건설업의 산업재해 특성분석을 위한 의사결정나무 기법의 상용 최적 알고리즘 선정

- Selection of an Optimal Algorithm among Decision Tree Techniques for Feature Analysis of Industrial Accidents in Construction Industries -

임 영 문 *
Leem Young Moon
최 요 한 **
Choi Yo Han

Abstract

The consequences of rapid industrial advancement, diversified types of business and unexpected industrial accidents have caused a lot of damage to many unspecified persons both in a human way and a material way. Although various previous studies have been analyzed to prevent industrial accidents, these studies only provide managerial and educational policies using frequency analysis and comparative analysis based on data from past industrial accidents. The main objective of this study is to find an optimal algorithm for data analysis of industrial accidents and this paper provides a comparative analysis of 4 kinds of algorithms including CHAID, CART, C4.5, and QUEST. Decision tree algorithm is utilized to predict results using objective and quantified data as a typical technique of data mining. Enterprise Miner of SAS and AnswerTree of SPSS will be used to evaluate the validity of the results of the four algorithms. The sample for this work chosen from 19,574 data related to construction industries during three years (2002 ~ 2004) in Korea.

Keywords : Optimal Algorithm, Decision Tree, Feature Analysis, AnswerTree.

† 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

* 강릉대학교 산업시스템공학과 교수
** 강릉대학교 산업시스템공학과 박사과정 수료

2005년 12월접수; 2005년 12월 수정본 접수; 2005년 12월 게재확정

1. 서 론

1.1 연구 배경 및 목적

최근 노동부에서 발표한 2004년 우리나라 산업재해 발생 현황을 살펴보면 산재보험 적용사업장 1,039,208개소에서 10,473,090명의 근로자가 작업을 하던 중 산업재해가 발생하여 산재로 판정된 근로자수는 88,874명이며, 이중 2,825명이 사망한 것으로 나타났다. 그러나 산재보험 미적용사업장까지 확대하면 산업재해로 인한 인명손실은 이보다 훨씬 많을 것으로 쉽게 예측 할 수 있다. 한국의 산업 재해율은 60년대 4~5%에서 80년대 2~3%로 감소하여 95년에 최초로 1% 미만으로 진입하였다. 그러나 98년 산업재해율이 0.68%로 사상 최저치를 기록한 이후 지속적인 증가세를 보이기 시작해 지난해 산업 재해율은 98년 이후 최고 수준인 0.9%로 나타났다. 이는 일본보다 2.5배나 높은 수준으로 나타나는 등 여전히 선진국과 비교하여 높은 산업 재해율을 나타내고 있다. 이러한 결과는 기존의 연구들이 주로 산업재해와 관련된 통계자료의 재해 구성 비율 분석과 같은 빈도 분석[2][3][9]에만 의존하여 관리적, 기술적, 교육적 등의 재해예방 대책만을 제시한 것에 따른 결과라고 판단되어진다. 이러한 접근 방법은 매우 많은 분석 시간을 필요로 하며, 산업재해와 관련된 재해 예측이나 예방에 있어서 중요한 변수가 어떤 것이며, 중요하지 않은 변수가 어떤 것인지 제공할 수가 없다. 이에 본 연구에서는 산업재해 예방을 위하여 선행 연구들의 접근방법과는 달리 기존에 발생된 산업재해 데이터를 데이터마이닝(Data Mining)의 의사결정나무 기법을 적용하여 여러 가지 알고리즘을 비교 분석한 후 건설업 분야에서의 재해관련 데이터에 대한 특성분석을 하는데 있어서 최적의 알고리즘을 선정하고자 한다.

1.2 연구 방법

본 연구에서는 강원도 관내 전 업종에서 2002년부터 2004년까지 3년간 산업재해 신청을 하여 산재로 결정 통지된 67,278건의 데이터 중 가장 대표적인 업종인 건설업의 데이터를 분석하기 위하여 SAS의 Enterprise Miner와 SPSS의 AnswerTree 소프트웨어를 이용한다. 먼저 의사결정나무의 대표적인 알고리즘인 CHAID, CART, QUEST, C4.5를 각각 산재 데이터에 적용하여 알고리즘별 정확도, 민감도, 특이도 값을 구하여 비교 분석한다. 또한 이를 토대로 최적의 알고리즘을 선택하고 건설업 데이터의 분류방법에 대하여 교차타당성(Cross Validation)을 실행하여 타당성 평가를 실시한다.

2. 알고리즘 특징 비교

다음의 < 표 1 >은 의사결정나무의 대표적인 알고리즘들의 특성을 비교한 것이다. 각 알고리즘들의 대표적인 특징을 살펴보면, CHAID[1][4] 알고리즘은 다지 분리를 사

용하여 트리를 형성하고, CART[5][6]와 QUEST[7] 알고리즘은 이지 분리를 사용하여 트리를 형성한다. 그리고 C4.5[8] 알고리즘은 예측변수에 따라 연속형, 순서형인 경우 이지 분리를 사용하여 트리를 형성하고, 명목형, 이산형인 경우는 다지 분리를 사용하여 트리를 형성한다는 것이다.

< 표 1 > 의사결정나무 알고리즘의 특징 비교

	CHAID	CART	C4.5	QUEST
목표, 예측 변수	- 목표변수 : 명목형, 순서형, 연속형에서 분석이 가능 - 예측변수 : 명목형, 순서형, 그룹화된 연속형 변수를 사용	- 모든 측도의 목표변수와 예측변수에 적용할 수 있으며 전 체 탐 색 법 (exhaustive search method)을 사용	- 목표변수 : 이산형만 가능 - 예측변수 : 연속형, 범주형	- 목표변수 : 명목형 (이산형) - 예측변수 : 순서형, 연속형, 명목형
분리 방법	- 다지분리 (multiway split)	- 이지분리 (binary split)	- 연속형, 순서형 예측 변수 : 이지분리 - 명목형, 이산형 예측 변수 : 다지분리	- 이지분리
분리 기준	- 목표변수가 명목형, 순서형 : 카이제곱 통계량을 이용 - 목표변수가 연속형인 경우 : F 통계량을 이용	- 연속형 목표변수 : 분산의 감소량, 절대편차의 감소량이 최대가 되도록 분리 - 이산형 목표변수 : 지니 지수, 엔트로피, Twoing 지수를 이용하여 불순도가 최소가 되도록 분리	- 엔트로피를 이용한 이 득 비 율 (g a i n ratio) 을 사용	- 예측변수가 순서형, 연속형 : ANOVA F-검증, Levene의 검증 - 예측변수가 명목형 : Pearson의 카이제곱 검증을 사용
가지치기, 결측치	- 가지치기 과정을 포함하지 않고 정지기준에 의해 분리를 중지 - 결측치의 대치기준을 포함하지 않음	- 가지치기와 결측치의 대치기준을 포함	- 가지치기를 포함 - 결측치의 대치기준을 포함하지 않음	- 가지치기와 결측치의 대치기준을 포함

3. 데이터 셋

본 연구에서 사용된 데이터 셋은 2002년부터 2004년까지 산업자원부에서 강원도를 대상으로 집계한 업종별 산업 재해자 통계자료이다. 이 데이터들의 특성은 업종에 따른 독립변수로는 사업장명, 재해자명, 재해일자, 재해자 구분, 발생형태, 규모, 진료일수, 입원일수, 통원일수, 재가 일수, 공사규모, 연령, 성별, 요양기간, 근속기간, 재해월, 재해시간 그리고 근로손실일수로 총 18개이다. 하지만 여기서 개인 신상보호를 위한

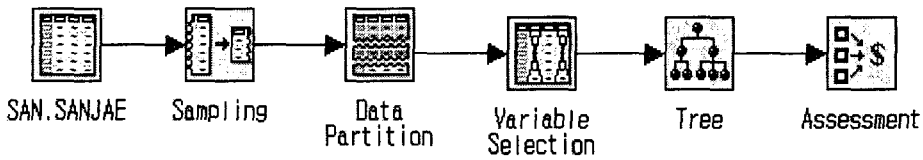
사업장명과 재해자명을 제외하고, 또한 결측치를 다수 포함하고 있는 독립변수를 제외하면 발생형태, 규모, 연령, 성별, 근속기간, 재해월, 재해요일 그리고 재해시간으로 함축된다. 아래 < 표 2 >와 같이 재해자 형태와 관련된 데이터는 총 67,278개로서, 부상은 60,249개이고, 사망은 7,029개이다. 그 중 건설업과 관련된 데이터를 중심으로 특성분석을 위한 최적 알고리즘을 선정하고자 한다.

< 표 2 > 업종에 따른 재해자 분포

업 종	재해자 형태		합 계
	부 상 자	사 망 자	
건설업	18,975	599	19,574
광업	12,903	5,459	18,362
금융보험업	450	24	474
기타산업	10,880	427	11,307
제조업	10,313	223	10,536
농업	269	3	272
어업	131	7	138
운수·보관업	2,946	203	3,149
임업	3,249	70	3,319
전기·상수도업	133	14	147
합 계	60,249	7,029	67,278

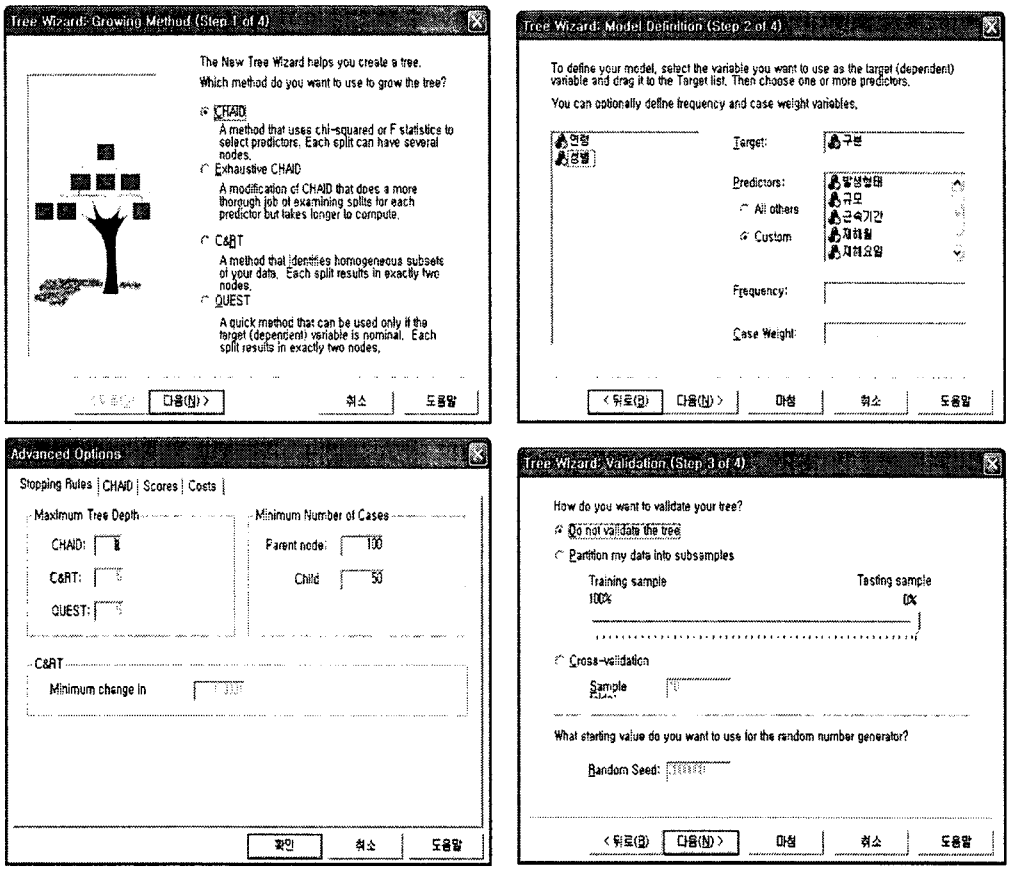
4. 분석 결과

의사결정나무 분석의 대표적인 알고리즘인 CHAID, CART, C4.5, QUEST를 비교분석한 후 최적의 모형을 선택하기 위하여 건설업에 대하여 각 알고리즘별 정분류율 또는 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity) 값을 구하여 비교 분석하였다. 본 연구에 사용된 건설업에 관련된 데이터 집합에 대하여 모형구축 자료(Training Data Set)와 모형검증 자료(Testing Data Set)[4]를 50 : 50의 비율로 하여 SAS의 Enterprise Miner를 이용하여 < 그림 1 >과 같은 데이터마이닝의 5단계 분석을 실행하였다.



< 그림 1 > SAS의 Enterprise Miner를 이용한 데이터마이닝의 5단계 분석

또한 SPSS의 AnswerTree에서는 아래 <그림 2>와 같은 과정을 통하여 알고리즘에 대한 분석을 실행하였다.

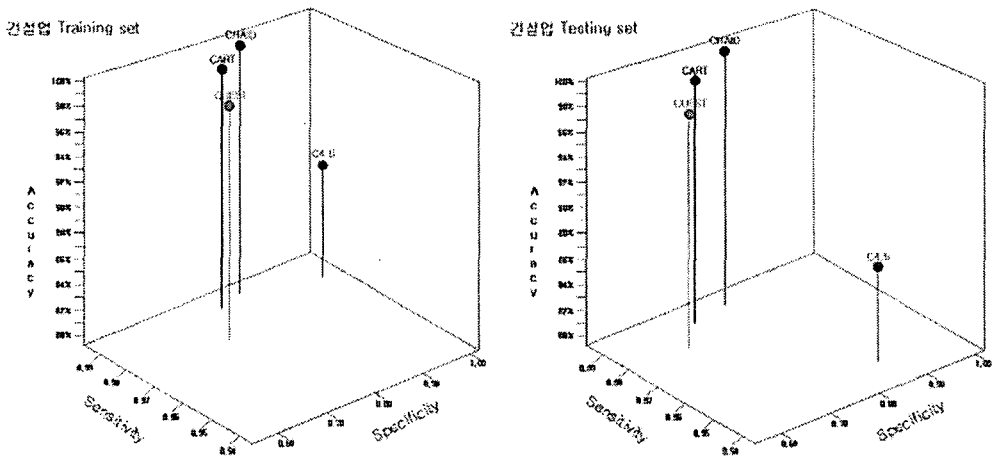


< 그림 2 > SPSS의 AnswerTree 분석 절차

이러한 과정을 통하여 알고리즘을 비교 분석한 결과는 다음의 < 표 3 >과 같다. 아래 < 표 3 >에서 볼 수 있듯이 건설업에 대한 알고리즘을 비교해 보면 CHAID가 모형구축 자료와 모형검증 자료에서 정확도(Accuracy)는 98.73907%와 98.19302%, 민감도(Sensitivity)는 99.16353%와 98.99588 %로 가장 높은 값으로 나타났다. 특이도(Specificity)는 C4.5가 모형구축 자료와 모형검증 자료에서 97.40249%와 86.34208%로 가장 높은 값을 나타냄으로 정확도, 민감도, 특이도 값을 비교 분석하여 보면 4개의 알고리즘 중 CHAID가 최적의 알고리즘인 것으로 나타났다. < 그림 3 >은 알고리즘별 성능비교를 쉽게 이해할 수 있도록 < 표 3 >의 결과를 그래프로 표현한 것이다.

< 표 3 > 건설업 데이터에 대한 알고리즘 성능비교

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



< 그림 3 > 건설업 데이터에 대한 CHAID, CART, C4.5, QUEST 알고리즘 비교

또한 건설업 데이터의 경우에서 CHAID 알고리즘에 대한 교차타당성은 < 그림 4 >와 같이 모형구축 오분류 값이 0.00837846, 교차타당성 오분류 값이 0.017932로 나타났다. 일반적으로 이론에 따르면 모형구축 오분류 값과 교차타당성 오분류 값의 차이가 10%이내이면 분류된 데이터들의 결과에 대한 타당성을 줄 수 있는 것으로 알려져 있는데, < 그림 4 >에서 나타난 두 값의 차이는 1%이내(0.017932-0.00837846)의 차이를 나타내고 있어서 본 연구에서 분류된 데이터들은 타당성을 갖고 있다고 할 수 있다.

Misclassification Matrix				
		Actual Category		
		1	2	Total
Predicted Category	1	18908	9	19005
	2	67	5	569
	Total	18975	5	19574
		Resubstitution		Cross-Validation
Risk Estimate		0.00837846		0.017932
SE of Risk Estimate		0.000651501		0.000948516

< 그림 4 > 건설업 데이터에 대한 교차타당성(Cross Validation)

5. 결론 및 추후 연구 사항

본 연구에서는 산업재해 예방을 위하여 데이터마이닝 기법 중 의사결정나무의 대표적인 알고리즘인 CHAID, CART, C4.5, QUEST를 SAS의 Enterprise Miner와 SPSS의 AnswerTree를 이용하여 대표적 업종인 건설업에 대하여 네 개의 알고리즘 별 성능평가를 위하여 정확도, 민감도, 특이도 값을 구하고 비교 분석하여 최적 알고리즘을 선정하였다.

건설업에 관련된 산업재해 발생 데이터를 가지고 알고리즘을 비교 분석한 결과 CHAID가 모형구축 자료와 모형검증 자료에서 정확도는 98.73907%와 98.19302%, 민감도는 99.16535%와 98.99588%로 다른 알고리즘에 비하여 가장 높은 값을 나타냈으며, 특이도는 C4.5가 모형구축 자료와 모형검증 자료에서 97.40249%와 86.34208%로 가장 높은 값을 보였으므로, 건설업에 관련된 데이터들에 대한 분석결과로는 CHAID가 최적 알고리즘임을 알 수 있었다.

추후 본 연구를 바탕으로 산업재해 예방을 위하여 재해 예측에 관련된 정보를 정량적, 정성적으로 제시할 수 있는 전문가 시스템을 개발하고자 한다.

6. 참 고 문 헌

[1] 강현철, 서두성, 최종후, Enterprise Miner의 의사결정나무분석 알고리즘, 아카데미, 2001.

[2] 김종현, "우리나라 산업재해 통계를 이용한 재해실태분석과 통계제도의 개선 방향", 경일대학교 석사학위논문, pp.40~60, 1998.

[3] 노동부, 산업재해현황분석, 2004.

[4] 최종후, 한상태, 강현철, 김은석, (AnswerTree를 이용한) 데이터마이닝 의사결정나무 분석, 고려 정보 산업, pp. 17~74, 1998.

[5] Breiman, L., J. H. Friedman, R. A.. Olshen, and C. J. Stone. Classification and regression trees, Belmont : Wadsworth, 1984.

[6] Dave S. Kerby, "CART analysis with unit-weighted regression to predict suicidal ideation from Big Five traits, Personality and Individual Differences", Volume 35, Issue 2, July, pp. 249-261, 2003.

[7] F. Berzal, iet al.,, "On the quest of easy-to-understand splitting rules, Data and Knowledge Engineering" Vol.44, pp. 31-48, 2003.

[8] M. Mulholland, D. B. Hibbert, P. R. Haddad and C. Sammut, "Application of the C4.5 classifier to building an expert system for ion chromatography", Chemometrics and Intelligent Laboratory Systems, Volume 27, Issue 1, January, pp. 95-104, 1995.

[9] R. Godin, R. Missaoui, "An incremental concept formation approach for learning from databases", Theoret. Comput. Sci. Vol. 133, pp. 387~419, 1994.

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.

최 요 한 : 강릉대학교에서 학사, 석사학위를 취득하였고, 현재 강릉대학교 정보전자공학부에서 박사과정을 수료하였으며 관심분야는 데이터마이닝, 산업안전, 정보시스템 등이다.