안전보건감독에서 빅데이터와 머신러닝의 역할에 관한 유럽 안전보건청(OSHA)의 토론자료(1)

번역 대외홍보팀

구성

※7월호에는 '서론'부터 '빅데이터와 머신러닝'이 연재되며 8월호에는 '근로감독 대상을 정할 때 빅데이터와 머신러닝 활용하기'부터 '결론'이 연재됩니다.

서론

근로감독은 각 기업이 산업안전보건 규정을 준수하기 위해 필요한 조치를 행하는지 여부를 확인하기 위해 근로감독기관이 사용할 수 있는 가장 중요한 정책수단이라 할 수 있습니다.

근로감독의 효과는 몇 가지 각기 다른 기본요소의 영향을 받습니다.

그러한 기본요소 중의 하나는 감독 대상, 즉 감독할 회사나 현장을 선택하는 절차입니다. 원칙적으로 유용하게 사용할 만한 최소한 세 가지 선택방법이 있습니다.

첫 번째 방법은, 모든 회사를 그 회사에 잠재된 위험성이나 회사 규모, 산업 유형 또는 여타 기준에 관계없이 감독 대상으로 선정하는 것입니다.

두 번째 방법은, 회사의 특성에 관계없이 모든 회사가 같은 확률로 감독 대상에 뽑힐 수 있는 무작위 표본추출 방법입니다. 예방적이거나 경제적인 측면에서 보았을 때, 이 두 가지 방법 모두 대체로 효과가 없는 것으로 보입니다(Blanc, 2013).

따라서 대부분의 근로감독기관은 위험기반 접근방식으로 불리우는 세 번째 방법을 통해 근로감독 대상을 선택합니다. 요컨대, 위험기반 접근 방식은 위험 수준을 기준으로 감독 대상 사업장을 선택하는 것입니다.

위험기반 접근 방식은 대부분의 선도적인 근로감독기관에게는 필수적인 원칙이지만, 실제 적용하기에는 상당한 어려움이 있습니다.

적용이 힘든 주된 이유는, 위험 분석을 위해 만들어진 정밀한 방법론이 부족하다는 것입니다(Mischke 등, 2013).

위험성에 기반하여 우선순위를 정할 수 있는 적절한 방법론이 개발되지 않으면, 위험기반 접근 방

식이라는 것이 자칫 가시적인 정책성과가 빠진 정책 결과 발표와 유사한 것이 되어 버릴 수도 있습니다.

따라서, 특히 고위험 기업군을 대상으로 하는 위험분석 방법론을 개발할 필요가 있습니다(Weil, 2008).

대부분의 근로감독기관은 감독 대상 기업 및 감독 활동과 관련하여 막대한 양의 데이터를 수집하고 저장합니다.

따라서 근로감독기관은 어마어마한 양의 그리고 빠른 속도로 늘어나는 데이터, 이른바 '빅데이터'를 보유하게 됩니다. 빅데이터와 머신러닝 기술을 결합함으로써, 그 데이터 안에 있는 숨겨진 의미나 흐름을 분석하여 여러가지 예측이나 예상의 목적으로 다양하게 활용하는 경우가 늘어나고 있습니다.

예를 들면, 빅데이터와 머신러닝 기술을 활용하여 암 발병가능성을 예측하여 환자 치료에 반영하거나, 기업의 파산가능성 예측, 유가 예측, 세금 관련 사기행위 탐지, 범죄발생 예측 또는 주식시장 변동 흐름 예상 등의 다양한 분야에서 활용 가능성 테스트가 진행중입니다.

그러나, 본 논문에서 언급되어야 하는 근본적인 의문 하나가 있습니다. 그것은 '빅데이터와 머신러닝 기술을 활용하여 위험성이 높은 근로감독 대상 기업을 선정하는 것이 과연 근로감독기관들에게 유망한 방법이 될 것인가'라는 것입니다.

위험성에 기반한 근로감독 대상 선정

OECD(경제협력개발기구)에서 2014년에 발표한 '규제정책을 위한 모범사례'에 따르면, 근로감독기관들이 근로감독 대상을 정할 때 위험성 분석과 위험성 평가가 기본적으로 실행되어야 한다고 강조하고 있습니다. 다시 말하면, 그 기업에 산재사고, 유해물질 사용 또는 불법적인 근로조건 등의 위험요소가 발생 또는 존재할 가능성이나, 해당 사안이 발생 또는 존재한 적이 있는지 등을 확인하거나 평가한 후에 근로감독 대상 기업을 선택해야 한다는 것을 의미합니다.

감독기관의 인적, 물적 자원이 유한하기 때문에 모든 위험 대상 기업의 모든 위험 부문을 통제하거나 확인하는 것은 불가능하다는 것이, 위험성에 기반하여 근로감독 대상을 정할 때의 기본적인 전제조건입니다. 다시 말하면, 근로감독기관이 안전보건감독을 실시할 때, 해당 기업의 몇몇 문제가 있는 부문이 다른 부문보다 근로감독 우선순위가 높아야 한다는 것을 인식해야 한다는 의미입니다. 같은 맥락으로 동일 산업군 내에서 몇몇 문제가 있는 기업이 다른 기업보다 근로감독 우선순위가 높아야 합니다.

위험성에 기반하여 근로감독 대상을 정한다는 원칙은 새로운 것이 아닙니다. 작업장에 대한 영국의 산업보건 감독 시스템에 대하여 거의 50년 전에 진행된 로벤스 위원회의 평가보고에 의하면, 위험성에 기반한 감독방식을 기업의 자율규제와 결합하여 근로감독을 진행하는 것이 안전보건감독 절차를 현대화하는 이상적인 방법 중의 하나라고 소개하고 있습니다(Robens, 1972).

근로감독기관의 인적,물적 자원을 효율적으로 사용하는 데 있어 로벤스 위원회의 보고서에서 권고하는 방법은, 해당 자원을 가장 심각한 문제가 있는 부문에 선택적으로 집중하고, 산재사고 통계나 감독 대상 기업의 기술정보 또는 감독관의 현지 지역 관련 지식 등 안전보건과 관련된 모든 유용한 데이터에 대한 체계적 분석을 통해 위험성이 높다고 확인된 기업과 문제에 근로감독의 우선순위를 높여서 적용하라는 것입니다. 로벤스 위원회의 보고서에서 권고한 내용은 국제적으로 다수의 근로감독기관 및 EU국가들에게 광범위하게 채택되었습니다. 근로감독에 있어서 위험성에 기반한 접근방법이 이렇게 널리 퍼지게 된 것은, 대부분의 근로감독기관이 위험성이 낮은 대상으로부터는 인적, 물적 자원을 거둬들여서 가장 위험성이 높은 근로감독 대상의 단속에 해당 자원을 집중해야 한다는 개념을 받아들였다는 것을 의미합니다. 이러한 방법이 가능해지려면, 일정한 방식의 데이터 분석이 필요합니다.

고위험성 기업이나 위험성에 노출된 노동자 집단을 확인하는 분석방법은 잘 개발되어 있습니다. 이러한 위험성 기반의 분석방법은, 대개 직업병이나 산재사고 또는 유해인자 노출 데이터 등과 같은 국가 단위의 통계자료를 바탕으로 합니다. 또한 해당 분석에서 도출된 결과물은 위험성 점검 캠페인이나 산재예방전략 수립 또는 국가 단위나 국제적인 규모로 위험성 점검 우선순위 등을 결정할 때 필요한 기본 자료가 됩니다.

국가 단위로 광범위하게 진행되는 위험성 기반의 분석이 동일 산업군 내에서 위험성에 기반하여 각 기업들에 근로감독 대상 우선순위를 부여하는 경우가 훨씬 자주 발생합니다. 근로감독기관들이 구체적인 위험성에 노출된 근로감독 대상 기업들을 선정하는 공통의 접근방법은, 근로감독관들이 현지에서 수집한 지역정보에 의존하는 것입니다. 스웨덴이나 덴마크의 몇몇 근로감독기관은 감독 대상 기업에 위험성 가산점을 부여하는 위험성 등급시스템을 개발하여 사용하고 있습니다. 각 기업의 특성(예를 들면 기업 규모, 산업 유형, 등록된 산재사고 횟수 등)에 따라 가산점이 부여되는 방식으로, 가산점을 적용하여 위험성 총점이 합산되

고, 총점이 가장 높은 기업의 근로감독 우선순위가 높아지는 방식입니다.

그러나 이렇게 가산점을 부여할 경우 근로감독 대상의 예측 타당성을 나타내는 총점이 상대적으로 낮게 나와서, 고위험성 기업과 저위험성 기업을 구분하는 데 있어 적절하지 않은 결과를 도출하는 문제가 생길 수도 있습니다.

빅데이터와 머신러닝

기업들 중에서 근로감독 우선순위를 매기는 과정은 건초 더미에서 바늘을 찾는 것과 비슷합니다. 이 경우, 건초는 수십만 개에 달하는 검색 대상 기업에 비유되고 이 중에 극히 일부 기업만이 용납할 수 없는 위험성 수준을 가진 바늘에 비유될 수 있습니다. '건초 더미에서 바늘을 찾는 것'은 빅데이타와 머신러닝이 효과적으로 능력을 발휘할 수 있는 분야입니다.

머신러닝 알고리즘을 사용하는 주된 목적은 예측, 재분류 또는 평가 등의 작업을 실행할 때 유용하게 사용될 수 있는 통계학적 모델을 제공하는 것입니다.

예를 들어 암과 관련된 예측을 하는 분야에서, 연구진들은 암 감염성이나 암 재발 가능성 또는 암 생존률 등을 예측하는 데 30년 이상 머신러닝 알고리즘을 사용해오고 있습니다.

내용적으로 볼 때, 암과 관련된 예측을 하는 것과 위험성에 기반하여 근로감독 대상을 선정하는 것은 아주 동떨어진 분야로 보입니다.

그러나, 두 분야는 건초더미에서 바늘을 찾는 것과 같은 고난이도의 예측작업이라는 면에서 공통점을 갖습니다.

머신러닝 알고리즘은 '지도학습'과 '비지도학습'으로 구분할 수 있습니다. 지도학습에서, 머신러닝 알고리즘은 다양한 (다른 요소의 영향을 받지 않는) 독립변수들로부터 예측치가 도출될 수 있는 (위험성 수준 등의) 종속변수들로 구성됩니다. 물론, 정밀한 예측치가 도출되려면 독립변수와 종속변수 사이에 일정 수준 이상의 상관관계가 있어야 합니다.

비지도학습에서, 머신러닝 알고리즘이 예측해야 할 종속변수는 없지만 각각의 위험성 데이터를 내용의 유사성에 따라 일련의 위험성 그룹으로 묶는 역할을 합니다.

스웨덴이나 덴마크의 근로감독기관에서 개발된 가산점 방식과 비교해 보면, 머신러닝에서 사용되는 알고리즘은 주로 반복적인 시행착오(trial and error) 과정을 거치면서 예측 정확도를 점진적으로 높여나가는 방식을 사용합니다. 다른 말로 표현하면, 머신러닝 시스템이 이전의 성공(올바른 예측)이나 실패(잘못된 예측)를 통해 학습을 하고, 이렇게 피드백되어 학습된 지식을 바탕으로 더욱 더 정확한 예측을 할 수 있게 되는 것입니다.

(2편에 계속)