 최근 데이터마이닝 기법을 이용한 연구에서는 단지 데이터마이닝 알고리즘 중에서 하나의 알고리즘을 선택하여 분석을 하였을 뿐, 최적 알고리즘을 찾고자 하는 노력은 없었다[1][2].

* 강릉대학교 정보전자공학부 교수
** 강릉대학교 정보전자공학부 석사과정

따라서 본 연구에서는 데이터마이닝 기법을 적용하여 총 10개 업종 중에서 빈도가 가장 높은 건설업을 중심으로 최적의 알고리즘을 선정하고자 통계 툴인 SPSS와 데이터마이닝 툴인 AnswerTree를 사용하였다.

## 2. 데이터 셋

본 연구에서 사용된 데이터 셋은 2002년부터 2004년까지 산업자원부에서 강원도를 대상으로 집계한 재해자 통계자료이다. 이 데이터들의 특성을 살펴보면, 개인 신상보호를 위한 사업장명과 재해자명을 제외하고, 또한 NULL 데이터를 다수 포함하고 있는 독립변수를 제외하면 발생형태, 규모, 연령, 성별, 근속기간, 재해월, 재해요일 그리고 재해시간으로 함축된다.

업종별 재해자 분포를 살펴보면 아래 <표 1>와 같다. 아래 <표 1>에서 보는 바와 같이 업종 중에서 가장 빈도가 높은 것은 건설업이기에 본 연구에서는 빈도가 가장 높은 건설업을 중심으로 최적 알고리즘을 선정하고자 한다. 건설업에서의 성별에 따른 재해자 분포를 살펴보면 아래의 <표 2>와 같다.

<표 1> 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 |

<표 2> 건설업에서 성별에 따른 재해자 분포

| 성 별 | 재해자 형태 | | 합 계 |
|---|---|---|---|
| | 부 상 자 | 사 망 자 | |
| 남 자 | 18,167 | 582 | 18,749 |
| 여 자 | 808 | 17 | 825 |
| 합 계 | 18,975 | 599 | 19,574 |

## 3. AnswerTree 결과 분석

본 연구에서는 여러 가지 알고리즘 중에서 AnswerTree에서 분석이 가능한 CART, CHAID, QUEST의 3가지 알고리즘만을 비교 분석하여 최적 알고리즘을 선정하였다.

### 3.1 오분류 확률 비교

최적 알고리즘을 선정하기 위해 우선 각 알고리즘별 오분류 확률을 평가해 본 결과 <표 3>과 같이 나타났다. 최초 부모 마디의 오분류 확률인 0.0306018에서 CART는 0.0184428로, CHAID는 0.0083274로, 그리고 QUEST는 0.0222234로 감소하였다. 오분류 확률의 감소량에서 볼 수 있듯이 CHAID가 가장 많이 감소하였기 때문에 3가지 알고리즘 중 CHAID의 정확도가 가장 높은 것을 알 수 있다.

<표 3> 알고리즘들의 오분류 확률 비교

| Algorithm | 오분류 확률 | |
|---|---|---|
| | Root Node | Final Node |
| CART | 0.0306018 | 0.0184428 |
| CHAID | 0.0306018 | 0.0083274 |
| QUEST | 0.0306018 | 0.0222234 |

### 3.2 모형구축 자료(Training Data Set)과 모형 검증자료(Testing Data Set) 비교

아래 <표 4>에서 Training Data Set에서 민감도(Sensitivity)가 가장 높은 것은 알고리즘은 CHAID로 99.16353%이다. 또한 특이도(Specificity)와 정확도(Accuracy)에서도 CHAID가 각각 83.70370%과 98.79907%로 가장 높게 나타났다. Testing Data Set에서는 민감도(Sensitivity)는 CHAID가 98,99588%로, 특이도(Specificity)에서는 CART가 71.12971%로, 정확도(Accuracy)에서는 CHAID가 98.19302%로 높게 나타났다.

이러한 결과에서 볼 수 있듯이 AnswerTree에서 분석이 가능한 3가지 알고리즘, CART, CHAID, QUEST 중 CHAID가 가장 적합하다고 볼 수 있다.

<표 4> 알고리즘 특성도 비교

| Algorithm | Training Data Set | | | Testing Data Set | | |
|---|---|---|---|---|---|---|
| | Sensitivity (%) | Specificity (%) | Accuracy (%) | Sensitivity (%) | Specificity (%) | Accuracy (%) |
| CART | 98.80146 | 79.91632 | 98.34249 | 98.70540 | 71.12971 | 98.02875 |
| CHAID | 99.16353 | 83.70370 | 98.79907 | 98.99588 | 70.96774 | 98.19302 |
| QUEST | 97.83661 | 75.59055 | 97.54932 | 97.96748 | 67.12329 | 97.50513 |

## 4. 결론 및 추후 연구

본 연구에서는 데이터마이닝 기법을 적용하여 빈도가 가장 높은 건설업을 중심으로 최적의 알고리즘을 선정하고자, 우선 오분류 확률을 비교하고, 그 다음으로 모형 구축 자료(Training Data Set)와 모형 검증자료(Testing Data Set)를 비교하였다. 마지막으로 교차타당성(Cross-Validation) 평가를 통해 타당성 평가를 해 보았다. 가장 낮은 오분류 확률을 갖는 알고리즘은 CHAID로서 0.83274%로 나타났다. 따라서 오분류 확률의 감소량에서 CHAID 알고리즘이 가장 적합한 알고리즘으로 나타났다.

모형 구축 자료(Training Data Set)에서 민감도(Sensitivity)가 가장 높은 알고리즘은 CHAID로 99.16353%이다. 또한 특이도(Specificity)와 정확도(Accuracy)에서도 CHAID가 각각 83.70370%와 98.79907%로 가장 높게 나타났다. 모형 검증 자료(Testing Data Set)에서 민감도(Sensitivity)는 CHAID가 98,99588%로, 특이도(Specificity)에서는 CART가 71.12971%로, 정확도(Accuracy)에서는 CHAID가 98.19302%로 높게 나타났다. 따라서 모형 구축 자료(Training Data Set)와 모형 검증 자료(Testing Data Set) 비교에서 CHAID 알고리즘이 가장 적합한 알고리즘으로 나타났다. 따라서 AnswerTree에서 지원 가능한 데이터마이닝 알고리즘 3가지(CART, CHAID, QUEST) 중 산업재해와 관련된 데이터를 분석하고 예측하는데 있어서 가장 적합한 알고리즘은 CHAID 알고리즘이다. 또한 여러 업종 중에서 건설업 데이터를 분석하고 예측하는데 있어서도 CHAID 알고리즘이 최적 알고리즘이다.

추후에는 단일 업종이 아닌 여러 업종에 대해서 방대한 데이터들을 확보하여 각 업종별 산업재해 예방에 적합한 최적인 알고리즘을 선정하고자 한다.

**Acknowledgement**
본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

## 5. 참고문헌

[1] 곽준구, 의사결정나무 기법을 이용한 업종별 산업재해의 특성 분석, 강릉대학교 석사학위, 2005, pp. 22~28.

[2] 황영섭, 데이터마이닝 기법을 이용한 산업재해 감소를 위한 재해자들에 대한 요인분석, 한국산업경영시스템학회 춘계학술대회, 2005.

[3] R. Godin, R. Missaoui, An incremental concept formation approach for learning from databases, Theoret. Comput. Sci. 133, 1994, pp. 387~419.

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