산업재해 데이터의 분석 및 분류를 위한 정확도 성능 평가

Evaluation on Performance of Accuracy for Analysis and Classification of Data Related to Industrial Accidents

임영문[1], 유창현[2]
Leem Young Moon, 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 performance of algorithms for data analysis of industrial accidents and this paper provides a comparative analysis 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. In this study, data on 67,278 accidents were analyzed to create risk groups for a number of complications, including the risk of disease and accident. The sample for this work chosen from data related to manufacturing industries during three years (2002~2004) in korea. According to the result analysis, NN has excellent performance for data analysis and classification of industrial accidents.

Keyword : Data Mining Techniques, ROC Chart, Lift Chart, Response Threshold

1. 서론

오래전부터 산업재해에 관련된 데이터는 국가기관을 중심으로 수집 보관 되어오고 있다. 업종별 유형별 산업재해에 관련된 데이터를 적절하게 분석하고 그 분석결과를 잘 활용할 수 있다면 상당량의 재해를 감소시킬 수 있을 것이라 생각된다. 그러나 기존의 산업재해 연구 자료들은 대부분 산업재해 데이터를 토대로 빈도 분석, 비교

† 본 연구는 한국과학재단 목적기초연구(R01-2003-000-00158-0)지원으로 수행되었음.
1) 강릉대학교 산업공학과 교수
2) 강릉대학교 산업공학과 석사과정

분석에만 의존 하여[1], 사후 관리적, 교육적인 결과들을 제시하고 있다. 이에 본 연구에서는 다양한 분야에서 데이터 분석, 분류, 예측 등에 활용되고 있는 데이터 마이닝(Data Mining) 기법을 산업재해 관련 데이터의 분석[5] 및 분류에 활용하고자 한다. 데이터 마이닝 기법은 대량의 과거 데이터로부터 자료의 예측 가능하다. 기존의 데이터 마이닝 적용분야를 보면 이동통신사의 이탈 고객 예측모형, 취업고객 분석 및 예측 모형, 의학적 진단 예측 모형 등으로 활용되어져 왔다[3,4,6,8,9]. 본 연구에서는 강원도 내 전 업종에서 발생한 재해자 총 67,278명의 자료를 바탕으로 재해형태인 사망 및 부상 예측을 위하여 의사결정나무(Decision Tree) 알고리즘인 CHAID, C4.5, CART와 로지스틱회귀모형(Logistic Regression), 신경망(Neural Network)등의 다양한 기법을 적용하여 결과를 비교 분석하여 최적의 모형을 제시하고자 한다.

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 구성되었다.

2.2 분석방법

본 연구는 재해구분(사망, 사고)을 분류하기 위하여 크게 데이터 입력, 데이터 분할, 변수 선택, 모형화, 평가 단계로 진행하였다. 데이터 분할단계에서는 데이터의 검증을 위하여 Training Set 과 Testing Set 데이터의 비율을 50:50으로 분할하였고, 변수 선택 단계에서는 효율적인 입력변수 선정을 위하여 카이제곱 통계량을 이용하여 입력변수를 선택하였다. 모형화 단계에서는 여러 모형들을 상정한 후 결과를 비교하여 최적의 모형을 선택할 수 있는 지표를 제시하였다. 평가단계에서는 오분류표를 이용한 각 예측모형들의 정분류율(Accuracy), 오분류율(Error Rate) 값을 측정하여 수치를 비교 하고, 분류기준 값의 변화에 따라 민감도와 특이도를 고려한 예측의 정도를 나타내는 ROC Chart와 사후확률을 이용하여 예측의 정확성을 알 수 있는 Lift Chart, 모형의 일치성 판별을 위한 Response Threshold 도표를 이용하여 모델을 비교, 평가 하였다. 분석도구로는 SAS Enterprise-Miner 4.3을 이용하였다.

3. 분석결과

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 보다 적은 나이, 재해시간을 제외한 재해구분, 발생형태, 업종, 규모, 성별, 근속기간, 재해월, 재해요일로 결정 되었으며, 이들 변수들 중에서 재해구분, 성별은 이산형(Binary) 변수로, 발생형태, 업종, 재해월, 재해요일은 명목형(Nominal) 변수로, 규모, 근속기간은 연속형(Interval) 변수로 구성하였다.

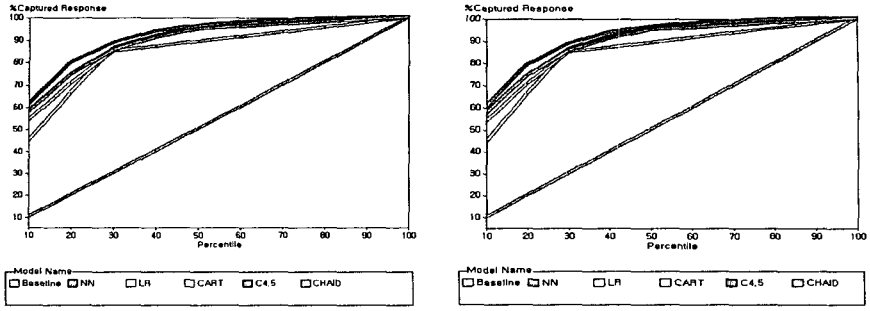
3.2 모델별 결과 비교

데이터 마이닝 모델들을 비교 분석하기 위하여 모델별로 정분류율, 오분류율을 분류표(Classification Tables)를 이용하여 계산하였다. 분류표란 목표변수의 실제범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 것으로 E-Miner를 사용하여 구해진 모델별 비교표는 <표 1>과 같다.

<표 1> 모델별 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%

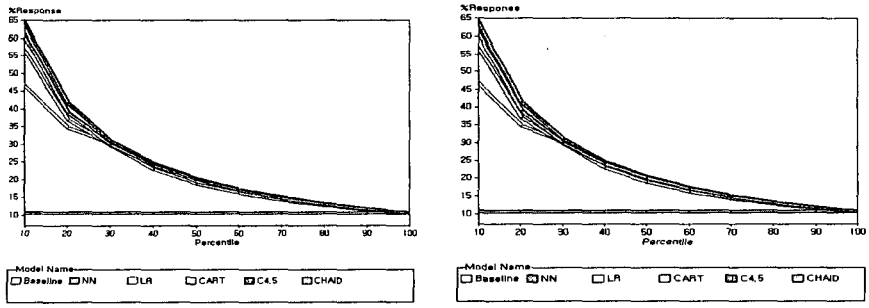
위의 <표 1>에서 볼 수 있듯이 모델들을 비교해 보면 정분류율은 NN이 Training Set에서 93.09%로 가장 높은 분류율을 나타내었고, LR은 90.75%로 가장 낮은 분류율을 보였다. 오분류율에서도 NN이 6.91%로 가장 우수하였으며, 가장 나쁜 LR과 비교했을 때 2.3%가 차이가 났다. Testing Set 데이터에서도 NN이 정분류율 92.94%, 오분류율 7.06%로 가장 좋은 성능을 나타내었다. 분류표에 의한 종합적인 결과를 보면 전체적으로 NN가 높은 분류 결과를 나타냈다.



(a) Training set (b) Testing set

<그림 1> 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 상단으로 더 가까운 모형을 성능 면에서 우수한 모형으로 판단하면 된다[7]. 위의 <그림 1>은 ROC Chart의 결과 모형을 보여주고 있다. ROC Chart에서 나타나는 결과를 보면 NN이 다른 모형들에 비해서 Training data 와 Testing data 모두에서 정확한 결과를 예측할 수 있는 모형으로 판단된다.



(a) Training set (b) Testing set

<그림 2> Lift Chart

Lift Chart는 사후확률을 이용하여 예측의 정확성을 알 수 있게 한다. Lift Chart는 각각의 관측치에서 사후확률을 구한 후 사후확률의 크기순서에 따라 전체 자료를 균일하게 N등분한 후 각 집단에서의 %Captured Response, %Response 그리고 Lift를 계산한다. 각각의 의미는 다음과 같다[2].

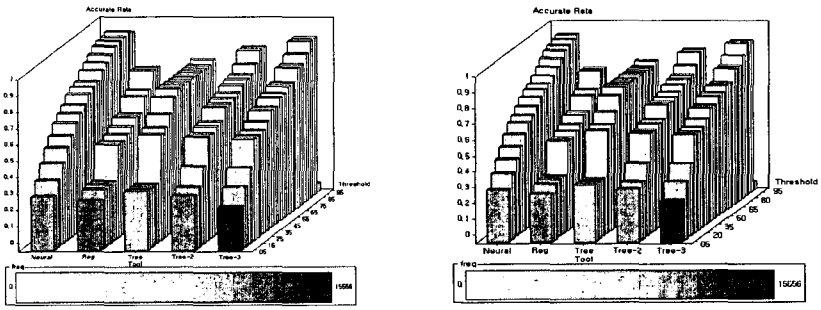
$$\%\text{Captured Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{전체에서 목표변수의 특정범주 빈도}} \times 100$$

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

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

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

위 <그림 2>의 Lift Chart를 보면 5가지 모형중에서 사망 가능성이 높은 상위 10%의 집단에서 NN이 62.87% 정도의 정확성을 가진다는 것을 보여주었고, 다음으로 C4.5가 60.35%, CHAID가 59.84%, CART가 55.95%, LR이 46.1%순으로 정확성을 가지는 것으로 나타났다.



(a) Training set (b) Testing set

<그림 3> Response Threshold 도표

Response Threshold 도표는 분류기준 값이 달라짐에 따라 모형의 분류결과가 얼마나 변동이 없는가를 알기 위해서 사용한다. 변화에 따른 모형분류결과의 변동여부를 모형의 일치성(Consistency)이라고 하는데 <그림 3>에서 첫 번째 모형인 NN을 보면 분류기준 값이 조금 변동함에 따라 분류결과가 달라지는 것을 볼 수 있는데 다른 모형들은 분류값이 변함에도 양극단으로 치우쳐져 있는 것을 볼 수 있다. 이러한 분류결과의 일치성 여부는 정확도와 직결된다. Response Threshold 도표를 비교했을 때 NN이 가장 뛰어난 일치성을 보였다.

5. 결론 및 추후연구

본 연구에서는 산업재해를 예측하기 위하여 Data Mining 기법을 이용하여 산업재해 예측모형을 비교하였고, 최적의 모델을 제시하였다. 분석결과를 보면 본 연구에서 언

급된 알고리즘 모두 전체적으로 뛰어난 분류값을 보여주고 있었다. 그중에서 Neural Network가 정확도와 오분류율에서 뛰어난 분류값을 보여 산업재해 예측에 가장 적합한 모델이라 사료된다.

본 연구에서는 분류표, ROC Chart, Lift Chart, Response Threshold 도표를 이용한 정분류율, 오분류율, 민감도, 특이도를 기반으로 비교를 하였다. 추후 산업재해 예측에서 뛰어난 분석 성능을 보인 Neural Network의 특성을 고려한 다양한 평가방법에 관한 연구가 필요할 것으로 사료되며, 좀더 장기적인 다량의 데이터를 기반으로 다양한 연구가 이루어 져야할 것이다.

5. 참고문헌

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

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

[3] 백귀훈, "의사결정나무를 이용한 취업고객분석 및 예측", 성균관대학교 석사학위논문, 2002.

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

[5] 임영문, 황영섭, 최요한, "데이터마이닝 기법을 활용한 산업재해자들에 대한 요인분석", 대한안전경영과학회지 제7권 4호, 2005.

[6] 조윤정, "데이터마이닝을 이용한 종합건강진단센터의 데이터베이스 마케팅에 관한 연구", 서울대학교 보건대학원 보건학석사학위논문, pp. 53~56, 2001.

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

[8] 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, pp. 583~588, 2005.

[9] 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, pp. 97~105, 2004.