웹 기반의 산업재해 예측시스템 개발에 관한 연구 - A Study on Development of A Web-Based Forecasting System of Industrial Accidents -

임 영 문 *
Leem Young Moon
황 영 섭 **
Hwang Young Seob
최 요 한 ***
Choi Yo Han

Abstract

Ultimate goal of this research is to develop a web-based forecasting system of industrial accidents. As an initial step for the purpose of this study, this paper provides a comparative analysis of 4 kinds of algorithms including CHAID, CART, C4.5, and QUEST. In addition, this paper presents the logical process for development of a forecasting system. Decision tree algorithm is utilized to predict results using objective and quantified data as a typical technique of data mining. The sample for this work was chosen from 10,536 data related to manufacturing industries during three years(2002~2004) in korea.

Keywords : Data Mining, Decision Tree, CHAID, CART, C4.5, QUEST, Forecasting System

본 연구는 산학연 공동기술개발 사업의 연구결과로 주식회사 여한테크와 공동으로 수행되었음.

* 강릉대학교 산업시스템공학과 교수
** 강릉대학교 산업시스템공학과 박사과정
*** (주)여한테크 기획이사

1. 서 론

다양한 산업현장에서 본인의 의지와 관계없이 사고를 당해 재해를 입는 경우가 빈번하게 발생되고 있는 현실이다. 예상치 못한 사고는 때와 장소 등에 관계없이 불특정 다수에게 언제나 발생 할 수 있으며, 또한 사고로 인한 재해는 인적, 물적 피해를 발생시키고 있다.

이러한 사고를 방지하기 위한 최선의 방법은 사고 발생 위험 요소를 사전에 제거하는 것이며, 차선의 방법으로는 사고의 예방이다. 그동안 산업재해 예방을 위하여 많은 선행 연구가 진행되어 왔으며, 많은 양의 데이터가 축적되었다.

그러나 산업재해와 관련된 연구의 대부분은 과거에 발생한 재해에 대하여 단순히 빈도 분석이나 비교 분석을 실시한 결과를 토대로 예방대책을 제시하는 경우가 대부분이었다. 대량의 데이터를 분석하여 의미 있는 정보를 찾아내는 것은 미래를 예측할 수 있는 객관적인 방법이 될 수 있을 것이다.

대용량의 데이터를 효과적으로 분석하기 위하여 데이터마이닝(Data Mining)이라는 분야가 부각되어 여러 분야에서 적용되고 있으며[5][6][7], 분석된 결과도 다양하게 활용[4][8]되고 있다. 이에 본 연구에서는 과거의 산업재해 데이터를 데이터마이닝(Data Mining)의 의사결정나무 기법[1]을 적용하여 알고리즘을 비교 분석한 후 최적의 알고리즘을 선정하고자 한다. 또한 선정된 알고리즘을 토대로 노드(Node) 분석 및 데이터베이스(Data Base)를 구성한 후 산업재해 예방을 위하여 정량적, 정성적 예측이 가능한 웹 기반의 산업재해 예측 시스템을 구축하고자 한다.

2. 연구 방법

본 연구에서는 강원도 관내 전 업종에서 2002년부터 2004년까지 3년간 산업재해 신청을 하여 산재로 결정 통지된 67,278건의 데이터에 대하여 SAS의 Enterprise Miner[2]와 SPSS의 AnswerTree[1] 소프트웨어를 적용하고자 한다. 먼저 데이터마이닝의 기법 중에서 의사결정나무의 대표적인 알고리즘인[3] CHAID, CART, QUEST, C4.5를 각각 발생된 산재 데이터에 적용하여 타당성 평가를 한 후 알고리즘별 정확도, 민감도, 특이도 값을 구하여 비교 분석한다. 또한 이를 토대로 최적의 알고리즘을 선택하여 교차타당성(Cross Validation)을 이용한 타당성 평가를 실시한다.

그리고 선택된 최적의 알고리즘을 실행하여 나타난 노드를 분석한 후 데이터테이블을 작성하여 데이터베이스를 구축한다. 이러한 과정들을 토대로 산업재해 예방을 위하여 정량적, 정성적 예측이 가능한 웹 기반의 예측 시스템을 개발하고자 한다.

3. 데이터 분석 결과

데이터가 수집된 전체 10개의 업종 중에서 가장 대표적인 업종인 건설업과 제조업을 샘플링으로 선택하여 총 67,278건의 원시 데이터 중에서 30,110건의 데이터를 관찰치로 사용하였다. 데이터에 대한 알고리즘을 비교 분석한 결과 다음의 <표 1>, <표 2>과 같은 값을 얻을 수 있었다.

다음의 <표 1>과 <표 2>에서 볼 수 있듯이 건설업과 제조업에 관련된 산업재해 데이터 분류에 대한 정확도, 민감도, 특이도 전체를 비교하여 볼 때 전반적으로 CHAID가 가장 높은 값을 나타냄으로 CHAID가 최적의 알고리즘임을 알 수 있었다.

<표 1> 건설업에 대한 알고리즘 비교

Algorithm	Training set			Testing set		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
CHAID	98.739	99.163	83.703	98.193	98.995	70.967
CART	98.342	98.801	79.916	98.028	98.705	71.129
QUEST	97.498	97.786	74.590	97.464	97.906	66.666
C4.5	87.738	98.338	97.402	86.594	94.630	86.342

<표 2> 제조업에 대한 알고리즘 비교

Algorithm	Training set			Testing set		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
CHAID	99.354	99.709	82.242	99.297	99.631	84.482
CART	95.698	96.321	64.485	98.862	99.864	54.310
QUEST	99.145	99.864	64.485	98.959	99.709	65.517
C4.5	97.266	95.327	97.306	96.962	87.931	97.166

또한 건설업과 제조업의 경우에서 CHAID 알고리즘에 대한 교차타당성(Cross Validation)을 분석하여 보면 <그림 1>과 같이 건설업에 대한 교차타당성은 모형구축 오분류 값이 0.00837846, 교차타당성 값이 0.017932이며, <그림 2>에서 볼 수 있듯이 제조업에 대한 교차타당성은 모형구축 오분류 값이 0.0603426, 교차타당성 값이 0.0774704로 값의 차이가 적게 나타남으로 타당함을 알 수 있다.

Misclassification Matrix

		Actual Category		
		1	2	Total
Predicted Category	1	18908	9	19805
	2	67	5	569
	Total	18975	5	19574

	Resubstitution	Cross-Validation
Risk Estimate	0.00037846	0.017932
SE of Risk Estimate	0.000651501	0.000948516

<그림 1> 건설업 데이터에 대한 교차타당성(Cross Validation)

Misclassification Matrix

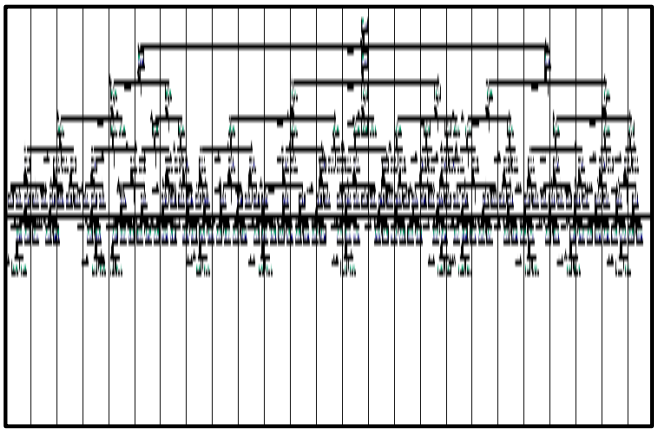
		Actual Category		
		2	1	Total
Predicted Category	2	141	2	167
	1	203	3	3628
	Total	344	3	3795

	Resubstitution	Cross-Validation
Risk Estimate	0.0603426	0.0774704
SE of Risk Estimate	0.00386537	0.00433962

<그림 2> 제조업 데이터에 대한 교차타당성(Cross Validation)

4. CHAID 알고리즘 분석 결과

SPSS의 AnswerTree를 이용하여 CHAID 알고리즘을 실행한 결과 (그림 3)과 같은 노드의 형태가 도출되었다.



<그림 3> CHAID 알고리즘의 모형 분석 결과

위의 <그림 3>을 살펴보면 뿌리마디는 19,574개의 관측치로 부상 18,945개(96.79%)와 사망 599개(3.06)%로 나타났고, 깊이(Depth)는 7이며, 끝마디는 총 124개로 형성되었다. CHAID 알고리즘을 이용한 노드 분석 결과 재해형태, 재해월, 재해요일, 재해시간, 규모, 근속기간, 발생형태의 7가지 변수를 모두 만족하는 노드는 전체 124개의 끝마디 중 41개의 끝마디로 나타났다. 이러한 41개의 끝마디를 토대로 산업재해 예측 시스템을 구축하고자 <그림 4>와 같이 데이터테이블을 구성하여 데이터베이스로 활용 하였다.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	연번	월	요일	시간	경우수	규모	근속기간	발생형태	부상n	사망n	all	부상%	사망%
2	1	3	월(1)	0-2(1)	1	2,3,4,5	1	5	193	7	19574	0.986	0.036
3	2	3	월(1)	0-2(1)	2	2,3,4,5	1	8	293	22	19574	1.497	0.112
4	3	3	월(1)	0-2(1)	3	2,3,4,5	1	1,2,3,4,6,9	551	0	19574	2.815	0.000
5	4	3	월(1)	0-2(1)	4	2,3,4,5	1	7	4	5	19574	0.020	0.026
6	5	3	월(1)	0-2(1)	5	6,7	1,4,7	1,3,4,5,7,9	122	0	19574	0.623	0.000
7	6	3	월(1)	0-2(1)	6	6,7	1,4,7	8	199	15	19574	1.017	0.077

<그림 4> 데이터테이블

5. 예측 시스템

개발 중에 있는 예측 시스템에 접속하면 아래의 <그림 5>와 같이 초기화면을 볼 수 있으며 메뉴는 조회, 취소, 종료로 구성되어 있고 입력변수는 업종, 월, 요일, 시간, 규모, 근속기간으로 구성되어 있다. 이에 대한 결과는 재해발생 예측 위험지수, 부상 사망 현황, 검색결과 Case로 구성되어 출력된다.

<그림 5> 웹 기반의 산업재해 예측 시스템

위의 <그림 5>를 보면 입력변수인 업종, 월, 요일, 시간, 규모, 근속기간을 입력할 경우 입력변수 값과 동일한 조건의 검색 결과를 발생형태와 발생형태별로 부상과 사망 확률, 경우(Case)별 부상 및 사망 현황을 정량적으로 나타내며, 이에 따른 재해발생 예측 위험지수를 매우 높음, 높음, 낮음, 결과 없음과 같이 정성적으로도 나타낸다.

이러한 과정을 토대로 하여 시스템 구성 요소에 대하여 현장 적용성을 검토한 후 웹 기반의 산업재해 예측시스템을 구축 하고자 한다.

6. 결론 및 추후연구

본 연구는 웹 기반의 산업재해 예측 시스템을 구축하고자 선행 작업으로 데이터마이닝 기법 중 의사결정나무의 대표적인 알고리즘인 CHAID, CART, C4.5, QUEST를 SAS의 Enterprise Miner와 SPSS의 AnswerTree를 이용하여 대표적 업종인 건설업과 제조업에 관련된 데이터에 대하여 알고리즘 별 타당성 평가와 정확도, 민감도, 특이도를 구하여 비교 분석하였다. 그 결과 CHAID가 최적의 알고리즘으로 선택되었다. 이를 토대로 웹 기반의 예측시스템이 개발 중에 있으며 본 연구에서 시스템 구축에 사용된 데이터베이스의 데이터양이 19,574개를 토대로 하였으므로 전 업종에 확대 적용하기엔 다소 무리가 있다고 생각되며, 더 많은 데이터를 확보하여 전 업종에 적용할 수 있는 웹 기반의 예측 시스템을 개발 할 예정이다.

7. 참고문헌

[1] 강현철, 서두성, 최종후, Enterprise Miner의 의사결정나무분석 알고리즘, 아카데미, 2001.
[2] 강현철, 최종후, 한상태, 김은석, Answer Tree를 이용한 데이터마이닝, SPSS아카데미, 2001.
[3] 권혜숙, 데이터마이닝 패키지에서 분류나무 알고리즘의 비교 연구, 서울대학교 석사학위논문, 2002.
[4] 임영문, 황영섭, 최요한, 데이터마이닝 기법을 활용한 산업재해자들에 대한 요인분석, 대한안전경영과학회지 제7권 4호, pp. 61-71, 2005.
[5] David Enke and Suraphan Thawornwong, The use of data mining and neural networks forforecasting stock market returns, Expert Systems with Applications 29, pp. 927-940, 2005.
[6] K. Kirchner, K. -H. Tölle and J. Krieter, "Decision tree technique applied to pig farming datasets", Livestock Production Science, Vol. 90, Issues 2-3, November, pp. 191-200, 2004.
[7] Paul R. Harper, David J. Winslett, "Classification trees; A possible method for maternity risk grouping", European Journal of Operational Research, Vol. 169, issue 1, pp. 146-156, 2006.
[8] S.K. Pal, A. Pal(Eds.), Paul R. Harper, "On learning to predict web traffic", Decision Support Systems, Vol. 35, No. 2, pp.213-229, 2003.