데이터 마이닝을 이용한 산업재해예측 알고리즘의 비교

대한산업안전협회 김성대 차장

I. 서론

연구배경 및 목적

우리나라는 산업화와 함께 급속한 경제성장을 이루었다. 산업화가 시작된 이래 사업장의 증가, 신종업종 출현, 근로자수의 증가로 인해 수많은 유형의 산업재해에 의한 매년 많은 재해자가 발생하고 있다.

산업재해 예방을 위한 연구는 1990년대 말부터 조금씩 시도되고 있으나 아직까지 큰 진전이 없는 실정이다. 이미 발생된 산업재해 발생자료를 이용하여 데이터 마이닝의 대표적 기계학습 알고리듬인 선형회귀분석(linear regression), 신경망(neural networks), C4.5, CHAID, QUEST 등에 적용, 분류성능을 비교한 연구는 주목할 만하다. 그러나, 현재까지 추가적인 기계학습 알고리듬에 대하여 분류성능을 비교한 연구가 없고, 예측성능의 주요 평가지표인 시계열 분석에 대한 성능평가가 없는 실정이다.

따라서, 본 논문은 (사)대한산업안전협회에서 2000년부터 2006년까지 7년간 집계한 산업재해 발생에 관한 자료를 데이터 마이닝 기법 중 널리 알려졌으나 국내에서 분류성능에 대한 검증작업이 아직 이루어지지 않은 기계학습 알고리듬인 BN, LR, SVM, C4.5, RF, MLP, RBFN 등에 확대 적용하고 시계열 분석 평가를 실험하여 각 알고리듬별 국내 산업재해 발생자료에의 적용가능성을 검토하여 보다 현실적인 산업재해 예방 제고 및 재해율 감소에 기여하는 것에 그 목적이 있다.

II. 관련연구

산업재해보상보험제도가 1964년 도입되고 산업재해조사가 1977년 국가통계로 지정된 이래 국가 및 민간단체 차원으로 대량의 산업재해 자료를 축적 및 관리해 오고 있다.

그간의 산업재해 연구는 과거에 발생한 산업재해에 대하여 현황 또는 그 특성을 분석한 것이 대부분이었으며, 산업재해 예측에 관한 연구는 비교적 근래인 1990년대 말부터 조금씩 시도되어 오고 있다. 산업재해 예측에 관한 대표적인 연구로는 연도별 산업재해 도수율, 강도율, 업종별 재해자수 등의 속성에 대하여 선형회귀분석(linear regression) 등의 통계기법을 적용하여 시간에 따른 추세를 예측한 것과 과거 발생한 산업재해 발생자료를 의사결정트리(decision tree) 등의 데이터 마이닝의 기계학습 알고리듬을 적용하여 그 성능을 비교분석한 것이 있다.

본 논문에서 초점을 두고 있는 데이터 마이닝은 매우 큰 대용량의 데이터로부터 자동적 혹은 반자동적인 방법을 이용하여 이들 데이터(data) 내에 존재하는 관계(associations), 패턴(patterns), 규칙(rules) 등을 탐색하고 찾아내어 사전에 알지 못했으나 의미있는 정보(inforamtion)를 추출하여 의사결정 등에 사용하는 과정들을 의미한다. 데이터 마이닝의 기법은 ID3, CHAID, CART, C4.5, QUEST 등의 의사결정트리(decision trees), MLP, RBFN, Kohonen Networks 등의 신경망(neural networks), 장바구니분석, 순차패턴 등의 연관규칙(association rules), k-Means, k-Nearest Neighbor 등의 군집분석(clustring), 선형회귀(linear regression), 로지스틱회귀(logistic regression), 시계열분석 등의 통계적 기법(statistical methods)으로 대별된다.

본 장에서는 상기 데이터 마이닝의 기법 중 의사결정트리(decision trees)와 신경망(neural networks)의 개념을 정리하고, 몇 가지 대표적인 기계학습 알고리듬에 대하여 기술한다.

1. 의사결정트리(Decision Trees)

의사결정트리는 일련의 규칙에 따라 모집합을 몇 개의 하위 집합으로 구성된 모형(model)을 구축하여 분류(classification)하거나 예측(prediction)을 목적으로 사용되는 전통적인 분석방법이다.

분석과정이 뿌리, 가지, 잎의 나무구조로 표현되기에 판별분석(discriminant analysis) 회귀분석(regression analysis), 신경망(artificial neural networks) 등의 방법에 비해 월등히 높은 해석력을 갖고 있다는 특성을 가진다. 이러한 특성으로 인해 결과의 정확도보다 분석과정의 설명이 필요한 경우에 더욱 유용하게 사용되기도 한다.

일반적인 의사결정트리 분석은 다음과 같은 절차를 가지며, 각 절차 중 정지기준, 분리기준, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정트리가 형성된다.

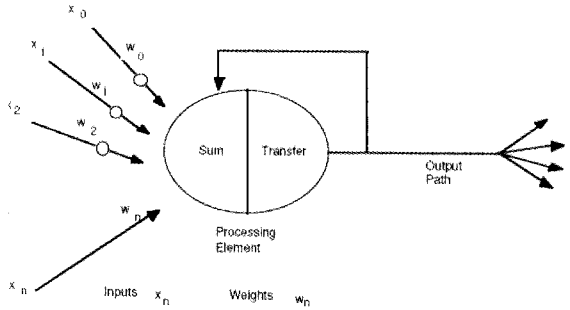
- 의사결정트리의 형성 : 분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)을 지정하여 의사결정트리를

구축한다.

- 가지치기 : 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 규칙을 가지고 있는 가지(branch)를 제거한다.
- 타당성 평가 : 이익도표(gain chart)나 위험도표(risk chart) 또는 검정용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정트리를 평가한다.
- 해석 및 예측 : 의사결정트리를 해석하고 분류 및 예측모형을 설정한다.

2. 신경망(Neural Networks)

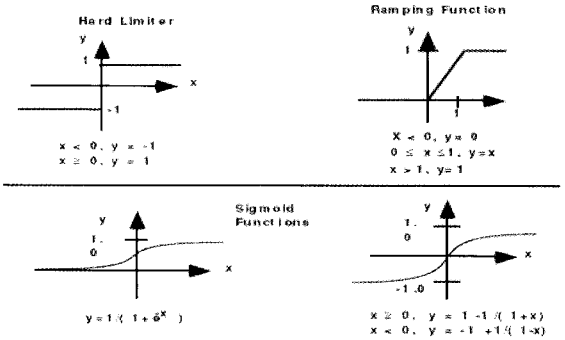
신경망은 다양한 응용분야 적용되어 성공적인 결과를 보여 왔으며, 고전적인 문자인식, 음성인식, 영상인식, 로봇제어 등의 공학적 응용 이외 최근에는 경제예측, 재무분석, 금융·신용조회, 기업도산 예측 및 주가예측 등의 광범위한 사회경제 분야에서도 응용되고 있다.



[그림 Ⅱ-1] 인공 뉴런 모델

신경망 이론은 1943년 신경생리학자 McCulloch와 수학자 Walter Pitts에 의해 인간의 사고활동에 관여하는 수많은 신경세포 뉴런(neuron) 중 하나를 [그림 Ⅱ-1] 같이 수학적 모델로 제시함으로써 태동하였으며, 1970년대 초 Werbor 등이 계층형 신경망에 대한 순방향 역전파 구조(feed-forward back-propagation architecture)를 제안함으로써 현대 신경망 이론이 정립되었다. 신경망은 다양한 외부 입력에 대한 정보처리를 수행하는 이 인공 뉴런들의 다수가 연결되어 상호작용하는 있는 구조를 갖는 것이다.

신경망 내의 각 인공 뉴런은 직접적인 외부 입력뿐만 아니라 특정 뉴런의 출력을 입력으로 재귀입력 또는 연결된 다른 인공 뉴런의 출력을 입력으로써 수용할 수 있다.



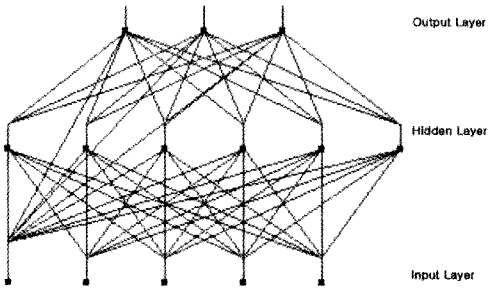
[그림 Ⅱ-2] 대표적인 전송 함수의 예

인공 뉴런의 출력은 각 입력값과 연결 가중치(connection weight)의 곱을 합한 후 [그림 Ⅱ-2]의 전송 함수(transfer function)에 의하여 계산된 결과이다. 인공 뉴런의 출력은 수식 (1) 으로 나타낼 수 있다.

$$O_n = T(\sum(X_i W_i - \Theta_i)) \quad \cdots\cdots \text{수식 (1)}$$

여기서 On은 인공 뉴런 n의 출력, T은 전송 함수, X은 입력값, W은 해당 입력에 대한 가중치, θ은 임계값(threshold)을 나타낸다.

대표적 계층형 신경망인 다층 구조 퍼셉트론(Multi Layer Perceptron, MLP)은 [그림 Ⅱ-3]와 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)을 갖는 3-계층 구조를 갖으며, 은닉층은 1개 또는 2개를 가질 수 있다. 데이터의 흐름은 예측 동작(recall mode)일 때 순방향(forward)이고, 학습 동작(learning mode)일 때 역방향(backward)이 된다.



[그림 Ⅱ-3] 계층형 신경망 구조

임의의 인공 뉴런 j은 다른 뉴런의 출력값 Oi을 입력받아 수식 (2)에 의해 총합 Sj를 계산한 후, 수식(3)에 의해 출력값 Oj을 계

산한다.

$$S_j = \sum(O_i W_{ij} - \Theta_j) \quad \cdots\cdots \text{수식 (2)}$$

$$O_j = \frac{1}{(1+e^{-S_j})} \quad \cdots\cdots \text{수식 (3)}$$

여기서 Sj은 총합, Oj 인공 뉴런 j의 출력값, Oi은 다른 인공 뉴런 i의 출력값, Wij은 인공 뉴런 i와 인공 뉴런 j간의 연결 가중치, 그리고 θj은 임계값을 나타낸다.

Ⅲ. 실험 및 고찰

1. 분석대상 자료

본 논문에 사용된 자료는 (사)대한산업안전협회에서 2000년 1월부터 2006년 12월까지 산업현장에서 발생한 산업재해를 기록한 84개월 분(7년) 16,556건의 자료이다. 아래 [표 Ⅲ-1]은 산업재해수를 연도별 · 업종별로 정리한 것이다.

[표 Ⅲ-1] 연도별 · 업종별 산업재해 집계 건 수

업종 \ 연도	2000	2001	2002	2003	2004	2005	2006	계
광업	6	7	8	14	9	2	5	51
제조업	1,773	1,656	1,970	2,404	2,022	1,678	1,753	13,256
전기가스 및 상수도업	0	1	3	8	0	5	4	21
건설업	1	0	0	2	1	0	0	4
운수창고 및 통신업	74	71	155	198	195	181	208	1,082
임업	4	1	1	17	7	5	2	37
어업	0	0	0	0	0	0	2	2
농업	0	0	0	0	0	2	5	7
기타의 사업	156	173	330	448	349	311	329	2,096
금융 및 보험업	0	0	0	0	0	0	0	0
계	2,014	1,909	2,467	3,091	2,583	2,184	2,308	16,556

이 자료는 지역, 사업장명, 업종대분류코드, 업종중분류코드, 업종대분류명, 업종중분류명, 근로자수, 사업장규모, 고용형태, 근무형태, 동시작업인원, 안전장치사용 작업여부, 재해자명, 성별, 연령, 학력, 근속기간, 재해발생시기, 월, 일, 요일, 시간대, 재해구분, 발생형태, 상해종류, 상해부위, 휴업예상일수 등 30개 속성로 구성되어 있다.

이 중 문자열로 구성되어 분석에 불필요한 사업장명, 업종대분류명, 업종중분류명, 재해자명과 다수의 결측치(missing value)를 포함하는 성별, 연령, 학력, 일, 요일, 발생형태를 제외한 지역, 업종대분류코드, 업종중분류코드, 근로자수, 사업장규모, 고용형태, 근무형태, 동시작업인원, 안전장치사용 작업여부, 근속기간, 재해발생시기, 월, 시간대, 재해구분, 상해종류, 상해부위, 휴업예상일수를 등 17개 속성을 선정하였다. 각 속성의 값은 별도 정의된 코드를 부여하여 가공하였다.

2. 실험방법

데이터 마이닝의 주요 기계학습 알고리듬들의 산업재해 자료에 대한 분류성능을 비교하기 위해 본 논문에는 (사)대한산업안전협회에서 2000년부터 2006년까지 집계한 총 16,556건의 산업재해 자료 중 높은 발생 빈도를 차지하는 제조업(80%), 운수、창고 및 통신업(6.5%), 기타의 사업(12.6%) 업종으로 한정한 총 16,434건을 실험자료로 사용하였다.

[표 Ⅲ-] 연도별 · 업종별 실험자료

업종 \ 연도	2000	2001	2002	2003	2004	2005	2006	계
제조업	1,773	1,656	1,970	2,404	2,022	1,678	1,753	13,256
운수창고 및 통신업	74	71	155	198	195	181	208	1,082
기타의 사업	156	173	330	448	349	311	329	2,096
계	2,003	1,900	2,455	3,050	2,566	2,170	2,290	16,434

실험자료는 전처리 과정을 통해 선정된 [표 Ⅲ-2]의 속성을 사용하였고 이 중 목표 속성을 휴업예상일수로 선택하였다. 실험은 SUN B2000 머신과 기계학습 공개 소프트웨어인 Weka 3.4.13을 이용하여 기계학습 알고리듬인 BN(Bayes Network), LR(Logistic Regression), SVM(Supportj Vector Machine), C4.5, RF(Random Forest), MLP(Multi-Layer Perceptron), RBFN(Radial Basis Function Network) 에 적용하여 분류성능 비교실험을 수행하였다.

분류성능의 실험은 연도별 평가와 시계열 평가의 두 가지 방식을 실시하였다.

각 알고리듬별 실험시 사용된 조건은 아래 [표 Ⅲ-4]와 같다.

[표 Ⅲ-] 기계학습 알고리듬별 실험조건

알고리듬	조 건
BN	weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 — -P 3 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator — -A 0.5
LR	weka.classifiers.functions.Logistic -R 1.0E-8 -M 100
SVM	weka.classifiers.functions.LibSVM -S 0 -K 3 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1

알고리듬	조 건
C4.5	weka.classifiers.trees.J48 -U -M 2
RF	weka.classifiers.trees.RandomForest -I 50 -K 0 -S 1
MLP	weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 1000 -V 0 -S 0 -E 20 -H t -G -B -R
RBFN	weka.classifiers.functions.RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1

3. 실험결과

(1) 연도별 분류성능 평가

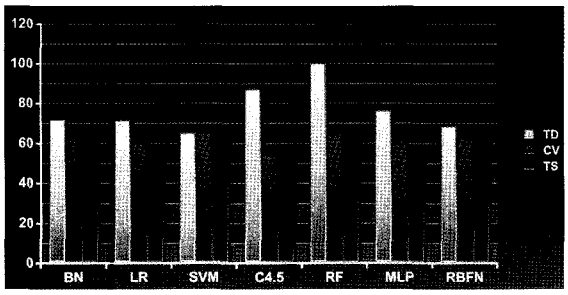
연도별 분류성능 평가는 2000~2006년 각 연도별 실험자료를 훈련자료로 이용하여 구축한 모형(model)에 대하여 훈련자료(trained data) 실험, 교차타당성 실험(cross validation) 및 분할 실험(test split)을 실시하였다. 훈련자료 실험은 모형구축에 사용된 훈련자료를 검증자료로써 재이용한 것이고, 교차타당성 실험시 조건은 전체 실험자료를 임의로 섞은 10개의 집합으로 구성한 뒤, 특정 집합의 원소를 이용하여 훈련하고, 다음 집합의 원소를 이용하여 검증하도록 10 폴드로 설정했으며, 분할 실험 시 조건은 실험자료를 훈련:검증 = 7:3 로 분할하여 10회 실행한 후 평균 정확도를 산출하였다. 또한, 모든 실험결과 수치는 소수점 셋째 자리에서 반올림하였다.

[그림 Ⅲ-1]은 각 알고리듬에 대하여 연도별 분류성능 실험한 결과의 평균값을 나타낸 것이다. 실험결과 전반적으로 알고리듬의 훈련자료 실험의 결과수치가 여타 교차타당성 실험 및 분할 실험에 비해 상대적으로 높게 산출된 것을 알 수 있다.

이는 훈련자료에 대한 과잉적합(over-fitting)의 전형적인 예라고 볼 수 있겠으며, RF 알고리듬의 결과수치는 특이할 만 하다. 훈련자료 실험과 분할 실험의 결과수치의 차이가 적은 알고리듬은 SVM, RBFN, BN, LR, MLP, C4.5, RF 순이며, SVM 알고리듬의 경우 결과 차이가 거의 없었다.

교차타당성 실험과 분할 실험의 결과수치는 모든 알고리듬이 비슷한 수준의 양상을 보였으며, 이는 학습에 사용되는 훈련자료가 적을 경우 보편적으로 사용되는 알고리듬 분류성능의 척도가 되는 교차타당성 실험의 유효성을 재확인한 것이라 볼 수 있겠다.

분할 실험의 결과를 보았을 때 각 알고리듬별 성능은 SVM, RF, BN, RBFN, LR, MLP, C4.5 순이다.



[그림 Ⅲ-1] 연도별 분류성능 평가

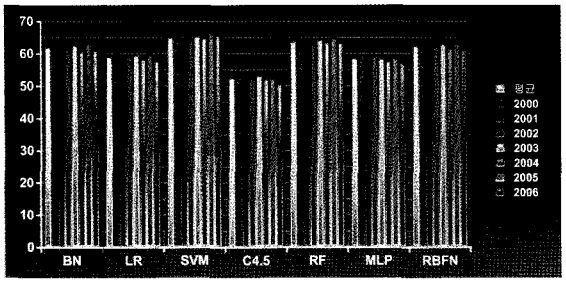
(2) 시계열 분석에 의한 분류성능 평가

시계열 분석에 의한 분류성능 평가는 크게 두 가지로 실험하였다.

첫째, 각 연도 1년분 산업재해 발생자료를 훈련자료로써 각 알고리듬에 학습시킨 뒤 타년도 산업재해 발생자료를 검증자료로써 적용하여 실험을 하였다. 이 실험은 단기간의 산업재해 발생자료를 가지고, 알고리듬별 가까운 미래 혹은 과거 시점의 예측력을 평가하는 것이다.

둘째, 2000년부터 차년도 산업재해 발생자료를 1년 단위로 누적시킨 훈련자료를 각 알고리듬에 학습시킨 뒤, 훈련자료 이후의 차년도 산업재해 발생자료를 검증자료로써 적용한 실험을 하였다. 이 실험은 산업재해 발생자료의 학습량에 따라 알고리듬별 미래 시점의 예측력을 평가하는 것이다.

따라서, 이 두 실험은 각 알고리듬이 산업재해 발생자료에 대하여 각각 단기적 및 장기적인 시계열 특성을 어느 정도 고려할 수 있는지 비교가 되는 실험이 되겠다.

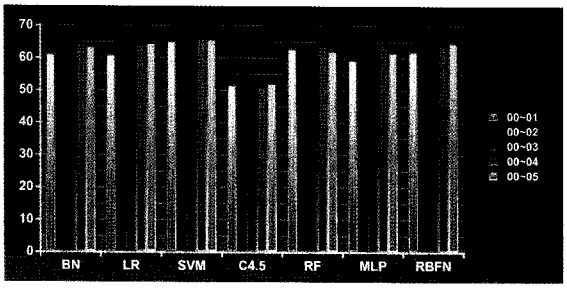


[그림 Ⅲ-2] 시계열 분석에 의한 분류성능 평가

상기 [그림 Ⅲ-2]은 시계열 분석에 의한 분류성능 평가실험 중 첫 번째 실험의 결과를 도표화한 것이다. 실험결과의 평균을 보았을 때, 각 알고리듬별 성능은 SVM, RF, RBFN, BN, LR, MLP, C4.5 순이다.

특이할 만 한 점은, 모든 알고리듬에 훈련자료의 연도 시점에서 멀어지는 연도의 검증자료를 적용하였을 때 분류정확도가 조금씩 하락하는 경향을 보이는 것이다. 이러한 결과는 사회적인

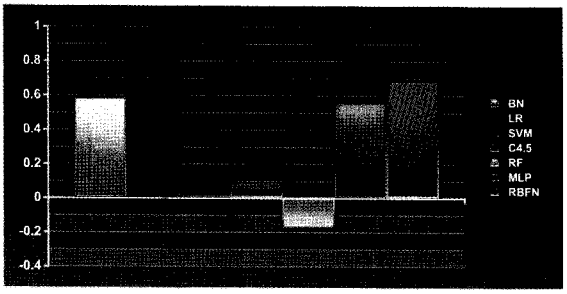
요인에 의해 산업재해 발생의 특성 및 원인이 변하는 것임을 유추할 수 있겠고, 장기간 훈련자료가 학습에 필요하다는 것을 나타내는 좋은 예라고 볼 수 있겠다.



[그림 Ⅲ-3] 시계열 분석에 의한 분류성능 평가

상기 [그림 Ⅲ-4]은 시계열 분석에 의한 분류성능 평가실험 중 두 번째 실험의 결과를 도표화한 것이다. 실험결과 산업재해 발생자료의 학습량에 따라 전체적인 알고리듬의 분류정확도가 향상되는 것을 볼 수 있다. 학습량이 가장 많은 6년간의 훈련자료(2000~2005년)을 사용한 경우의 알고리듬별 평균 정확도는 SVM, RBFN, LR, BN, RF, MLP, C4.5 순이다.

알고리듬 중 BN, RF, C4.5, MLP는 편차가 크고, 이 중 C4.5 알고리듬의 경우 학습량의 증가에도 분류정확도의 향상이 다른 알고리듬에 비해 크지 않았다.



[그림 Ⅲ-4] 시계열 분석에 의한 분류성능 평가

상기 [그림 Ⅲ-4]은 훈련자료의 학습량에 따른 분류정확도의 상승량의 평균을 도식화한 것으로, 학습율이 높은 알고리듬은 LR, RBFN, BN, MLP, SVM, C4.5, RF 순이다.

통계기법의 알고리듬인 BN, LR 및 신경망 알고리듬인 MLP, RBFN의 학습율이 결정트리 알고리듬인 C4.5, RF 알고리듬에 비해 월등히 높았다. 특히, SVM 알고리듬은 전체실험에 대한 평균 정확도가 가장 우수했으나 학습량에 비례한 학습율이 저조한 편이고, RF 알고리듬은 학습율의 편차가 커서 평균수치가 음수를 갖는 결과를 보였다.

Ⅳ. 결론

본 논문에서는 산업재해 발생자료에 대하여 데이터 마이닝의 주요 기계학습 알고리듬에 적용하여 휴업예상일수를 대상으로 하는 분류성능, 즉 예측률을 비교 분석하였다.

연도별 분류성능 평가는 분할 실험의 결과를 보았을 때, 각 알고리듬별 성능은 SVM, RF, BN, RBFN, LR, MLP, C4.5 순이다.

시계열 분석에 의한 분류성능 평가실험 중 첫 번째 실험의 결과의 평균을 보았을 때, 각 알고리듬별 성능은 SVM, RF, RBFN, BN, LR, MLP, C4.5 순이다. 특히, 모든 알고리듬에 훈련자료의 연도 시점에서 멀어지는 연도의 검증자료를 적용하였을 때 분류정확도가 하락하는 경향을 보였다. 이는 사회적인 요인에 의해 산업재해 발생의 특성 및 원인이 변하는 것임을 유추할 수 있겠다.

장기간 훈련자료가 학습에 반영될 경우 미치는 영향을 실험한 시계열 분석에 의한 분류성능 평가실험 중 두 번째 실험의 결과 산업재해 발생자료의 학습량에 따라 전체적인 알고리듬의 분류정확도가 향상되는 것을 볼 수 있다. 학습량이 가장 많은 6년간의 훈련자료(2000~2005년)을 사용한 경우의 알고리듬별 평균 정확도는 SVM, RBFN, LR, BN, RF, MLP, C4.5 순으로, 통계기법의 알고리듬인 BN, LR 및 신경망 알고리듬인 MLP, RBFN의 학습율이 결정트리 알고리듬인 C4.5, RF 알고리듬에 비해 월등히 높았다. 훈련자료의 학습량에 따른 분류정확도의 상승률을 평가했을 때, 학습율이 높은 알고리듬은 LR, RBFN, BN, MLP, SVM, C4.5, RF 순이다. 알고리듬 중 BN, RF, C4.5, MLP는 편차가 컸고, 최근의 산업재해 예측에 관한 연구에 많이 사용하는 의사결정트리의 대표적 알고리듬인 C4.5의 경우 학습량의 증가에도 분류정확도의 향상이 다른 알고리듬에 비해 매우 저조했으며 이를 보완하기 위한 추가적인 연구가 필요하다.

전통적으로 예측모형으로 가장 많이 활용하고 있는 신경망 알고리듬 중 MLP은 높은 분류 정확도를 기대했던 실험 전의 예상과 달리 전체적인 실험결과가 저조했다. 이는 신경망의 특성상 조정해야 할 파라미터의 수가 너무 많고, 특히 MLP의 경우 구조(입력층 노드수, 은닉층 노드수 및 수준)에 많은 영향을 받기 때문이라 짐작할 수 있으며, 본 연구에 사용된 산업재해 발생자료에 대해 적합한 MLP의 구조를 찾을 필요가 있다.

전반적인 실험결과를 보았을 때, SVM 알고리듬은 학습량에 비례한 학습율이 비록 저조한 편이었으나, 전체실험의 평균 정확도가 가장 우수했기에, 향후 산업재해 발생예측에 있어 통계기법과 신경망을 대체할 수 있는 유용한 기법임을 확인할 수 있었다.