C4.5 알고리즘을 이용한 산업 재해의 특성 분석

임영문† · 곽준구 · 황영섭
강릉대학교 산업시스템공학과
(2005. 7. 20. 접수 / 2005. 11. 16. 채택)

A Feature Analysis of Industrial Accidents Using C4.5 Algorithm

Young-Moon Leem† · Jun-Koo Kwag · Young-Seob Hwang
Department of Industrial Systems Engineering, Kangnung National University
(Received July 20, 2005 / Accepted November 16, 2005)

Abstract : Decision tree algorithm is one of the data mining techniques, which conducts grouping or prediction into several sub-groups from interested groups. This technique can analyze a feature of type on groups and can be used to detect differences in the type of industrial accidents. This paper uses C4.5 algorithm for the feature analysis. The data set consists of 24,887 features through data selection from total data of 25,159 taken from 2 year observation of industrial accidents in Korea. For the purpose of this paper, one target value and eight independent variables are detailed by type of industrial accidents. There are 222 total tree nodes and 151 leaf nodes after grouping. This paper provides an acceptable level of accuracy (%) and error rate (%) in order to measure tree accuracy about created trees. The objective of this paper is to analyze the efficiency of the C4.5 algorithm to classify types of industrial accidents data and thereby identify potential weak points in disaster risk grouping.
Key Words : A Feature Analysis, Industrial Disasters, C4.5 Algorithm

1. 서 론

우리나라의 산업재해율은 1995년도에 선진국 수준인 1%미만으로 낮아졌으나, 1998년부터 조금씩 지속적인 증가추세를 보이고 있으며, 재해로 인한 실질적인 근로손실을 나타내는 재해 강도율도 2000년을 기점으로 매년 증가추세를 보이고 있다. 이러한 이유에는 여러 가지 복합적인 요소들이 작용하고 있으나, 우리의 재해예방에 대한 활동들이 아직도 구체적이고, 근본적인 위험요인을 찾아 제거하지 못하고 피상적인 활동에 머무르고 있는 것으로 분석되고 있다[4]. 우리는 산업 현장에서 근로자들이 작업이나 업무에 기인하여 재해가 발생되어 상해를 입는 일 그리고 각종 질병에 이환되거나 이상이 생겨 악화되는 일이 발생하지 않도록 최선을 다하여야 할 것이다. 그러나 이와 같은 재해나 각종 질병의 이환에는 정도의 차이는 있겠지만 어느 정도의 수준까지는 받아들일 수밖에 없는 것이 우리 사회의 현실이다. 왜냐하면 산업재해를 전혀 발생하지 않게 하기 위해서는 막대한 자금과 인력을 투자해야 하고, 또한 자금과 인력을 투자 할지라도 근로자들의 과실 등에 의해 산업재해를 완전히 제거할 수 없는 경우가 종종 발생하기 때문이다. 산업재해를 예방하기 위한 노력으로 현재 축적되어 있는 다량의 데이터를 이용하여 다양한 방법의 데이터 분석이 많은 연구자들과 기관에 의해 이루어지고 있다. 기존에 이루어졌던 통계 분석을 이용한 재해에 관한 연구를 살펴보면 규모별, 성별, 연령별, 근속기간별, 재해발생시기별, 발생형태별로 데이터를 분석한 것을 볼 수 있다. 하지만, 이러한 연구의 분석방법을 살펴보면 데이터의 빈도분석을 통해 분석한 내용이 주를 이루고 있으며 빈도분석을 통해 나온 요인별 특성만을 토대로 재해를 줄이기 위한 개선방법에 대해 언급한 것을 볼 수 있다[2-4]. 이에 본 논문에서는 데이터의 효율적이면서 체계적 분석을 위해서 Data Mining Algorithm 중에서 C4.5 알고리즘을 사용하

† To whom correspondence should be addressed.
ymleem@kangnung.ac.kr

였고, 데이터 가공을 위해서 SAS의 Enterprise Miner를 사용하였는데, Enterprise Miner는 C4.5 알고리즘을 분석하는 가장 보편적인 도구로서[5] 데이터 가공이 쉽고, 결과 분석 역시 쉽게 행할 수 있고, C4.5 알고리즘을 이용하여 재해자의 성향과 특성을 분석하는데 중요한 도움을 줄 것이기 때문이다.

2. 분석 방법

의사 결정나무(Decision Tree)는 한정된 수의 클래스로 예를 분류하는 것으로 데이터의 군집화와 분류에 적합한 방법이다[1,10]. 군집화는 모집단을 유사한 특성에 따라 세그먼트로 나누는 것이다[8]. 예를 들면, 보험회사는 어떤 고객속성이 높은 클레임과 관계가 깊은지를 알고 싶어 한다. 알고리즘을 적용하였을 때 가장 영향력이 있는 속성이 결혼 상태라고 결정한다면, 모집단을 결혼과 미혼의 두개의 클러스터로 나누고, 그 다음 영향력이 있는 나이, 소유차종, 거주지 등으로 나누어 분할한다. 이 알고리즘에서는 예측을 보다 정교하게 하기 위해 각 분할된 클러스터에 통계적인 유의성을 부여한다. 하나의 대상이 어떤 클래스에 속하는지를 결정하기 위해 개념을 나무구조로 표현한다. 일반적으로 효율성을 위해 한 번에 전체 데이터에 대하여 의사결정나무를 구성하는 방법보다는 초기에 랜덤하게 선택된 작은 부분집합을 설정하는 것으로부터 출발하여 전체 데이터로 확대해 나간다[6]. 나무외 각 노드는 속성명칭을, 각 노드에서 분기되는 아크는 이들 속성이 가질 수 있는 값들을 의미하며 최하위 노드는 여러 클래스로 분류되었음을 나타낸다. 본 연구 논문에서는 C4.5 알고리즘을 사용하여 Tree를 구성하였고 C4.5 알고리즘을 사용하기 위한 툴로 SAS의 Enterprise Miner[8]를 사용하였다.

2.1. C4.5 알고리즘

C4.5 알고리즘은 J. Ross Quinlan에 의해 수정 발전된 의사결정 알고리즘이다[9]. 이것의 초기버전인 ID3 (Interactive Dichotomizer 3, 1986) 알고리즘은 기계학습(Machine Learning) 분야에 많은 영향을 주었다. CART가 각 마디에 이원분할을 형성하며 이지분리 나무구조를 만드는데 반하여 C4.5는 연속형 예측변수에 관해서는 이지분리를 하지만, 명목형 예측변수에 관해서는 각 범주가 하나의 마디를 가지는 다지 분리 구조를 갖는 나무로 구성된다. C4.5에서 의사결정나무를 형성하기 위하여 처음 수행하는 작업은 분할 정복(Divide and Conquer)이다. 입력되는 훈련 집합(Training Set)이 성공적으로 분할되도록 모든 하부집합에 하나의 Class가 속하는 경우들로 구성될 때까지 나무를 형성한다. C4.5는 정보(Information)라는 개념을 사용한다. p가 한 메시지(Message)의 확률일 때, 이 Message로 전달되는 Information은 $-\log_2 p$로 측정한다. 예를 들어 8개의 동일한 확률을 갖는 메시지(Equally Probable Message)가 있을 경우, 한 메시지의 Information은 $-\log_2 \frac{1}{8} = 3$이 된다. 이는 작은 확률로 일어나는 메시지일수록 이를 알기 위해서는 보다 많은 정보가 필요하다는 뜻이다. Case들의 집합인 S에서 무작위로 한 Case를 선택할 때, 이 Case가 C_j에 속할 확률은 다음과 같다.

$$\frac{freq(C_j, S)}{|S|} \tag{1}$$

여기서, |S|는 S에 속하는 모든 Case의 개수이고, $freq(C_j, S)$는 집합 S에서 C_j에 속하는 Case들의 개수이다. 따라서 이 Case가 전달하는 정보(Information)는 다음과 같다.

$$-\log_2\left(\frac{freq(C_j, S)}{|S|}\right) \tag{2}$$

집합 S에서 기대 정보(Expected Information)를 구하기 위해선, 각 case가 전달하는 정보를 가중평균하면 된다.

$$\text{info}(S) = -\sum_j^k \left(\frac{freq(C_j, S)}{|S|} \times \log_2\left(\frac{freq(C_j, S)}{|S|}\right)\right) \tag{3}$$

위의 info(S)와 비슷한 개념으로, T가 X에 의해 n개로 분할 된 후의 기대 정보(Expected Information)를 구하려면 식 (4)를 이용할 수 있다.

$$\text{info x}(T) = \sum_i^n \left(\frac{|T_i|}{T} \times \text{info}(Ti)\right) \tag{4}$$

X에 의한 분할로 얻어진 정보(Information)는 다음 식 (5)에 의해 얻을 수 있다.

$$\text{Gain}(X) = \text{info}(T) - \text{info x}(T) \tag{5}$$

기존 알고리즘인 ID3에서는 이 Gain을 최대로 하는 test를 선택했었다. 그러나 이 경우에는 범주의 수가 많은 변수로의 심각한 bias가 생기는 문제점이 있다. 예를 들어, 각 Terminal Node에 한 Case만을 포함하며, 모든 Case들이 1의 확률로 배정되는 분리 변수가 있다고 하자. 이 경우에는 info x(T) = 0 일 것이다. 따라서 어떤 변수를 사용하는 것보다 Information Gain이 최대가 될 것이다. 그러나 이러한 분리는 전혀 의미를 갖지 못한다. 그래서 T에 있는 한 Case가 속하는 Subset(Class 대신)을 정의(Identify)하는데 필요한 평균 정보의 양(Split Info)으로 정규화(Normalize)시켜줄 필요가 있다.

$$\text{Split Info}(X) = -\sum_{i=1}^{n}\left(\frac{|T_i|}{|T|}\times\log_2\left(\frac{|T_i|}{|T|}\right)\right) \quad (6)$$

Split Info는 T가 n개의 Subset으로 분할됨에 따라 발생하는 정보(Information)의 양이다. Gain을 Split Info로 나누어 주면, Split에 의해 생성된 유용한 정보의 비율인 Gain Ratio가 된다.

$$\text{Gain Ratio}(X) = \frac{gain(X)}{splitinfo(X)} \quad (7)$$

변수별로 Gain Ratio를 최대화 시켜주는 Split Point를 찾고, 이를 각 변수 별로 실시하여 변수끼리 최대 Gain Ratio를 또 비교하여, 그중에서 가장 큰 변수를 선택하면 된다. 즉 C4.5도 변수의 선택과 Split Point의 선택이 동시에 이루어진다[7].

3. 데이터 셋

본 연구에서 사용 되어진 데이터는 산업자원부에서 발간된 통계집에서 발췌한 자료로서 2003년, 2004년 2년 동안 강원도 관내에서 각 업종별 재해 발생에 관련된 데이터이다. 연구에 사용되어진 초기 Raw Data의 경우 재해 일자, 재해자 구분, 발생형태, 업종, 규모, 직종, 진료 일수, 입원 일수, 통원 일수, 재가 일수, 연령, 성별, 요양 기간, 근속기간, 재해월, 재해 요일, 재해 시간, 근로 손실 일수 등 총 32개의 변수와 총 25,159개의 데이터로 구성되어 있다.

이 데이터 중에서 본 연구에서 필요로 하는 데이터를 얻기 위해 결측치, 즉 값이 없거나 전혀 상관없는 값들이 들어간 데이터를 제거하는 데이터 정제 과정을 거쳐 발생형태, 업종, 규모, 나이, 근속기간, 재해 요일, 성별, 재해월, 재해시간 등 9개의 변수와 24,887개의 데이터를 사용하였다. 아래의 Table 1은 Tree를 구성하고자 할 때, 목표 변수로 선택되어진 업종의 데이터 구성을 보여주고 있다.

Table 1. Constituent of data on target value

No	발생 형태	데이터 수
1	광업	8,120
2	제조업	3,205
3	금융보험업	224
4	건설업	7,081
5	운수보관업계	994
6	임업	1181
7	기타산업계	4,086

Table 1에 나온 업종의 데이터 구성에서 광업이 가장 많은 재해자가 발생한 것을 볼 수 있고, 그 다음이 건설업, 기타, 제조업 순으로 재해자가 발생한 것을 볼 수 있다.

발생 형태 데이터 구성에서 기타는 감전, 광산사고, 분류불능 데이터, 붕괴, 도괴, 익사, 빠짐 등 다양한 형태의 재해 사고를 포함하고 있다. 본 연구에서는 위에서 보여준 9개의 변수를 이용하여 목표 변수를 업종으로 하는 의사결정나무 분석을 이용한 업종별 산업재해 데이터의 Tree를 만들고 이를 분석하여 산업재해자의 특성을 분석하고자 한다.

4. 분석 결과

4.1. SAS Enterprise Miner를 이용한 데이터 분석

SAS E-Miner를 이용한 분석을 하기 위해 Fig. 1의 분석 흐름도를 작성할 때, 본 연구에서 사용한 데이터는 2년 동안의 데이터라고는 하지만 24,000여개의 데이터로서 많은 데이터는 아니므로 각 업종별 재해자의 개략적 특성을 알아보기 위해 모든 범주에 대해서 Tree를 구성해 나가도록 하였다. 또한 최하위 노드에 포함되어 있는 데이터의 개수가 1%인 노드까지 트리를 하여 분리를 한 결과 특정한 분리가 나오지 않아 최대로 나눌 수 있는 Depth까지 Tree를 나누었고 이 Tree를 분석하여 어떤 인자가 업종별 재해자의 특성을 분석하는데 도움을 주는 변수인지 알아보고자 하였다.

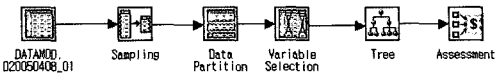

Fig. 1. Flow of analysis.

4.2. Tree 분석 결과

본 연구에서 나눈 Tree의 경우 총 222개의 node로 구성되어 있고 최종 노드인 Leaf node는 151개의 node로 구성되었으며 이처럼 Tree의 수가 많은 것은 범주의 수가 많기 때문인 것으로 사료된다. 범주의 수가 많기 때문에 Tree의 수가 많아진 것이다. 만약 CART (Classification and Regression Tree) 알고리즘으로 Tree를 구성 하였다면 2진 분리로 하는 CART 알고리즘의 특성상[7] Tree node의 개수가 이렇게 많이 나오진 않았을 것이라 추정할 수 있다. C4.5 알고리즘을 이용하여 구성한 트리의 결과를 보면, 정확도와 오분류 확률을 먼저 확인해야 한다[11]. 이 트리의 정확도와 오분류 확률은 다음 Table 2와 같다.

Table 2. Accuracy and error rate of tree

	정확도	오분류 확률
Train 데이터	0.6921 (69.21%)	0.3079 (30.79%)
Validation 데이터	0.6853 (68.53%)	0.3147 (31.47%)

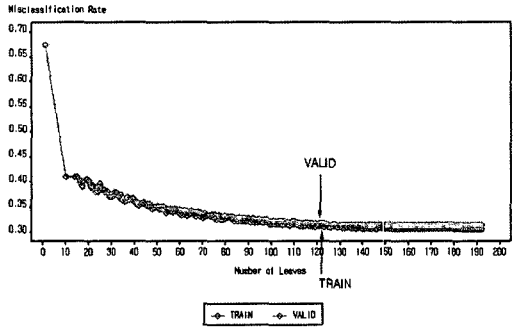

Fig. 2. Error rate of each node.

위의 Fig. 2를 보면 처음 최상위 노드에서 그 다음 노드로 Tree가 분리 될 때 오분류 확률이 현격하

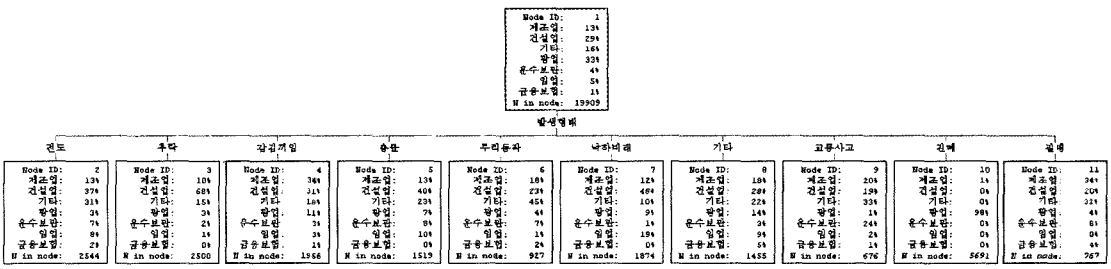

Fig. 3. Tree according to occurrence type.

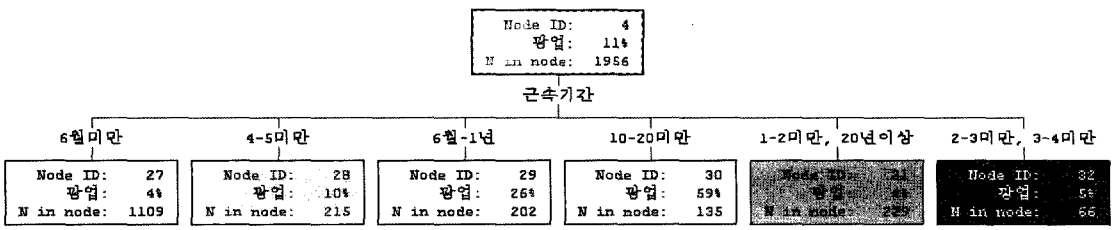

Fig. 4. Entanglement, ejection(node4).

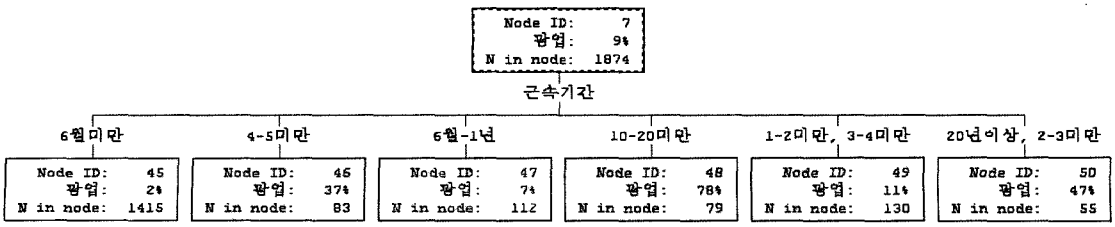

Fig. 5. Fall or flying object from scaffolds(node7).

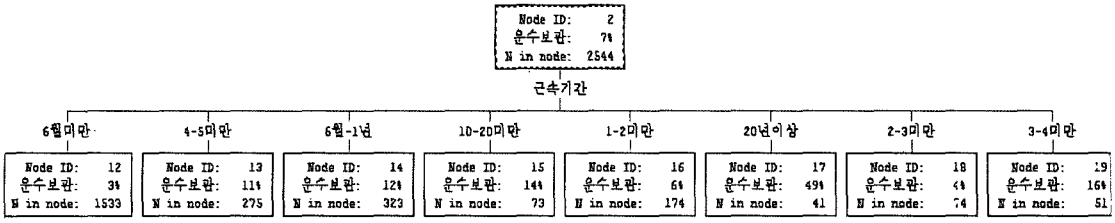

Fig. 6-1. Business on traffic service and custody(node2).

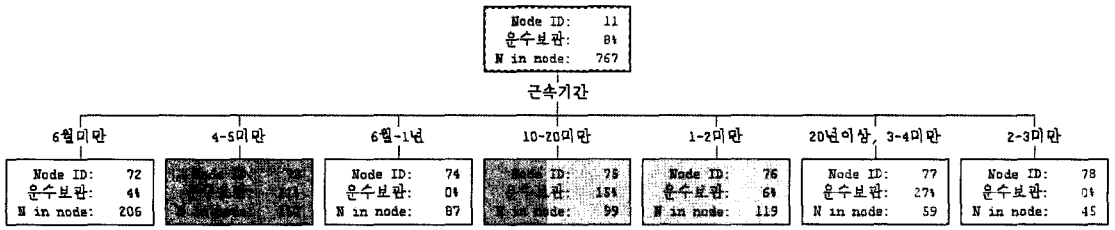

Fig. 6-2. Business on traffic service and custody(node11).

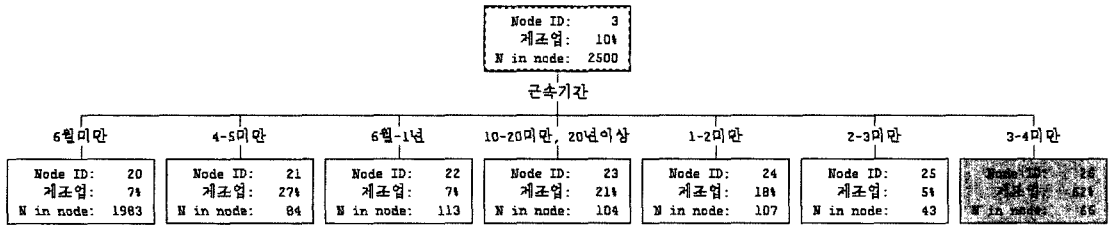

Fig. 7. Manufacturing(node3).

게 떨어지는 것을 볼 수 있다. 또한 계속해서 Tree가 분리되면 분리 될수록 오분류 확률이 감소하고 트리가 계속해서 분리 될수록 감소는 거의 없어지고 거의 직선에 가까워지는 것을 볼 수 있다. 데이터 분석 결과는 Fig. 3부터 Fig. 8-6에 나타나 있다.

Fig. 3은 E-Miner를 실행했을 때 생성되는 Tree중 첫 번째 Depth의 Tree를 보여준 것이다. 위 Fig. 3의 하위 Tree들은 양이 많고 한 번에 다 표현할 수 없

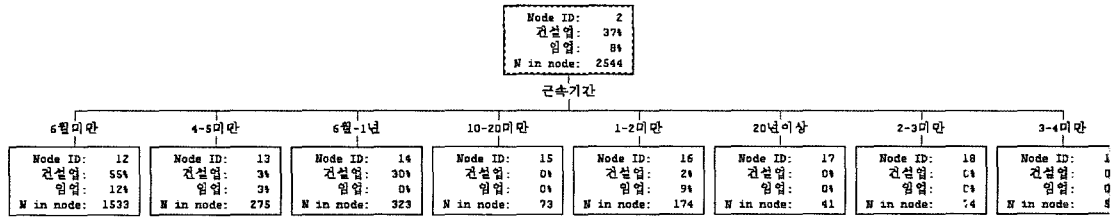

Fig. 8-1. Forestry and construction(node2).

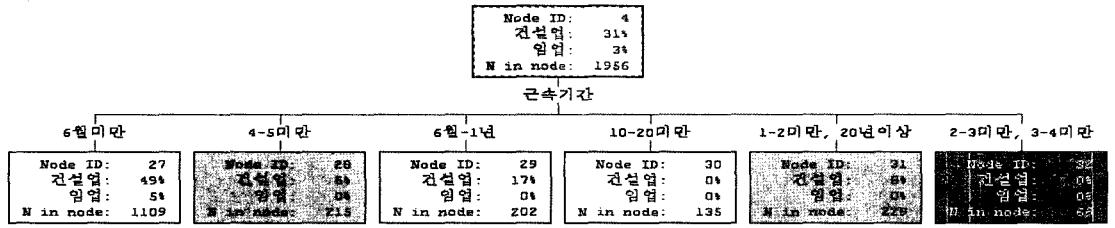

Fig. 8-2. Forestry and construction(node4).

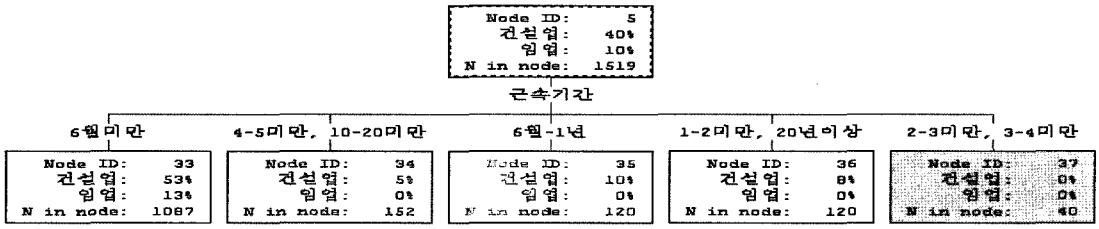

Fig. 8-3. Forestry and construction(node5).

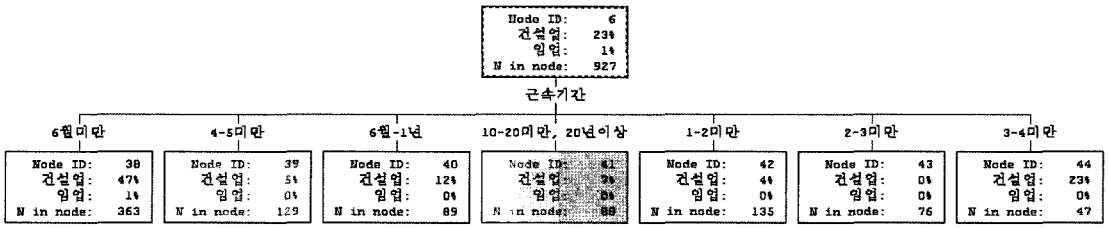

Fig. 8-4. Forestry and construction(node6).

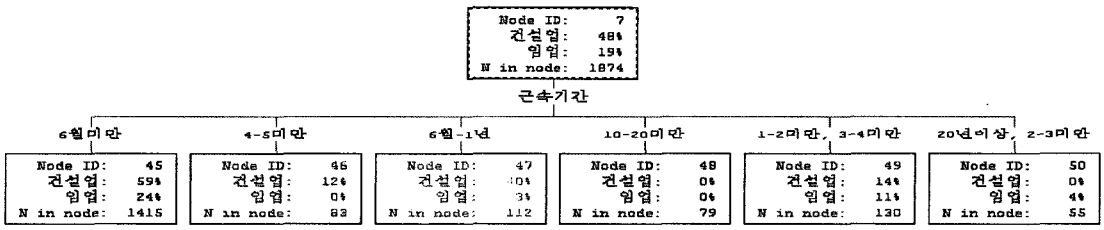

Fig. 8-5. Forestry and construction(node7).

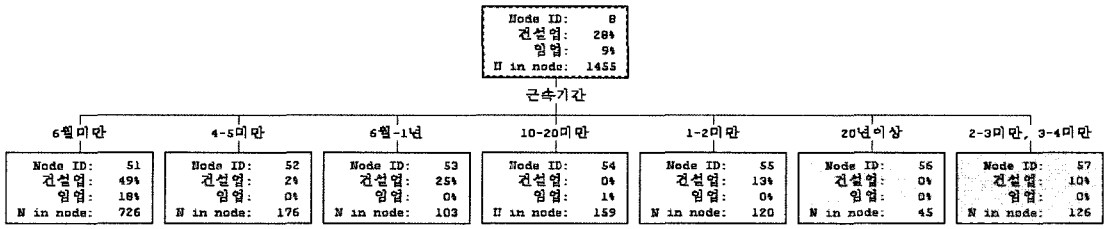

Fig. 8-6. Forestry and construction(node8).

기에 Tree들 중에서 발췌하여 표기하였다. Fig. 3 트리의 결과를 살펴보면 Node의 10번에서 진폐로 분리된 Tree에서 대부분의 업종 데이터가 광업에 속해 있는 것을 볼 수 있다. 또한 광업 분야에서는 재해의 발생형태가 진폐, 추락, 감김·끼임, 충돌, 낙하·비래, 교통사고, 기타 등으로 나타났다. 진폐의 경우는 장기간 종사한 사람에게서 나타나는 질병으로 먼지나 이물질이 호흡기관을 통해 폐로 들어가 생기는 질병을 말한다[4]. 진폐의 경우 98%가 광업에 종사하는 근로자에게서 나타난 것으로 보여졌으며 재해자의 수도 다른 재해에 비해 그 수가 상당히 많은 것으로 나타났다. 추락의 경우 광업에서 10년 이상 종사한 사람들에게 많이 나타났고, 이는 앞서 얘기한 바와 같이 오랜 경력에서 오는 부주의에 의한 것으로 추정된다. 감김·끼임의 경우 근속기간이 10년~20년 미만에 속하는 사람들에게서 나타났고, 충돌은 근속기간이 10년~20년 미만인 사람들에게서, 낙하·비래는 4년~5년 미만, 10년 이상 근무한 종사자들에게서 나타났다. 교통사고의 경우는 주로 7월과 8월에 많이 나타나는 것으로 나타났다. 임업의 경우는 낙하·비래의 재해 발생이 주로 일어나는 것으로 나타났으며 근속기간은 6개월 미만의 근로자

들이며, 재해가 발생한 월은 주로 8월과 12월이고, 재해 시간은 6시~10시 사이와 16~18시 사이에 재해가 발생한 것으로 나타났다. 금융 보험의 경우는 주로 전도가 발생한 것으로 나타났고 4년에서 5년 미만 근로한 근로자들이며 주로 5월과 10월에 나타난 것으로 분석되었다. 운수 보관업계에서는 전도, 충돌, 교통사고, 질병 등이 발생한 것으로 나타났으며 전도의 경우 근속기간이 20년 이상인 사람에게 나타난 것으로 보여졌다. 또한 전도에서 근속기간이 4년에서 5년 미만의 근로자가 1월에 재해가 발생한 것으로 나타났고, 6개월에서 1년 미만 근로한 근로자의 경우 9월과 10월에 재해가 발생했다. 충돌의 경우 근속기간이 20년 이상인 사람들에게서 나타났고, 교통사고의 경우 주로 11월에 발생했다. 질병은 근속기간이 3년에서 4년 정도인 근로자에게서 나타났다. 제조업의 경우는 추락, 감김 · 끼임, 충돌, 무리한 동작, 낙하 · 비래, 질병, 전도, 교통사고 등 다양한 재해가 발생한 것을 볼 수 있으며, 추락의 경우 근속기간이 3년~4년 미만인 사람에게 나타났고, 감김 · 끼임의 경우 1년~5년 미만, 20년 이상 근무한 사람에게서 나타났다. 또한 근속기간이 6개월 미만인 경우는 주로 1월과 2월, 4월에 재해가 발생했다. 충돌의 경우 2년~4년 미만, 10~20년 미만인 경우에, 무리한 동작은 10년 이상인 근로자에게서, 낙하 · 비래는 1년~2년 미만인 사람에게서 나타났다. 질병의 경우 1년 이상 근무한 근로자에게서 나타났다. 전도의 경우 근속 기간이 4년~5년 미만인 경우는 2, 3, 7, 8, 9월에 주로 재해가 나타났고, 1년 미만인 경우 2, 4, 6, 12월에 나타난 것으로 분리되었다. 데이터의 양이 두 번째로 많았던 건설업의 경우 거의 모든 재해의 발생 형태를 갖고 있었고, 재해 또한 가장 많이 일어난 것으로 나타났다.

건설업의 분리에서는 특이하게도 거의 모든 재해 발생 형태가 근속기간이 6개월 미만인 비 숙련자에게서 가장 많이 나타났다. 특히 추락과 충돌의 경우 거의 모든 연령대에서 나타났다. 즉, 건설업에서의 경우 6개월 미만의 비 숙련자들의 관리 소홀과 부주의로 인한 재해가 가장 많았던 것으로 추정된다. 기타 산업계의 경우도 건설업과 마찬가지로 거의 모든 재해의 발생형태를 갖고 있지만 이것을 어떤 하나의 산업계로 분류하고 그룹화 하기에는 기타 산업계에 들어가 있는 업종의 종류가 많고 또한 연관성이 없기 때문에 어려울 것이다. 또한 여기서 여러 가지 발생형태가 나타난 것은 업종의 종류가 다양하여 그에 따른 재해의 발생 형태 또한 다양하기 때문에 이러한 분류가 나온 것으로 보인다.

5. 결론 및 추후 연구 사항

본 연구에서 업종별 산업 재해 데이터를 분석해 본 결과 오분류 확률은 약 30% 정도로 나타났고 트리가 생성되면 생성 될수록 오분류 확률은 줄어드는 것을 알 수 있으며, 데이터 분석의 결과 제시할 수 있는 결론을 요약하면 다음과 같다.

1) 광업을 분류할 때 영향을 주는 변수들로는 발생 형태, 근속기간, 재해 발생월등이 있다. 광업에 종사하는 근로자들 중에는 10년 이상 종사한 경우 진폐에 해당하는 사람이 많았고 감김 · 끼임, 충돌, 낙하 · 비래의 경우도 진폐와 마찬가지로 10년 이상 종사하는 사람들에게 많이 나타난 것을 볼 수 있다. 진폐는 광업에 종사하는 사람들에게서 가장 많이 나타나는 것으로 초기에는 증상이 없고 5년, 10년 이상 근무하는 사이 서서히 발병하는 것으로 알려져 있다. 이 질병은 작업장의 환경 개선과 작업 마스크 착용 등 미리 예방하는 방법 밖에는 없다. 감김 · 끼임, 충돌, 낙하 · 비래의 경우 10년 이상 종사하는 사람들에게서 많이 나타나는 이유는 오랜 경력에서 나오는 숙련도 과신으로 인한 주의 소홀, 작업에 대한 작업 절차 무시 등을 원인으로 추정할 수 있다.

2) 광업뿐만 아니라 운수 보관업계와 제조업 또한 3년 이상 근무한 숙련자에게서 재해의 비율이 높은 것을 알 수 있다. 이는 관리자의 주의 깊은 관리와 숙련자에게 안전에 대한 인식 그리고 작업의 위험성을 각인 시켜 주는 것이 재해를 예방하는데 도움이 될 것이다. 운수 보관업계의 경우 광업과 마찬가지로 발생형태, 근속기간, 재해 월에 의해서 Tree가 분리되는 것을 볼 수 있고, 제조업의 경우는 발생형태, 근속기간, 재해월과 재해 시간에 의해 분리가 되는 것을 볼 수 있다.

3) 이에 반해, 임업과 건설업의 경우 6개월 미만의 비숙련 근로자에게서 더 많은 재해자가 나온 것을 알 수 있다. 재해자의 근속기간이 짧은 것을 빼면 위에서 제시한 광업과 운수 보관업계, 제조업계와 같이 부주의에서 비롯된 재해로 예상 할 수 있다. 이는 관리자의 숙련자와 비 숙련자에 대한 세심한 관리가 필요함을 알 수 있고 또한 작업 환경을 개

선하여 작업장 내에서 일어나는 일에 대한 주의가 필요함을 알 수 있다. 건설업의 분리에 영향을 준 변수들로는 발생형태, 근속기간, 재해 월, 나이, 재해 시간이 있고, 임업의 경우 발생형태, 근속기간, 재해 월, 재해 시간에 의해 분리되었다.

의사결정 알고리즘을 이용한 데이터 분석 결과를 보면 각 업종에 대한 재해자들의 발생 형태, 재해 요일, 근속기간, 나이와 재해 시간 등 재해 발생의 특성을 알 수 있었다. 본 논문에서 분석한 데이터 중 업종과 발생형태에서 기타를 보면 다량의 업종과 다량의 발생형태가 포함되어 있고, 데이터의 수 또한 많아서 세밀한 분석에 어려움을 주었다. 기타에 포함되는 발생형태의 경우 감전, 유해화학, 이상온도 등과 같이 연관성도 없고 종류 또한 많아 분석을 하는데 있어서 어려운 점이 많았다. 좀 더 효율적인 분석을 위해서는 체계화 된 대량의 데이터로 정확도 및 분석에 대한 비교 분석 및 최적의 알고리즘 선정 등이 필요하다고 사료된다.

감사의 글 : 본 연구는 산업자원부 지역혁신 인력양성사업의 연구결과로 수행되었음.

참고문헌

1) 권혜숙, "데이터마이닝 패키지에서 분류나무 알고리즘의 비교 연구", 서울대학교 석사 학위 논문, 2002.
2) 김경배, 산업재해의 예방대책에 관한 연구, 관동대학교 석사학위논문, 2004.
3) 김종현, 우리나라 산업재해 통계를 이용한 재해 실태분석과 통계제도의 개선방향, 경일대학교 석사학위논문, pp. 40~60, 1998.
4) 노동부, 2004년판 노동백서, pp 56~65, 2004.
5) 오희경, 최형인, "데이터 마이닝 분류 모델 비교 및 분석", 서울대학교 석사 학위 논문, 2002.
6) 임영문, 최영두, "연관규칙을 이용한 데이터 분석에 관한 연구", 산업경영시스템학회지, Vol. 23, No. 61, pp 115~126, 2001.
7) 최종후, 한상태, 강현철, 김은석, "(Answer Tree를 이용한) 데이터 마이닝 의사결정나무 분석", 고려 정보 산업, pp. 17~74, 1998.
8) 최종후, 한상태, 강현철, 김은석, 김차용, 김미경, "SAS Enterprise Miner를 이용한 데이터 마이닝-방법론 및 활용", 자유아카데미, pp. 230~300, 2001.
9) Chong Yau Fu.; "Combining Loglinear Model with Classification and Regression Tree (CART) : an Application to Birth Data", Computational Statistics & Data Analysis 4, 2004.
10) Kamber, M., Winstone, L., Gong, W., Cheng, S., Han, J.; "Generalization and Decision Tree Induction", Efficient Classification in Data Mining. Proceedings of the International Workshop Issue of Data Engineering (RIDE' 97) Birmingham, UK. pp. 111~120, 1997.
11) Pietersma, D., Lacroix, R., Lefebvre, D., Wade, K. M., "Induction and Evaluation of Decision Tree for Lactation Curve Analysis", Comput. Electron. Agric. 38, pp. 19~32, 2003.