산업재해의 최적 예측모형을 위한 근사모형에 관한 연구

- A Study on Approximation Model for Optimal Predicting Model of Industrial Accidents -

임 영 문 *
Leem Young Moon
유 창 현 **
Ryu Chang Hyun

Abstract

Recently data mining techniques have been used for analysis and classification of data related to industrial accidents. The main objective of this study is to compare algorithms for data analysis of industrial accidents and this paper provides an optimal predicting model of 5 kinds of algorithms including CHAID, CART, C4.5, LR (Logistic Regression) and NN (Neural Network) with ROC chart, lift chart and response threshold. Also, this paper provides an approximation model for an optimal predicting model based on NN. The approximation model provided in this study can be utilized for easy interpretation of data analysis using NN. This study uses selected ten independent variables to group injured people according to a dependent variable in a way that reduces variation. In order to find an optimal predicting model among 5 algorithms, a retrospective analysis was performed in 67,278 subjects. The sample for this work chosen from data related to industrial accidents during three years (2002 ~ 2004) in korea. According to the result analysis, NN has excellent performance for data analysis and classification of industrial accidents.

Keywords : NN, LR, Decision Tree, ROC Chart, Lift Chart, Approximation Model

† 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구 결과로 수행되었음.
* 강릉대학교 산업공학과 교수
** 강릉대학교 산업공학과 석사과정
2006년 5월접수; 2006년 6월 수정본 접수; 2006년 6월 게재확정

1. 서 론

발전하는 산업사회의 영향으로 산업의 공업화가 이루어지면서 산업재해의 유형은 점점 다양화 되고 있다. 특히 그 발생형태가 제조업의 공정 및 작업과정 뿐만이 아니라 토목업, 광업, 운수업등 여러 가지 직업유형에서 발생되고 있다. 2004년도 노동부에 따르면 산업재해로 인한 경제적 직접손실액은 전년도 대비 15.23% 증가하였고, 직·간접손실을 포함한 경제적 손실 추정액이 14,299,570백만원으로 나타났다[3]. 아직까지 산업재해는 여러 가지 측면에서 사회에 큰 비중을 차지하고 있으며 산업재해로 인한 사회간접자본의 손실과 근로자의 안전을 보호 할 수 있는 대책이 마련되어야 한다. 기존의 산업재해 연구 자료들은 대부분 산업재해 데이터를 토대로 사고의 유형분석, 빈도분석, 비교분석에만 의존 하여[2], 사후 관리적, 교육적인 결과들만 제시하고 있다.

데이터 마이닝(Data Mining)기법은 대량의 과거 데이터로부터 자료의 예측이 가능한 기법으로 여러 가지 산업예측모형에 활용되어져 왔다. 데이터 마이닝 적용분야를 보면 이동통신사의 이탈고객 예측모형, 취업고객 분석 및 예측모형, 의학적 진단 예측모형 등으로 활용되고 있다[5,6,8,10,11].

본 연구에서는 강원도에서 3년 동안 발생한 산업재해자 총 67,278명의 자료를 바탕으로 재해형태인 사망 및 부상예측을 위하여 의사결정나무(Decision Tree) 알고리즘의 CHAID, C4.5, CART와 로지스틱회귀모형(Logistic Regression: LR), 신경망(Neural Network: NN)등의 다양한 기법들을 적용하여 결과를 비교 분석하였고, 최적의 성능을 보인 예측모형을 제시하였다. 이러한 예측모형은 산업재해 예방을 위한 대처방안으로 활용 되어질 수 있을 것이다.

2. 연구내용 및 방법

2.1 연구 자료

본 연구에서는 강원도 관내 전 업종(건설업, 제조업, 광업, 금융보험, 농업, 어업, 운수보관, 임업, 전기상수, 기타산업)에서 2002년부터 2004년까지 3년간의 산재로 결정된 67,278건의 데이터를 사용하였다. 데이터는 총 17개의 항목 중 분석에 불필요한 항목을 제외한 재해구분, 발생형태, 업종, 규모, 연령, 성별, 근속기간, 재해월, 재해요일, 재해시간 총 10가지 항목으로 구성하였다.

2.2 분석방법

본 연구에서는 데이터 분류 기법들의 성능을 측정하기 위하여 크게 데이터 입력, 데이터 분할, 변수선택, 기법들의 분류규칙 입력단계, 평가단계로 진행하였다. 데이터 분할 단계에서는 데이터의 검증을 위하여 Training Set과 Testing Set 데이터의 비율을 50:50으로 분할하였고, 변수선택 단계에서는 효율적인 입력변수 선정을 위하여 카이제곱 통계량을 이용하여 입력변수를 선택하였다. 기법들의 분류규칙 입력단계를 거쳐 평가단계에서는 오분류표를 이용한 각 기법들의 정분류율(Accuracy), 오분류율(Error

Rate) 값을 측정하여 수치를 비교 하고, 분류기준 값의 변화에 따라 민감도와 특이도를 고려하여 예측의 정도를 나타내는 ROC Chart와 사후확률을 이용하여 예측의 정확성을 알 수 있는 Lift Chart를 이용하여 모델을 비교, 평가 하였다. 이러한 과정을 통해 얻어진 결과 값들로부터 최적의 분류기법을 선정하였고, 선정기법에 적합한 이해와 해석을 위하여 근사모형 접근방법을 사용하여 예측모형을 구축하였다. 분석도구로는 SAS Enterprise-Miner 4.3[4]을 이용하였다.

3. 분석 결과

3.1 변수 선택

본 연구에서는 총 10개의 입력변수들의 효율적인 입력변수 선정을 위하여 카이제곱 통계량(χ^2)을 이용하여 입력변수를 선택 하였다. 10개의 입력변수들 중 카이제곱 통계량이 3.84 (유의수준 5%)보다 적은 연령, 재해요일을 제외한 재해구분, 발생형태, 업종, 규모, 성별, 근속기간, 재해월, 재해시간으로 결정 되었으며, 이들 변수들 중에서 재해구분, 성별은 이진형(Binary) 변수로, 발생형태, 업종, 재해월, 재해요일은 명목형(Nominal) 변수로, 규모, 근속기간은 연속형(Interval) 변수로 구성하였다.

3.2 모델별 결과 비교

데이터 마이닝 기법들을 비교 분석하기 위하여 정분류율, 오분류율의 값을 분류표(Classification Tables)를 이용하여 계산하였다.

		예측된 변수 0	예측된 변수 1	
원래 목표변수	0	실제0 예측0	실제0 예측0	실제0
	1	실제1 예측0	실제1 예측1	실제1
		예측0	예측1	

<그림 1> 분류표의 구성

분류표란 목표변수의 실제범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 것으로 위의 <그림 1>과 같고, 분류표에 의한 정분류율, 오분류율의 개념은 다음과 같이 정의할 수 있다[1,7].

$$\text{정분류율} = \frac{(\text{실제}0, \text{예측}0)\text{의빈도} + (\text{실제}1, \text{예측}1)\text{의빈도}}{\text{전체빈도}}$$

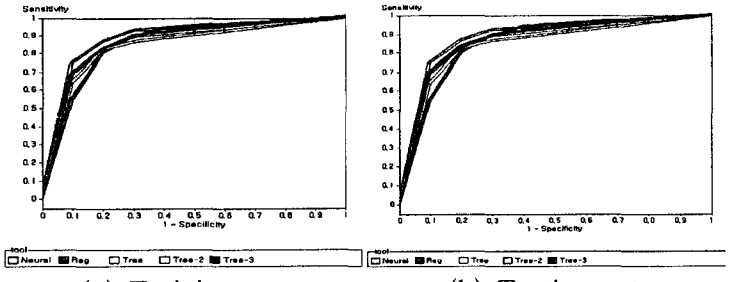
$$\text{오분류율} = \frac{(\text{실제}0, \text{예측}1)\text{의빈도} + (\text{실제}1, \text{예측}0)\text{의빈도}}{\text{전체빈도}}$$

분류표를 통해 얻은 데이터 마이닝 기법들의 결과 값은 <표 1>과 같다.

<표 1> 기법들의 분류결과

	Training Set		Testing Set	
	정분류율	오분류율	정분류율	오분류율
NN	93.09%	6.91%	92.94%	7.06%
LR	90.75%	9.25%	90.75%	9.25%
CHAID	92.19%	7.81%	91.92%	8.08%
C4.5	92.37%	7.63%	92.25%	7.75%
CART	92.84%	7.16%	92.40%	7.60%

위의 <표 1>에서 볼 수 있듯이 기법들을 비교해 보면 정분류율은 NN이 Training Set에서 93.09%로 가장 높은 분류율을 나타냈고, LR은 90.75%로 가장 낮은 분류율을 보였다. 오분류율에서도 NN이 6.91%로 가장 우수하였으며, LR과 비교했을 때 2.3%가 차이가 났다. Testing Set 데이터에서도 NN이 정분류율 92.94%, 오분류율 7.06%로 가장 좋은 성능을 나타내었다. 분류표에 의한 결과를 보면 전체적으로 NN이 높은 분류 결과를 나타냈다.



(a) Training set (b) Testing set

<그림 2> ROC Chart

ROC(Receive Operating Characteristic) Chart는 이진형의 목표변수를 가지는 모형들의 성능을 비교, 평가하는데 매우 유용한 도표로 사용된다. ROC Chart는 X축(1-특이도)과 Y축(민감도)으로 각 분류기준 값에 대해 나타내며, 이러한 결과에 따라 그래프가 도표의 왼쪽 상단에 더 가까운 모형을 성능 면에서 우수한 모형으로 판단하면 된다[9]. 위의 <그림 2>는 ROC Chart에 의한 기법들의 성능을 보여주고 있다. ROC Chart에서 나타나는 결과를 보면 NN이 다른 기법들에 비해서 Training Set과 Testing Set 모두의 데이터에서 정확한 결과를 예측할 수 있는 기법으로 판단된다.Lift Chart는 사후확률을 이용하여 예측의 정확성을 알 수 있다. Lift Chart는 각각의 관측치에서 사후확률을 구한 후 사후확률의 크기순서에 따라 전체 자료를 균일하게 N등분한 후 각 집단에서의 누적 %Captured Response, %Response 그리고 Lift를 계산한다. 각각의 의미는 다음과 같다[1].

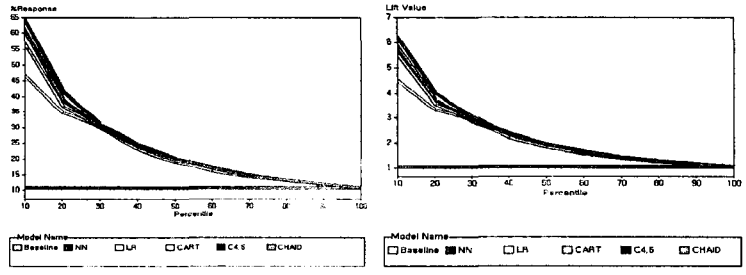
$$\%\text{Captured Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{전체에서 목표변수의 특정범주 빈도}} \times 100$$

$$\%\text{Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{해당 등급에서 전체 빈도}} \times 100$$

$$\text{Base Line \%Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{전체 빈도}} \times 100$$

$$\text{Lift} = \frac{\text{해당등급의 \%Response}}{\text{Base Line Lift}} \times 100$$

<그림 3>에서 볼 수 있듯이 누적 %Response 도표의 수평축은 사후확률 값으로 전체 데이터를 정렬하여 10%씩 나눈 각 집단을 나타내고, 맨 좌측의 10은 사후확률이 가장 높은 10% 집단을 나타낸다. %Captured Response는 특정범주 내에서 특정 등급이 차지하고 있는 점유율로 해석할 수 있으며, Lift Value는 전체 집단에 비해 해당 등급에서 예측력이 향상된 정도를 나타낸다.



%Response Lift Value

<그림 3> Lift Chart

사망 가능성이 높은 상위집단 10%, 20%, 30%의 누적이익도표는 다음의 <표2>에 나타난 것과 같다.

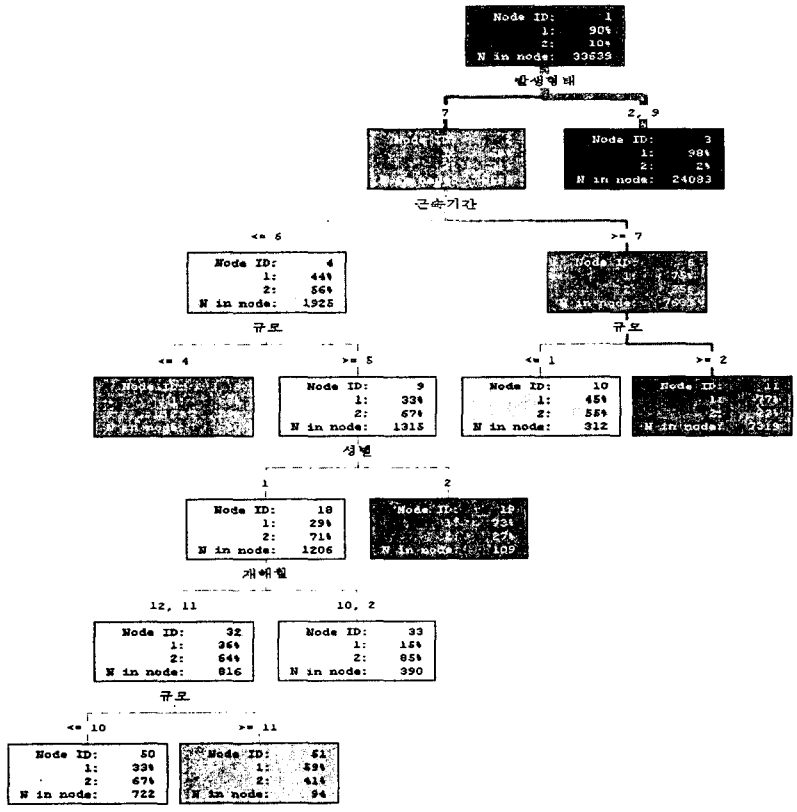
<표 2> 누적이익도표

모델명	백분위수	Captured Response Rate(%)	Response Rate(%)	Lift Value
NN	10	60.9888	64.4383	6.0989
	20	79.9967	41.6048	3.9998
	30	90.0543	31.2237	3.0018
LR	10	44.3153	41.0952	4.4315
	20	66.2398	34.4501	3.3020
	30	86.3071	29.2945	2.8769
CHAID	10	57.5295	59.8400	5.7530
	20	74.3310	38.6581	3.7166
	30	85.4526	29.6281	2.8484
C4.5	10	57.0216	59.3117	5.7022
	20	74.3975	38.6927	3.7199
	30	86.0847	29.8473	2.8695
CART	10	54.3378	56.5201	5.4338
	20	71.0563	36.9550	3.5528
	30	86.3072	29.4402	2.8303

누적이익도표를 살펴보면 상위 10% 집단에서 NN이 64.44%의 응답률(Response Rate)을 보이고 있다. 이것은 상위 10% 집단에서 NN이 64%의 사망확률이 높은 근로자를 포함한다는 것을 의미한다. 이것은 NN이 다른 기법들인 LR(41%), CHAID(60%), C4.5(59%), CART(57%)보다 높은 예측의 정확성을 가진다는 것을 보여주고 있다. 상위 10% 뿐만이 아니라 상위 20%, 30% 부분에서도 NN이 가장 높은 정확성을 보여주고 있다.

3.3 의사결정나무를 이용한 신경망 근사모형 분석

신경망분석의 단점 중 하나는 얻어진 모형을 해석하기가 어렵다는 것이다. 이러한 문제를 해결 할 수 있는 방법 중 하나는 다른 형태의 비선형모형이라 할 수 있는 의사결정나무를 이용하여 신경망 예측모형에 대한 근사모형을 만들고 이를 통해 해석해 보는 것이다. 이렇게 얻어진 의사결정나무는 신경망 예측모형에 대한 근사모형(Approximation Model)이라고 할 수 있으며, 이 결과를 살펴봄으로써 신경망 예측모형을 부분적으로나마 이해할 수 있을 것이다[1]. 신경망의 근사모형을 만들기 위해서는 변수변형(Transform Variables) 단계를 거쳐야 한다. 변수변형 노드는 기존의 변수를 변환하거나 사용자의 지정에 의해 새로운 변수를 만들 수 있다. 이러한 변수변환은 변수의 안정화, 비선형성의 제거, 비정규성의 수정 등에 사용될 수 있다.



<그림 4> 신경망 예측모형에 대한 근사모형

위의 <그림 4>는 산업재해 예방을 위한 예측 분류성능에서 가장 뛰어난 성능을 보인 NN에 대한 근사모형을 보여주고 있다. 신경망 예측모형에 대한 근사모형의 노드들을 분석해보면 총 8가지의 유형을 파악할 수 있으며 그 노드들을 정리해보면 <표 3>과 같다. 노드들 중에서 가장 뚜렷한 유형을 보여주는 트리로는 유형1과 유형2가 있다. 유형1의 발생형태(교통사고, 충돌)일 경우 24,083건으로 가장 높은 사망유형을 보였고, 두 번째로 유형2의 발생형태 → 근속기간 → 규모의 경우를 살펴보면 총 7,319건으로 발생형태가 직업병, 근속기간이 10년 이상, 회사규모가 9명 이상인 사업장 근로자들에게서 많이 발생하는 것을 알 수 있다. 이 결과는 교통사고, 충돌이 사망에 가장 큰 영향을 주는 요인이라 할 수 있으며, 강원도의 지리적 특성상 10년 이상 장기근로자에게서 직업병(진폐 : 폐질환의 일종)의 비율이 상당히 높게 나타나고 있는 것을 알 수 있다.

<표 3> 신경망 예측모형에 대한 근사모형 노드

노드유형	노드 경로	depth
유형1	node1 → node3 발생형태(교통사고, 충돌)	1
유형2	node1 → node2 → node5 → node11 발생형태(직업병) → 근속기간(5년~10년 이상) → 규모(5~9인 이상)	3
유형3	node1 → node2 → node5 → node10 발생형태(직업병) → 근속기간(5년~10년 이상) → 규모(5인 미만)	3
유형4	node1 → node2 → node4 → node8 발생형태(직업병) → 근속기간(4~5년 이하) → 규모(16~29인 이하)	3
유형5	node1 → node2 → node4 → node9 → node19 발생형태(직업병) → 근속기간(4~5년 이하) → 규모(30~49인 이상) → 성별(여자)	4
유형6	node1 → node2 → node4 → node9 → node18 → node33 발생형태(직업병) → 근속기간(4~5년 이하) → 규모(20~49인 이상) → 성별(남자) → 재해월(2월,10월)	5
유형7	node1 → node2 → node4 → node9 → node18 → node32 → node50 발생형태(직업병) → 근속기간(4~5년 이하) → 규모(20~49인 이상) → 성별(남자) → 재해월(11월,12월) → 규모(500~999인 이하)	6
유형8	node1 → node2 → node4 → node9 → node18 → node32 → node51 발생형태(직업병) → 근속기간(4~5년 이하) → 규모(20~49인 이상) → 성별(남자) → 재해월(11월,12월) → 규모(1000~1999인 이상)	6

5. 결 론 및 추 후 연 구

본 연구에서는 산업재해 예측 및 분석을 위해 Data Mining 기법을 적용하였는데 여러 기법들을 비교하여, 가장 좋은 분류성능을 보인 신경망기법을 선정하였고, 신경망 분석의 모형화를 위하여 예측 및 분석이 용이한 근사모형을 제시하였다. 이 근사모형은 모형화가 어려운 신경망분석의 이해에 도움을 줄 것으로 사료된다.

신경망 예측모형에 대한 근사모형의 노드를 보면 가장 큰 비중을 차지하는 유형1의 경우 발생형태가 교통사고, 충돌일 경우 사망 확률이 가장 높게 나타났으며, 유형2의 경우 근속기간이 10년 이상의 장기근로자와 근로자가 9명 이상의 사업장에서 직업병

(진폐)의 발생 확률이 높게 나타난다는 것을 알 수 있었다. 이에 강원도의 사업장에서는 교통사고와 충돌을 방지할 수 있는 예방대책과 근로자가 9명 이상의 사업장을 대상으로 10년 이상의 근로자에 대한 직업병을 방지할 수 있는 예방대책이 이루어 져야 할 것으로 사료되며, 추후 여러 가지 직업군별 재해방지를 위한 예방대책에 관한 연구가 진행되어야 할 것이다.

6. 참 고 문 헌

[1] 강현철, 한상태, 최종우, 김은석, 김미경, "SAS Enterprise Miner 4.0을 이용한 데이터마이닝 방법론 및 활용", 자유아카데미, 2001

[2] 김종현, "우리나라 산업재해 통계를 이용한 재해실태분석과 통계제도의 개선방향", 경일대학교 석사학위논문(1998) :40-60.

[3] 노동부, "산업재해현황분석", (2004)

[4] 배화수, 조대현, 석경하, 김병수, 최국렬, 이종언, 노세원, 이승철, 손용희, "SAS Enterprise Miner를 이용한 데이터 마이닝", 교우사(2005)

[5] 백귀훈, "의사결정나무를 이용한 취업고객분석 및 예측", 성균관대학교 석사학위논문(2002)

[6] 이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구", 한국지능정보시스템학회논문지 제9권 1호(2003)

[7] 임영문, 황영섭, 최요한, "데이터마이닝 기법을 활용한 산업재해자들에 대한 요인분석", 대한안전경영과학회지 제7권 4호(2005)

[8] 조윤정, "데이터마이닝을 이용한 종합건강진단센터의 데이터베이스 마케팅에 관한 연구", 서울대학교 보건대학원 보건학석사학위논문(2001) :53-56

[9] 최종우, 한상태, 강현철, 김은석, 김미경, 이성건, "SAS Enterprise Miner 4.0을 이용한 데이터마이닝 기능과 사용법", 자유아카데미(2001)

[10] Mevlut Ture, Imran Kurt, A. Turhan Kurum and Kazim Ozdamar, "Comparing classification techniques for predicting essential hypertension", *Expert Systems with Applications*, Volume 29, Issue 3(2005) :583-588

[11] Seung Hee Ho, Sun Ha Jee, Jong Eun Lee and Jong Sup Park, "Analysis on risk factors for cervical cancer using induction technique", *Expert Systems with Applications*, Volume 27, Issue 1(2004) :97-105

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.

유 창 현 : 현재 강릉대학교 산업공학과 대학원 석사과정에 재학 중이며 관심분야는 데이터마이닝, 알고리즘 분석 및 활용 등이다.

저 자 주 소

임 영 문 : 서울시 서초구 서초4동 아크로비스타 C동 910호

유 창 현 : 경기도 연천군 전곡읍 전곡 4/9 310-30번지