산업안전 향상을 위한 전문가 시스템 구축에 관한 연구

A Study on Construction of an Expert System for Enhancement of Industrial Safety

임 영 문*
Young-Moon, Leem
최 요 한**
Yo-han, Choi

Abstract

급속도로 발전하는 산업의 고도화와 이에 따른 업종의 다양화, 이에 동반되는 예상치 못한 산업재해는 불특정 다수에게 인적, 물적 피해를 야기 시키고 있다. 산업재해 예방을 위해 다양한 선행 연구들이 진행되었으나 이들 연구는 기존의 산업재해 데이터를 토대로 빈도분석, 비교분석을 통한 관리적, 교육적 등의 대책만을 제시하고 있다. 본 연구에서는 산업재해 예방을 위해 객관적이고 정량화된 데이터를 통한 예측 분석이 가능한 데이터마이닝을 적용하여 대표적인 기법인 의사결정나무의 CHAID, CART, C4.5, QUEST 4가지 알고리즘 비교분석하여 산업재해 예방 및 전문가 시스템 구축을 위해 적용할 수 있는 최적의 알고리즘을 제시하도록 한다.

Keyword : 데이터마이닝, 의사결정나무, CHAID, CART, C4.5, QUEST

1. 서론

급속도로 발전하는 산업의 고도화와 이에 따른 업종의 다양화, 이에 동반되는 예상치 못한 산업재해는 불특정 다수에게 인적, 물적 피해를 야기 시키고 있다. 이러한 사고를 방지하기 위한 최선의 방법은 사고 발생 위험 요소에 대한 사전제거이며, 차선의 방법은 사고의 예방이다. 사고의 예방을 위한 방법이나 대책으로 여러 면에서 많은 연구나 법규, 방호장치, 기법, 보호구등이 개발되어 현실화 되고 있으나, 예측하지 못한 상황, 조건 등 여러 변수들에 의해 사고는 지속적으로 발생되고 있다. 그러나 기존에 발생된 사고에 대한 많은 정보들을 분석하여 보면, 보다 객관적이고 정량화된 데이터

* 강릉대학교 정보전자공학부 교수
** 강릉대학교 정보전자공학부 박사과정 수료

를 얻을 수 있을 것이고 이를 통한 예측 분석이 가능할 것이다. 그동안 산업재해 예방을 위해 많은 연구가 선행되어 졌으며, 많은 양의 데이터가 축적되었다. 그러나 산업재해와 관련된 연구의 대부분은 과거 발생된 재해에 대해 단순한 빈도분석이나 비교분석을 통한 예방대책의 제시가 주를 이루고 있다. 대량의 데이터를 분석하기 위한 기존의 데이터 분석방법으로는 여러 면에서 한계를 나타내고 있다. 이에 대용량의 데이터를 효과적으로 분석하기 위에 데이터마이닝(Data Mining)[3][4]이라는 분야가 부각되어 여러 분야에서 적용되고 있다.

따라서 본 연구에서는 데이터마이닝을 적용하여 대표적인 기법인 의사결정나무의 CHAID, CART, C4.5, QUEST 4가지 알고리즘을 SAS의 Enterprise Miner[2]와 SPSS의 AnswerTree[1] 소프트웨어를 이용하여 비교 분석한 뒤 산업재해 예방 및 전문가 시스템 구축을 위해 적용할 수 있는 최적의 알고리즘을 제시하도록 한다.

2. 연구방법

본 연구에서 사용된 데이터는 아래 <표1>과 같이 강원도 관내 전 업종에서 2002년부터 2004년까지 3년간 산업재해를 신청하여 산재로 결정 통지된 가장 대표적인 업종인 건설업과 제조업의 30,110개의 관찰치를 대상으로 하였다. SAS의 Enterprise Miner와 SPSS의 AnswerTree 소프트웨어를 이용하여 의사결정나무의 대표적인 알고리즘인 CHAID, CART, QUEST, C4.5의 타당성 평가를 한 후 알고리즘별 민감도, 특이도, 정확도를 구하여 비교 분석한 후 산업안전 향상을 위한 전문가 시스템 구축에 적용할 최적 알고리즘을 선택하고자 한다.

<표 1> 업종에 따른 재해자 분포

업 종	재해자 형태		합 계
	부 상 자	사 망 자	
건설업	18,975	599	19,574
제조업	10,313	223	10,536
합 계	29,288	822	30,110

3. 데이터 분석 방법

본 연구에 사용된 총 30,110건의 원시 데이터는 재해 일자, 재해자 구분, 발생형태, 업종, 규모, 직종, 진료 일수, 입원 일수, 통원 일수, 연령, 성별, 요양 기간, 근속기간, 재해월, 재해 요일, 재해 시간, 근로 손실 일수 등의 총 18개의 변수들로 구성되어 있다. 그러나 의미 있는 정보를 찾기 위해 총 18개 변수들 중 분석에 불필요한 변수들을 제외한 총 9개의 변수를 선택하여, 재해형태를 목표변수로 선택하고, 발생형태, 규모, 연령, 성별, 근속기간, 재해월, 재해요일, 재해시간을 예측변수로 구성하였는데 이들 변수들 중에서 재해형태와 연령은 이산형(Binary) 변수로, 발생형태, 재해월, 재해요일,

재해시간은 명목형(Nominal) 변수로, 규모, 연령, 근속기간은 연속형(Interval) 변수로 구성하여 의사결정나무 분석의 대표적인 알고리즘인 CHAID, CART, C4.5, QUEST를 비교 분석하여 최적의 모형을 선택하기 위하여 건설업과 제조업에 대한 각 알고리즘별 정분류율 또는 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity)를 구하여 비교 분석 하였다.

4. 데이터 분석 결과

본 연구에 사용된 건설업과 제조업의 알고리즘을 비교 분석한 결과 다음의 <표 2>, <표 3>와 같이 나타났다.

<표 2> 건설업의 알고리즘 비교

Algorithm	Training set			Testing set		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
CHAID	99.16353	83.70370	98.73907	98.99588	70.96774	98.19302
CART	98.80146	79.91632	98.34249	98.70540	71.12971	98.02875
QUEST	97.78624	74.59016	97.49848	97.90669	66.66667	97.46407
C4.5	98.33887	97.40249	87.73884	94.63087	86.34208	86.59446

<표 3> 제조업의 알고리즘 비교

Algorithm	Training set			Testing set		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
CHAID	99.70925	82.24299	99.35435	99.63135	84.48276	99.29791
CART	96.32188	64.48598	95.69833	99.86437	54.31034	98.86299
QUEST	99.86431	64.48598	99.14546	99.70981	65.51724	98.95932
C4.5	95.32710	97.30672	97.26651	87.93103	97.16615	96.96279

위의 <표 2> 건설업의 알고리즘을 비교해 보면 CHAID가 Training set(분석용 데이터)과 Testing set(평가용 데이터)에서 모두 정확도(Accuracy)가 98.73907%와 98.19302%, 민감도(Sensitivity)는 99.16353%와 98.99588%로 가장 높게 나타났으며, 특이도(Specificity)는 C4.5가 Training set(분석용 데이터)과 Testing set(평가용 데이터)에서 97.40249%와 86.34208%로 가장 높게 나타냄으로 최적의 알고리즘으로 나타났다. 또한 <표 3> 제조업의 알고리즘을 비교해 보면 정확도(Accuracy)는 CHAID가

Training set(분석용 데이터)와 Testing set(평가용 데이터)에서 모두 99.35435%와 99.29791%로 가장 높게 나타났으며, 민감도(Sensitivity)는 QUEST가 Training set(분석용 데이터)과 Testing set(평가용 데이터)에서 모두 99.86431%와 99.70981%로 가장 높게 나타났고, 특이도(Specificity)는 C4.5가 Training set(분석용 데이터)과 Testing set(평가용 데이터)에서 97.30672%와 97.16615%로 가장 높게 나타냄을 알 수 있으나, CHAID가 민감도, 특이도에서 다른 알고리즘 보다 높게 나타냄으로 정확도, 민감도, 특이도 전체를 비교하여 볼 때 제조업에서도 CHAID가 최적의 알고리즘임을 알 수 있다.

5. 결론 및 추후 연구

본 연구에는 데이터마이닝 기법 중 가장 대표적인 의사결정나무인 CHAID, CART, C4.5, QUEST 4가지 알고리즘에 대한 타당성 평가 및 비교 분석을 통해 산업재해 예방을 위한 전문가 시스템 구축에 적용할 최적의 알고리즘 제시하였다.

SAS의 Enterprise Miner와 SPSS의 AnswerTree 소프트웨어를 이용하여 알고리즘의 타당성 평가를 한 후 알고리즘별 민감도, 특이도, 정확도를 구하여 비교 분석한 결과, 건설업에서는 CHAID가 정확도, 민감도에서 가장 높게 나타났고, C4.5가 특이도에서 가장 높게 나타났다. 제조업에서는 정확도에서는 CHAID, 민감도에서는 QUEST, 특이도에서는 C4.5가 가장 높게 나타났으나, CHAID가 민감도, 특이도에서도 높게 나타남으로 전체적인 평가를 할 때 제조업에서도 CHAID가 높게 나타남으로 건설업, 제조업 모두 CHAID가 최적의 알고리즘임을 알 수 있으며, 이를 토대로 다른 업종에 적용하여도 동일한 결과를 나타낼 것으로 추측된다. 본 연구에서 제시한 연구결과를 토대로 산업안전향상을 위한 전문가 시스템을 구축할 예정이다.

Acknowledgement

본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

6. 참고문헌

[1] 강현철, 최종후, 한상태, 김은석, Answer Tree를 이용한 데이터 마이닝, SPSS아카데미, 2001.

[2] 강현철, 한상태, 최종우, 김은석, 김미경, SAS Enterprise Miner 4.0을 이용한 데이터마이닝, 자유아카데미, 2001.

[3] 문정호, 사례연구를 통한 데이터마이닝 수행과정 연구, 서울대학교 석사학위논문, 2002.

[4] 장남식, 홍성완, 장재호, 데이터마이닝, 대청, 1999.