ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 September: 21(3): 1~17 http://dx.doi.org/10.13088/jiis.2015.21.3.01

데이터 마이닝과 텍스트 마이닝의 통합적 접근을 통한 병사 사고예측 모델 개발

윤승진
국방대학교 군사운영분석학과
(ysstudyhard@naver.com)

김수환
국방대학교 군사운영분석학과
(ksuhwan@kndu.ac.kr)

신경식
이화여자대학교 경영대학
(ksshin@ewha.ac.kr)

최근, 군에서 가장 이슈가 되고 있는 문제는 기강 해이, 복무 부적응 등으로 인한 병력 사고이다. 이 같은 사고를 예방하는 데 있어 가장 중요한 것은, 사고의 요인이 될 수 있는 문제를 사전에 식별・관리하는 것이다. 이를 위해서 지휘관들은 병사들과의 면담, 생활관 순찰, 부모님과의 대화 등 나름대로의 노력을 기울이고 있기는 하지만, 지휘관 개개인의 역량에 따라 사고 징후를 식별하는 데 큰 차이가 나는 것이 현실이다. 본 연구에서는 이러한 문제점을 극복하고자 모든 지휘관들이 쉽게 획득 가능한 객관적 데이터를 활용하여 사고를 예측해 보려 한다. 최근에는 병사들의 생활지도기록부 DB화가 잘 되어있을 뿐 아니라 지휘관들이 병사들과 SNS상에서 소통하며 정보를 얻기 때문에 이를 데이터화 하여 잘 활용한다면 병사들의 사고예측 및 예방이 가능하다고 판단하였다. 본 연구는 이러한 병사의 내부데이터(생활지도기록부) 및 외부데이터(SNS)를 활용하여 그들의 관심분야를 파악하고 사고를 예측, 이를 지휘에 활용하는 데이터마이닝 문제를 다루며, 그 방법으로 토픽분석 및 의사결정나무 방법을 제안한다. 연구는 크게 두 흐름으로 진행하였다. 첫 번째는 병사들의 SNS에서 토픽을 분석하고 이를 독립변수화 하였고 두 번째는 병사들의 내부데이터에 이 토픽분석결과를 독립변수로 추가하여 의사결정나무를 수행하였다. 이 때 종속변수는 병사들의 사고유무이다. 분석결과 사고 예측 정확도가 약 92%로 뛰어난 예측력을 보였다. 본 연구를 기반으로 향후 장병들의 사고예측을 과학적으로 분석, 맞춤식으로 관리한다면 군대 내 각종 사고를 미연에 예방하는데 기여할 것으로 기대된다.

주제어 : 데이터 마이닝, 의사결정나무, 토픽분석, SNS, 병력사고

논문접수일 : 2015년 6월 5일 논문수정일 : 2015년 6월 16일 게재확정일 : 2015년 6월 16일
투고유형 : 국문급행 교신저자 : 김수환

1. 서론

우리 군은 아직까지도 북한과 첨예하게 대립하고 있고 이러한 현 상황으로 인해 불가피하게 징병제를 시행하고 있다. 한창 혈기가 왕성한 젊은이들이 군이라는 특수한 집단에 들어와 복무하며 자신의 의무를 다하고 있으며 대부분은 훌륭히 군복무를 마치고 다시 가족의 품으로 돌아간다. 하지만 그 이면에는 끊임없이 발생하는 군내에서의 사건사고들이 있다는 것을 부인할 수 없다. 과거나 지금에서나 마찬가지로 군내에서는 크고 작은 사고들이 끊임없이 발생하고 있고 이러한 사고를 예방하기 위한 노력도 지속되고 있다. 하지만 작년만 하더라도 윤일병 폭행 사망사건, 임병장 총기난사 사고 등 군의 노력과는 무관하게 우리의 뇌리에 오래 기억될 큰 사건 사고들이 지속적으로 발생했다. 이에 국방부에서는 이러한 사고들을 예방하기 위해 많은 대책들

을 쏟아내지만 그 대책들이 미봉책에 그치는 경우가 많은 것이 현실이다. 사실 가장 근본적인 문제는 국방부 차원이 아닌 말단 중·소대급에서 '병사들을 어떻게 관리하고 사고를 예방할 것인가'이다. 그러한 측면에서 예하부대에서 사고 발생 가능성이 있는 인원들을 정확히 분류하고 그들에게 더욱 관심을 가져주는 것은 집중의 효과를 극대화 하고 사고를 예방할 수 있는 가장 좋은 방법이다.

그렇다면 실제 현장에서는 이러한 병사들을 구분하기 위해 어떠한 노력을 하고 있을까? 우선 가장 기본이 되는 것은 병사들의 생활지도 기록부 활용과 면담이다. 이것을 통해 특이사항을 파악하고 각종 검사들을 실시하여 사고를 유발할 가능성이 높은 병사를 식별하려 노력한다. 하지만 이렇게 충분한 데이터를 가지고 있음에도 불구하고 그 데이터를 효과적으로 활용하는 사례는 많지 않다. 또한 현재 사고를 식별하는데 가장 중요한 역할을 하는 면담으로는 얻을 수 있는 정보가 한정되어 있고 그들의 진심을 이끌어 내는 데는 많은 시간과 노력이 필요하다는 문제점이 있다. 하지만 사이버지식정보방이 확대되고, 컴퓨터를 사용할 기회가 많아지다 보니 병사들이 자연스럽게 SNS상에서 꽤나 자신들의 감정을 솔직하게 표출하고 외부인과 대화를 한다. 이를 적절히 활용한다면 보다 쉽게 그들의 속마음을 확인할 수 있고 이를 지휘참고 자료로도 충분히 활용할 수 있을 것이다. 이런 자료들을 활용, 데이터 마이닝을 통해 병사들의 사고를 예측하는 모델을 만들 수 있다는 아이디어 하에 본연구를 시작하게 되었으며, 기존의 데이터 및 병 SNS 자료를 과학적으로 분석하여 사고를 예측할 수 있는 방안을 제시하였다.

데이터 마이닝 기술은 일반적으로 '대량의 데이터 집합으로부터 유용한 정보를 추출하는 것'(Hand. D, 2001)으로 정의된다. 기존에는 구조화된 정형 데이터를 활용하는 형태에서 최근에는 웹과 소셜 미디어 등을 통해 급증하고 있는 텍스트 형태의 비구조화된 비정형 데이터를 분석하여 새롭고 유용한 정보를 얻기 위한 노력이 확산되고 있다.(Liu, 2012) 본 연구에서는 병사들의 기존 생활지도기록부 상 내부 데이터와 그들의 SNS를 토픽분석한 자료를 포함하여 데이터 마이닝을 하여 모델을 만들고 모델을 통해 사고를 예측할 수 있는 방안을 제시하고자 한다. 구체적으로는 (1) 병사들의 SNS 토픽 분석을 통해 그들의 관심 분야를 파악하고 (2) 이를 독립변수화 하여 기존의 내부 데이터들과 함께 의사결정나무를 만들고, 그 결과를 분석할 것이다.

본 논문의 이후 부분은 다음과 같이 구성된다. 제 2장에서는 본 연구의 수행을 위한 선행 연구들을 간략히 요약하고 군과 민간에서의 예측 및 데이터 마이닝에 관한 논문들을 리뷰하며, 제 3장에서는 본 연구의 모형 및 연구방법을 제시한다. 제 4장에서는 데이터를 기반으로 모델을 구축하고 실험 및 결과를 분석하며, 마지막 5장에서는 본 연구의 기여 및 한계 그리고 향후 발전방향을 제시한다.

2. 관련 연구

데이터 마이닝을 통한 예측모형은 다양한 분야에서 많이 연구되었으며 그중 대표적인 것은 부도예측 및 주가예측 분야다. 국가적, 개인적으로 많은 손실을 가져오는 기업의 부도를 예측모형화 하고자 하는 시도는 1930년대 이후 지속적으로 발전하였다. 특히 1966년 Beaver는 통계방

법론을 이용하여 연구방법을 적용하였다.(Beaver, 1966) 이후에도 다변량 판별분석, 로짓분석, 프로빗 분석, 다중회귀분석과 같은 다양한 통계방법론이 예측 정확도를 높이기 위하여 많은 연구자들에 의해 사용되었다.(Martin, 1977; Hanweak, 1977; Johnson, 1979; Emery and Cogger, 1982; Casey et al., 1986). 이후 1980년대부터는 인공신경망 기법과 인공지능 기법들이 연구에 많이 도입되었고 2000년대에는 유전자 알고리즘을 활용한 인공신경망 모형(Hong, 2003; Ok, 2009)까지 등장하여 이를 통해 모형의 예측력을 향상시키는 방안들이 제시되고 있다.

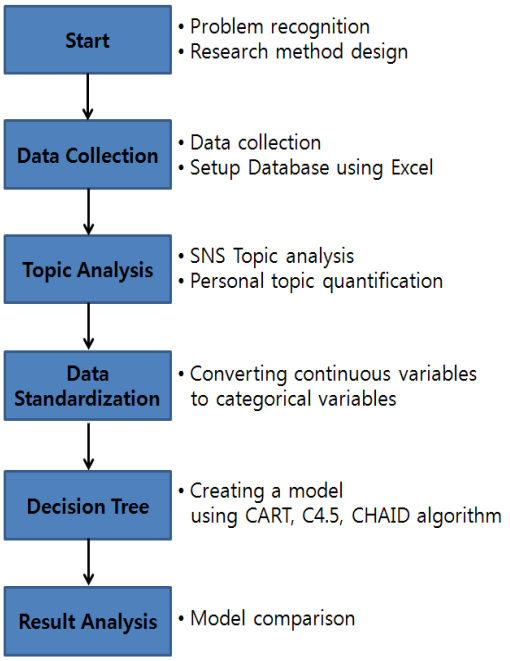
주가예측분야에서도 마찬가지로 데이터 마이닝을 활용한 활발한 연구가 지속되고 있는데 인공신경망을 이용하여 모의 주식투자를 하여 높은 수익률을 달성하기도 하였고 (Bergerson and Wunsch, 1991) 퍼지시스템과 신경망을 결합시킨 주식 예측 시스템(Wei, 2007)도 개발되었다. 최근에는 오피니언 마이닝을 기반으로 하여 소셜 네트워크를 분석해서 그 다음날의 주가를 예측하는 방법론까지 등장(Kim, 2012)하였다.

군에서의 데이터 마이닝 연구는 주로 인력할당과 관련된 연구가 다양하게 진행되어 왔다. 개인의 임무 수행 능력을 고려하여 개인의 임무를 할당하는 모형, 정해진 기간 내에 임무를 수행하기 위하여 개인의 임무 순서와 일정을 결정하는 모형, 다양한 현실적 제약 조건을 고려하여 최소의 인원으로 최적의 임무를 수행하는 모형 등 다양한 목적을 가진 모형이 연구되었다. 구체적으로 AHP기법과 합의방법을 이용하여 군사특기 분류 시 고려해야 할 의사결정요소의 종합가중치 산출방법과 신병 개인의 정성적 속성(임무 충족도)을 정량화 하는 방법을 제시한 논문(Kim, 1998), 공군 비행단에서 복무기간 단축에 따른 정원확보 부담을 경감하고 부대내 숙련급 병사 비율을 증가시켜 전투력 저하의 문제점 해결을 위해 병사의 배속방안을 클러스터링과 의사결정나무를 활용하여 분류한 모델(Kang, 2010)과 통계적 클러스터링을 활용한 새로운 신병분류 방법 제안(Jung, 2009)등이 있다.

앞에서 언급했던 데이터 마이닝 연구들은 단일형태 데이터(정형 혹은 비정형데이터)를 가지고 모델을 개발하였다. 하지만 정형데이터만을 기준으로 한 분석은 과거의 자료를 근거로 구성한 것이기 때문에 실시간성이 떨어지고 비정형데이터만을 기준으로 한 분석은 내부의 신뢰성 및 정보력이 있는 데이터들을 사용하지 않는다는 한계가 존재한다.

3. 사고예측모델 구성

군에서는 아직까지 객관적이고 정량적인 근거자료에 의한 합리적인 의사결정체계가 정립되어 있지 않고 많은 부분 주관적이고 정성적인 의사결정이 주를 이룬다.(Kim, 2014) 이에 본 논문에서는 병사 SNS를 통해 핵심 키워드를 도출하고 기존의 생활지도기록부 데이터와 접목하여 사고를 예측할 수 있는 모델을 만들어 보고자 한다. 병 SNS 핵심 키워드 추출을 위한 방법으로는 텍스트 마이닝 기법 중 토픽분석을 수행하고 사고예측 모델은 의사결정나무 방법을 사용한다. 본 논문에서 제시하는 사고예측모형 구성 절차를 세부단계별로 나타내면 <Figure 1>과 같다.



〈Figure 1〉 Prediction model configuration

분석 모델에서 데이터는 정형데이터 뿐 아니라 비정형 데이터까지 포함되어 있다. 이 비정형 텍스트 데이터는 토픽분석을 통해 의미를 도출하였다. 텍스트는 사용자들이 스마트 시대에 정보를 표현하고 획득하는 가장 일반적인 방식(Witten, 2004)인데 최근에는 이러한 텍스트에 대해 분석을 하고 의미있는 정보를 추출하기 위한 연구가 지속적으로 수행되고 있다. 먼저 텍스트 형태의 데이터를 분석하기 위해서는 텍스트 전처리 단계(형태소 분석, 의미정보 변환 및 추출)를 실시한 후 텍스트 문서 집합 내에 잠재된 주제를 도출하는 토픽분석을 수행한다. 토픽분석은 문서 집합 내에서 동시 출현빈도가 높은 단어들을 기준으로 유사한 주제로 문서들을 그룹화 한다. 특히, 개별문서와 주제가 일대일 매칭 개념이 아닌 여러 주제를 다룰 수 있다는 점을 가정하고 있다. 병사들의 SNS에서 추출된 토픽은 단어들의 집합으로 파악할 수 있으며, 문서집합 내에서의 해당 토픽의 출현 빈도는 그것에 대한 관심도를 반영한다고 볼 수 있다. 이러한 관심도를 산출함으로써 병사들의 관심분야 도출 및 병사 개개인의 토픽에 대한 관심도 도출이 가능하다. 조금 더 구체적으로 살펴보면 분석의 최소 단위는 각 문서가 벡터공간모델(Vector Space Model) (Albright, 2006)을 이용하여 표현되며, 각 문서에 사용된 용어의 빈도에 따라 해당 문서의 주제 및 특성이 요약된다. 대부분의 경우는 용어의 빈도수 보다는 TF-IDF(단어 빈도-역문서 빈도, Term Frequency-Inverse Document Frequency) 값을 많이 활용한다.(Salton et al.,1983) 이는 여러 문서 집합으로부터 특정 단어가 얼마나 중요한가를 판단할 수 있는 값이다. 여기서 IDF는 단어 빈도뿐만 아니라 DF(문서빈도, Document Frequency)의 역수를 취한 IDF(역문서 빈도, Inverse Document Frequency)를 고려한다. IDF는 특정 단어가 문서 집합 내에서 얼마나 공통적으로 출연하는지를 나타낸 값이다. TF-IDF에 기반한 분석에서 각 문서는 용어수 만큼의 차원을 갖게 되며, 이는 (문서)×(용어) 형태의 행렬로 표현될 수 있다. 하지만 이러한 행렬은 그 크기가 너무 크기 때문에 문서간 유사성 측정을 위해 각 문서는 SVD(Singular Value Decomposition)등의 차원축소기법을 통해 저장된다.(Albright, 2006). 이러한 비정형 텍스트 문서가 정형화 되면 다음 과정은 일반적인 정형 데이터에 대한 대이터마이닝 기법(인공신경망, 의사결정나무, SVM 등)을 활용하여 분류하거나 텍스트 군집분석 또는 토픽분석을 통해 문서들을 군집화 한다. 하지만 토픽 분석은 기존의 군집분석 등의 유사문서 그

룹화 기법과는 다르게 앞에서 언급한 바와 같이 다대다 대응이 가능하다는 차이가 있다. 하나의 문서에 꼭 하나의 주제가 아니라 여러 주제를 포함할 수 있다는 관점에서 그 현실을 잘 반영한다고 이야기 할 수 있다.

데이터표준화는 의사결정나무에 사용하기 용이하고 설명력을 높일 수 있도록 범주형으로 표준화 하였고 그 후 의사결정나무를 구성하였다. 의사결정나무 구성은 기존에 주로 사용되고 있는 CHAID, C4.5, CART 알고리즘을 사용하여 모형을 구성후 오분류율이 가장 작은 모델을 선택하기로 한다. 의사결정나무를 통해 구성한 모델은 어떤 독립변수가 모델을 구성하는데 영향을 미치는 지가 한눈에 파악되기 때문에 지휘관들이 지휘에 활용하기에 가장 적합하다고 할 수 있다.

4. 실험 결과 및 분석

4.1 실험데이터

본 논문에서 사용되어진 데이터는 전방 GOP 부대 4개사단 350명의 생활지도기록부 및 SNS 중 Facebook 자료이다. 데이터 형태로, 생활지도기록부는 정형데이터, Facebook은 비정형 텍스트형태의 데이터로 이루어져 있다. 이때, 각 사단별 사고자 및 비사고자는 무작위로 수집되었고 사고자의 평균 비율은 18%였으며 이 모든 데이터는 익명으로 처리하여 개인 정보 유출에 대한 문제가 발생하지 않도록 하였다. 최근 병사들은 대부분 SNS를 사용하고 예전과는 다르게 전 부대에 사이버지식정보방이 보편화 되어 있고 자신들의 자유시간에는 인터넷 사용에 제약이 없다. 최근에는 지휘관들도 병사들과의 의사소통을 위해 SNS 친구맺기가 많이 일반화 되어있어 지휘관의 아이디로 수집대상 병사들의 SNS를 수집할 수 있었다. 이 연구는 GOP부대에서 적용 가능한 사고예측모델로 제한하였다. 이유는 군은 그 유형에 따라 근무 환경 및 특성이 다르기 때문에 군 전체를 대상으로 모델을 구성하기에는 일반화를 달성하기 어렵다. 또 GOP 부대의 특성상 북한과 인접하여 대치하며 임무를 수행하고 있어 그 긴장도가 높으며 소총과 실탄, 수류탄 등의 무기를 소지하고 근무를 서기 때문에 여기서 발생하는 사고들은 자칫 대형사고로 이어질 수 있어 어느 부대보다 더 사고예방이 중요하기 때문이다.

4.2 병사 SNS 토픽분석

본 단계에서는 데이터 마이닝 상용 도구 중 하나인 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 병사 350명의 SNS글 약 20,000건에 대한 토픽분석을 수행하였다. 이때 SNS데이터는 엑셀로 정리하여 DB화 하였다.

수집한 데이터의 전처리 과정(형태소 분리, 불용어 처리, 어간 추출, 단어별 가중치 산출 등)을 수행하였고 이때 키워드별 가중치로써 TF-IDF를 사용하였으며, 토픽 분석을 위해 총 250개의 키워드를 추출하였다.

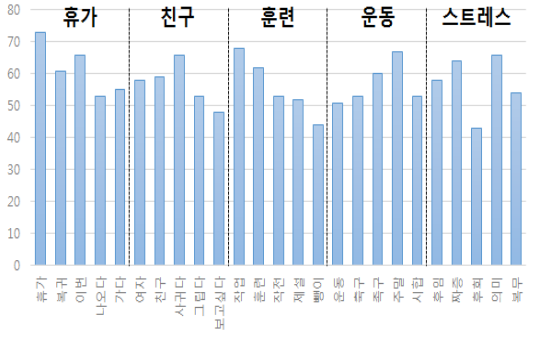
토픽분석 시 전체 토픽의 수는 10개, 토픽별 키워드의 수 또한 5개로 한정하였으며, 분석된 토픽 중 의미 도출이 가능하였던 상위 5개를 뽑아 사용하였고 그 결과는 <Table 1>와 같은 형태로 나타난다. 이 때, 토픽명에 대한 명명(Naming)은 연구자가 직접 판단해 작성해야 한다.

〈Table 1〉 SNS topic analysis

Rank	Topic	Keywords	Document threshold
1	휴가	휴가, 복귀, 이번, 나오다, 가다	0.245
2	친구	여자, 친구, 사귀다, 그립다, 보고싶다	0.227
3	훈련	작업, 훈련, 작전, 제설, 빵이	0.207
4	운동	운동, 축구, 족구, 주말, 시합	0.192
5	스트레스	후임, 짜증, 후회, 의미, 복무	0.191

여기서 문서임계치는 SAS에서 제시하는 문서가 해당 토픽에 포함되어야 하는 최소 토픽값을 의미하고 분석 시에도 이 문서임계치를 적용하였다.

해당 토픽 별 단어의 빈도는 <Figure 2>와 같다.



〈Figure 2〉 Word frequency of SNS topic

이렇게 토픽분석을 통해 핵심 토픽 5가지를 도출하였으며 이는 병사들의 현재 관심사라고 볼 수 있다. 병사들은 예상했던 바와 같이 휴가와 친구에 관심이 많았으며 생각보다 작전, 훈련, 작업에 대한 부담감을 SNS를 통해 토로하는 모습도 자주 보였다. 확실한 것은 면담 시에는 얻기 힘든 그들의 솔직한 감정과 관심주제가 잘 파악됨을 볼 수 있다.

이 토픽분석 결과를 바탕으로 개인별 SNS글의 초기대응도가 파악되는데 각 글별 <Table 1>의 문서 임계값 이상의 대응도를 갖는 경우 “1”, 그렇지 않은 경우 “0”으로 카운팅 하여 각 인원을 <Table 2>와 같이 토픽별로 점수화 한다. 여기서 ID는 각 병사가 쓴 글들이고, 표 안의 점수는 각 글들이 각 토픽에 대해 갖는 초기대응도이다.

〈Table 2〉 Example of personal topic quantification

(a) Initial Matrix

	ID	Topic1	Topic2	Topic3	Topic4	Topic5
A	1	0.324	0	0.373	0	0
	2	0.432	0.034	0	0	0.311
	3	0	0	0	0.321	0
B	1	0.234	0.337	0.423	0	0
	2	0.121	0.342	0	0	0
	3	0.021	0	0.134	0.579	0.035
	4	0.421	0	0.032	0.341	0

(b) Refined Matrix

	ID	Topic1	Topic2	Topic3	Topic4	Topic5
A	1	1	0	1	0	0
	2	1	0	0	0	1
	3	0	0	0	1	0
B	1	0	1	1	0	0
	2	0	1	0	0	0
	3	0	0	0	1	0
	4	1	0	0	1	0

〈Table 3〉 Converting each enlisted men' s topic points to percentage

N0.	Topic1	Topic2	Topic3	Topic4	Topic5
1	7	4	2	3	1
2	3	2	5	2	3
3	10	15	5	20	4
4	3	2	4	3	5
5	9	1	2	2	6
6	4	6	3	5	10
7	21	1	5	2	4
8	3	4	1	5	9
9	13	7	3	4	7
10	3	2	1	1	5
11	9	2	3	1	4
12	5	1	3	1	2

⇨

Topic1	Topic2	Topic3	Topic4	Topic5
41%	24%	12%	18%	6%
20%	13%	33%	13%	20%
19%	28%	9%	37%	7%
18%	12%	24%	18%	29%
45%	5%	10%	10%	30%
14%	21%	11%	18%	36%
64%	3%	15%	6%	12%
14%	18%	5%	23%	41%
38%	21%	9%	12%	21%
25%	17%	8%	8%	42%
47%	11%	16%	5%	21%
42%	8%	25%	8%	17%

이렇게 점수화 된 토픽점수는 병사들의 SNS 사용량에 따라 그 값이 천차만별이기 때문에 이를 표준화 하는 작업도 필요하다. 그래서 <Table 3>에서 보는바와 같이 개인별 토픽 전체 수에 각 토픽의 수를 나누어 %로 나타내었다.

4.3 데이터 표준화

각 데이터의 변수는 범주형 변수와 수치형 변수가 혼합되어 있다. 하지만 실제 부대에서 어떠한 변수에 대해 평가할 때 점수화 보다는 우수, 보통, 저조 등과 같이 범주화 시켜 사용하는 것이 대부분이고 의사결정나무의 특성상 연속형 데이터 보다는 범주형 데이터를 처리하는 것이 결과를 더욱 명확히 해석할 수 있기 때문에 <Table 4>, <Table 5>와 같이 데이터를 범주형으로 표준화하였다.

먼저 나이, 체력, 종교, 키, 몸무게 등의 변수는 이진화로 표현하였다. 나이와 같은 경우 병사들의 기본 입대 나이인 21살을 기준으로 그보다 많은 인원들을 1로 분류하였다. 체력은 일병 진급 기준인 2급을 기준으로 분류하였다. 키와 몸무게는 수집집단의 평균을 적용하였다. 종교와 부모는 양친이 아동의 정서와 관련한 논문자료(Lee, 2004)를 참고하였다. 이때, 욕, SNS 글수, 친구수는 병사들의 facebook을 통해 확인하였고 욕은 존재 유무로, 글수 및 친구수는 수집대상 평균을 기준으로 이진화 하였다.

〈Table 4〉 Categorical variables 1

Variables	Classification standard	
	Low(0)	High(1)
Age(x1)	$x_1 \leq 21$	$x_1 \geq 22$
Strength(x2)	$x_2 \leq Level3$	$x_2 \geq Level2$
Religion(x3)	None	Otherwise
Height(x4)	$x_4 \leq 175$	$x_4 > 175$
Weight(x5)	$x_5 \leq 65$	$x_5 > 65$
Parents(x6)	Single parents	Otherwise
Brother(x7)	Only son	Otherwise
Curse(x8)	None	Exist
Number of SNS (x9)	$x_9 \leq 45$	$x_9 > 45$
Number of friends(x10)	$x_{10} \leq 383$	$x_{10} > 383$

〈Table 5〉 Categorical variables 2

Variables	Classification standard		
	Low(0)	Normal(1)	High(2)
Training(x11)	$x < 70$	$70 \leq x \leq 90$	$x > 90$
Barrack score(x12)			
Tenacity(x13)	$x < 40$	$40 \leq x \leq 50$	$x > 50$
Research(x14)			
Creativity(x15)			
Consideration(x16)			
Initiative(x17)			
Sincerity(x18)			
Topic(Vacation)(x19)	$x < 20$	$20 \leq x \leq 30$	$x > 30$
Topic(Friend)(x20)			
Topic(Training)(x21)			
Topic(Sports)(x22)			
Topic(Stress)(x23)			

교육훈련에서부터 토픽분석까지는 부대에서 주로 사용하는 저조, 보통, 우수의 등급으로 구분하였다. 이때, 신인성검사 점수(강인성, 탐구성, 창의성, 배려성, 주도성, 성실성)는 KIDA에서 만든 매뉴얼(KIDA, 2012)을 기준으로 점수를 구분하였고 교육훈련 및 내무생활은 진급측정 시 사용되는 기준을 적용하였다. 토픽분석 후 생성된 각 토픽은 위에서 언급한 방법과 같이 %로 표현하였고 그 범위는 4분위수를 기준으로 분류하였다.

4.4 의사결정나무 구성

사고예측모델 개발을 위한 분석 접근 방식은 크게 두 가지이다. 첫 번째는 기존 생활지도기록부 데이터만을 독립변수로 활용하여 의사결정나무를 구성해 보는 것과 두 번째는 비정형 데이터를 정형 데이터화 한 토픽분석결과를 포함하여 의사결정나무를 구성해 보는 것이다. 이 두 실험을 통해 군대에 축척되어 있는 병사들의 데이터에서 data driven 기법으로 사고를 예측해 볼 것이고, 또한 기존의 사고예측 모형에서는 볼 수 없었던 실시간 SNS 비정형 데이터를 정형데이터와 같이 활용해 사고 예측 정확도를 더욱 높일 수 있는 모델을 구축하려 한다.

이때, 총 350개의 데이터 중 의사결정나무의 학습을 위한 training set(279개), 학습을 통해 구축된 모델의 검증을 위한 validation set(71개)으로 구분하여 실험을 실시하였다. training, validation 데이터 군은 모두 각각 사고자와 비사고자를 동일한 비율로 구성하였다. 또한 각 실험시 사용 모델은 CART(지니지수 이용), CHAID(카이제곱 통계량의 p값 이용), C4.5(엔트로피 지수 이용)를 모두 사용해서 결과를 보고 오분류율을 기준으로 모델을 선택하기로 하였다.

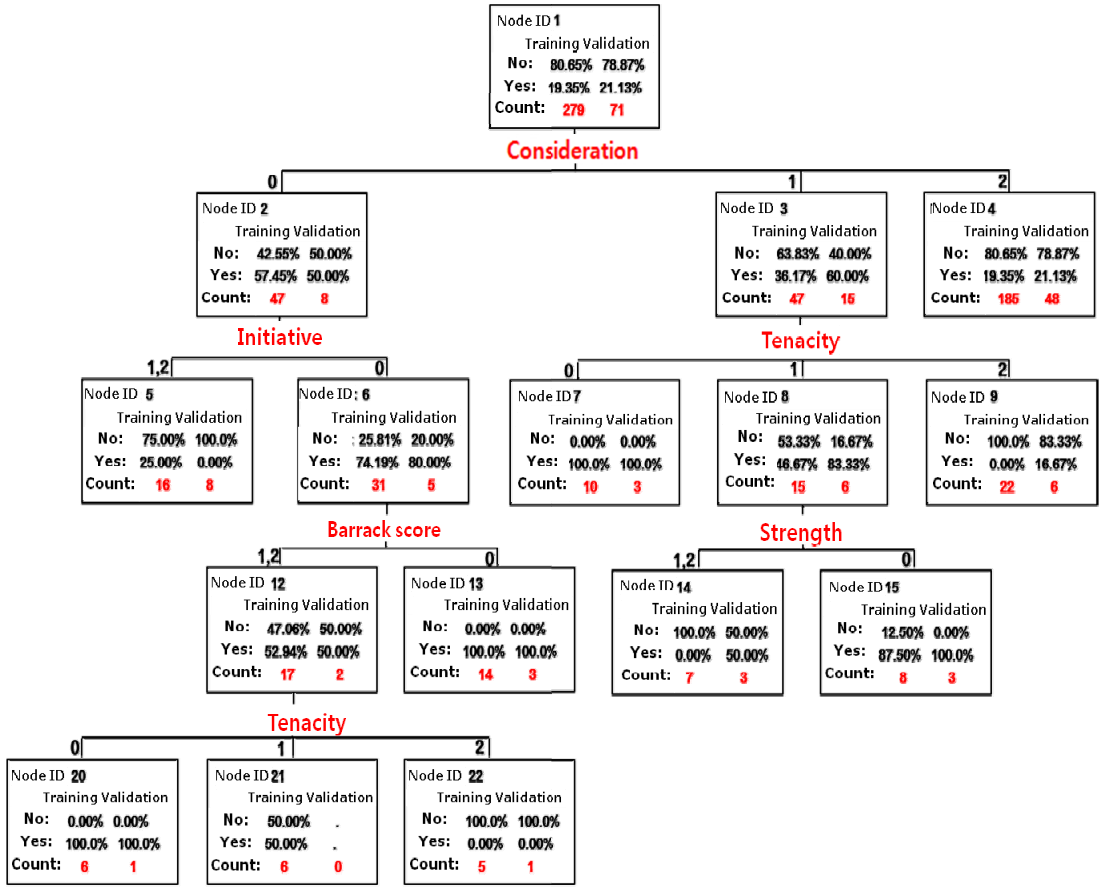
4.4.1 내부데이터를 활용한 의사결정나무 구성

앞에서 언급한 바와 같이 첫 번째 모형은 병사들의 생활지도기록부에서 확인할 수 있는 내부데이터만을 가지고 의사결정나무를 구성하였고 이에 따른 오분류율은 <Table 6>와 같다.

여기서 오분류율이라 함은 사고자를 비사고자

〈Table 6〉 Misclassification of algorithms

	CHAID		CART		C4.5	
data	training	validation	training	validation	training	validation
misclassification rate	12.1%	11.5%	13.5%	13.3%	13.5%	13.2%



〈Figure 3〉 Decision tree #1

로 비사고자를 사고자로 잘못 식별하는 비율을 의미하고 당연히 오분류율이 낮은 모델이 예측 정확도(1-오분류율)가 높다고 할 수 있다.

첫 번째 실험에서의 모형은 오분류율이 가장 낮은 CHAID(카이제곱 통계량 사용)모델을 선택하였고 트리 구성 결과는 <Figure 3>와 같다. 선정된 주요 변수들을 살펴보면 신인성검사 배려성, 주도성, 강인성, 내무생활점수, 체력점수가 선택되었다. 특히 신인성검사 적성적응도 점수인 배려성, 주도성, 강인성 점수는 트리를 구성하는데 중요한 역할을 하였다. 이를 통해 신인성검사 결과에 포함되어 있는 적성적응도 점수는 중요한 지휘참고자료라는 것을 알 수 있었다.

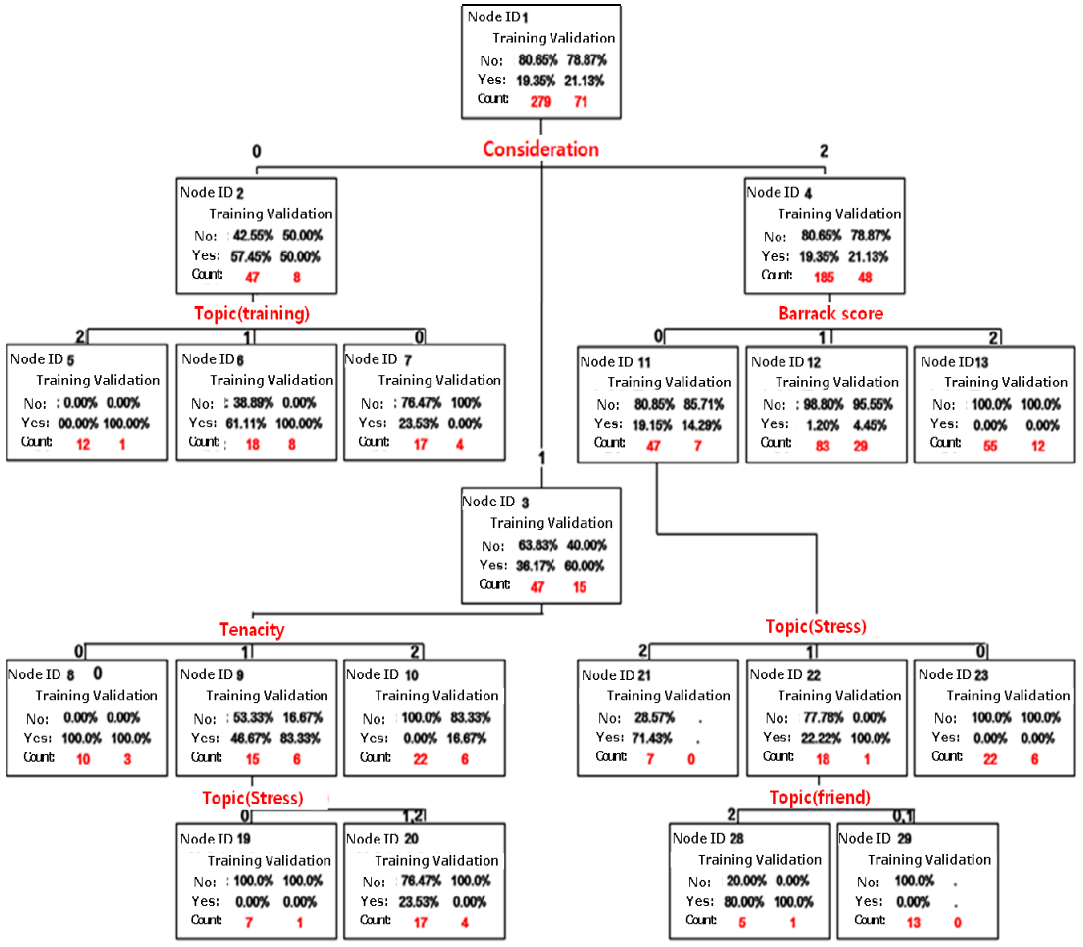
특히, 배려성이 50점 이상인 인원들은 매우 높은 비율(80%)로 사고자에 포함되지 않았다. 이는 신인성검사 시 배려에 대해서 긍정적으로 설문을 한 인원들은 비교적 사고와 연관성이 적다는 것을 보여준다.

4.4.2 병사 SNS를 포함한 의사결정나무 구성

그렇다면 이제 토픽분석을 통해 확보한 SNS

〈Table 7〉 Misclassification of algorithms

	CHAID		CART		C4.5	
data	training	validation	training	validation	training	validation
misclassification rate	9.3%	10.1%	9.3%	10.2%	7.8%	8.3%



〈Figure 4〉 Decision tree #2

자료를 독립변수로 추가하여 두 번째 의사결정 나무 모형을 구성해 보자. 그 방법은 앞 절의 방법과 동일하고 독립변수에 토픽분석결과(훈련, 친구, 휴가, 운동, 스트레스/짜증)가 추가되는 형태이다.

모형구축 결과 오분류율이 가장 작은 C4.5 모델을 채택하였고 세부내용은 <Table 7>과 같다.

이 실험에서는 C4.5(엔트로피지수 사용)모델

이 선택되었고 트리구성 결과는 <Figure 4>과 같다. 선정된 주요 변수들을 살펴보면 내부데이터로는 배려성, 강인성, 내무생활 변수, 외부데이터로는 훈련, 스트레스, 친구가 포함되었다. 여기서 훈련과 스트레스에 관한 언급에서 주목해 볼 점은 이러한 훈련 및 스트레스의 언급 비율이 높은 인원들이 사고의 확률이 높다는 것이다. 실제로 데이터를 직접 확인해 본 결과 훈련 및 스트레스 언급은 군 생활 불만, 훈련에 대한 부담감, 짜증과 많이 연결되는 것으로 확인되었다. 특히, 배려성이 낮은 인원들 중 훈련과 관련된 언급이 SNS상에 많았던 인원들은 사고 확률이 100%로 나타났다. 이 병사들의 SNS를 구체적으로 살펴 본 결과 훈련에 대한 부담감과 자신의 고충을 SNS에 많이 언급하고 있다는 것을 확인할 수 있었다.

이와 같이 내부 데이터만을 활용하여 데이터 마이닝을 했을 때에도 의미있는 결과(예측 정확도 87%)가 도출되었으나 SNS데이터를 추가하여 데이터 마이닝을 실시하니 예측 정확도 측면에서 4.3%(예측 정확도 92%)정도 값을 향상 시킬 수 있었다.

4.4.3 모형의 통계적 검증

이와 같은 두 모형간의 성과가 통계적으로 유의한지 알아보기 위하여 McNemar Test를 실시하였다. McNemar Test는 비모수통계의 일종으로 Chi-Square분산을 이용하여 실험 전후의 차이가 의미가 있는지를 찾는데 유용하게 이용된다. 검증 결과는 <Table 8>과 같다.

〈Table 8〉 Result of McNemar Test

Categories	chi-squared	p-value
Result	13.0667	0.0003006

McNemar's chi-squared 값은 13.0667, p-value는 0.0003006으로 0.01이하로 내부데이터로만 구성한 모델에 비해 SNS데이터를 포함하여 구성한 두 번째 모델의 효과가 매우 유의하다고 확인할 수 있다.

4.4.4 로지스틱 회귀분석과 예측력 비교 실험

범주형 데이터를 분석하는 방법에는 의사결정나무를 제외하고도 판별분석, 로지스틱 회귀분석방법 등이 있다. 그중 이항 로지스틱 회귀분석은 종속변수와 독립변수가 모두 범주형 변수일 경우 적용이 가능하다.

데이터는 의사결정나무에 사용되었던 범주형 변수를 동일하게 사용하였다. 변수 선택 방법은 stepwise selection을 활용하였고 각 모델별 선택된 변수와 오분류율은 제시된 <Table 9>와 같다.

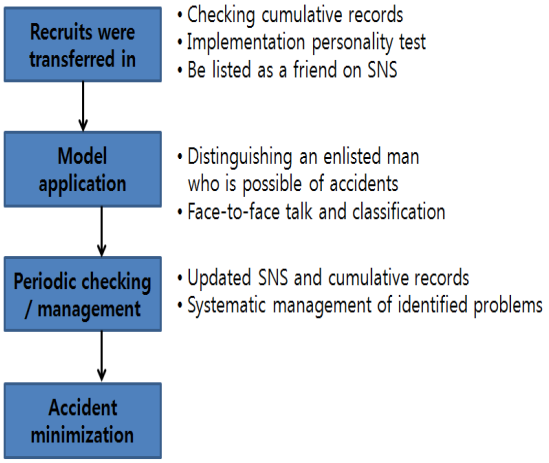
〈Table 9〉 Result of Logistic regression analysis

Categories	Model #1	Model #2
Selected Variables	Consideration Tenacity, Training Score, Strength	Consideration, Topic(Training), Tenacity, Topic(Sports)
misclassification rate	16.7%	12.3%
Difference of previous model	3.2%	4.5%

기존 의사결정나무 모델에 비해 로지스틱 회귀분석 결과 오분류율이 각각 3.2%, 4.5% 만큼 높아졌다. 이는 의사결정나무가 오분류율이 더 낮아 예측력이 높고 각 변수들에 대한 설명력 또한 높아 이를 방법론으로 사용하는 것이 타당할 것으로 판단된다.

4.4.5 모델의 적용

지금까지 데이터를 기반으로 하여 병사들의 사고예측모델을 구성하였다. 이 사고예측모델은 아래 <Figure 5>와 같이 중·대대급에 즉시 적용할 수 있다.



〈Figure 5〉 Application of accident prediction model

사고예측모델에 사용되는 데이터 중 생활지도기록부 데이터는 최초 병사들이 입대하여 신병교육대에 들어가게 되면 연대통합행정시스템에 축적된다. 이는 지휘관들이 열람할 수 있도록 관리된다. 또한 병사들은 SNS를 비교적 자유롭게 사용하고 있어 지휘관들은 친구맺기를 통해 SNS상에서 그들과 대화가 가능하고 데이터 획득 또한 어렵지 않다. 이렇게 축적된 데이터들은 구축된 사고예측모델을 통해 사고 가능성이 있는 병사가 식별해 준다. 이렇게 식별된 병사는 관심과 사랑이 필요한 병사로 분류하고, 의사결정나무를 통해 확인된 부족한 부분에 대한 선택적 관리가 가능하다. SNS 및 내부데이터는 계속해서 업데이트 되기 때문에 주기적으로 예측모델에 적용하여 그 변화 추이를 살펴 지휘에 참고한다면 효과적인 병력 관리가 가능할 것이다.

5. 결론

본 연구에서는 지휘관들이 손쉽게 획득 가능한 데이터를 기반으로 데이터 분석방법인 토픽분석과 의사결정나무를 적용하여 사고 가능성이 있는 병사를 예측하는 문제를 모델링 하였다. 특히, 생활지도기록부에서 획득 가능한 변수만을 활용한 것이 아니라 SNS와 같은 비정형 데이터를 토픽분석을 통해 정형데이터로 변형, 독립변수로 추가하여 92%의 높은 예측 정확도를 보이는 모델을 구축하였다.

본 연구에서의 의의로는 첫째, 최근 이슈가 되고 있는 텍스트 분석을 기존 정형 데이터 마이닝과 접목시켰다는 점과 둘째는 군에서 사고를 예측하는 정량적 분석방법을 제시하였다는 것이다. 이 모델을 통해 군에서는 사고 가능성이 있는 병사를 조기에 식별하고 이를 집중 관리함으로써 병영 내 사고를 획기적으로 줄일 수 있다고 확신한다.

본 연구의 한계점은 다음과 같다. 의사결정나무의 특성상 연속형 데이터를 처리하는 능력이 신경망이나 통계기법에 비해 떨어지며, 결과적으로 예측력도 감소한다. 따라서 이번 연구 시에도 모든 변수를 범주형으로 변화시켜 사용하였는데 그룹화하는 과정에서 발생하는 치우침을 배제할 수 없다. 또한 의사결정나무 자체가 표본의 크기에 지나치게 민감하다는 점이다. 보다 정확한 모형을 구축하기 위해서는 가능한 한 많은 데이터를 확보하고 모형을 구축해야 한다.

차후 진행될 연구로는 병사 SNS 토픽분석의 구체화다. 논문에서 토픽분석으로 얻었던 결과는 휴가, 친구, 훈련, 운동, 스트레스 5가지 토픽이다. 하지만 이 같은 토픽은 전체의 데이터에 대해서 가장 많이 언급된 단어들을 중심으로 분석한 결과이기 때문에 일반적이고 예측 가능한 단어들만이 도출되었다. 하지만 실제 지휘 간 SNS를 참고할 때 지휘관들이 확인 하는 것은 이러한 일반적인 관심사는 물론이고 자살, 충동, 사고, 탈영 등 사고와 직접적으로 연결되는 단어들이다. 이것을 고려해 본다면 토픽분석 시에도 비록 빈도수는 낮으나 사고와 직접적으로 연관되는 단어들을 도출하고 이를 몇 개의 토픽 그룹으로 분류하여 분석을 하게 된다면 더욱 정확하고 예측력이 높은 분석 모델이 될 수 있을 것이라 생각한다. 또한 최근 그 유용성이 입증되고 있는 인공신경망 모형을 이용한 사고 가능 병사 예측이다. 이때 독립변수 선정은 의사결정나무에서 식별된 변수들을 사용할 수 있다. 이렇게 구성한 모델과 의사결정나무의 예측력을 비교해 보는 것이 향후 연구과제가 될 수 있을 것이다.

참고문헌(References)

Albright, R., *Taming Text with the SVD*, SAS Institute Inc., 2006.

Beaver, W., "Financial ratios as predictors of failure. Empirical research in Accounting; Selected studies," *Journal of Accounting Research*, Vol. 5(1966), 71～111.

Bergerson, K. and D. C. Wunsch, "A Commodity Trading Model Based on a Neural Network-Expert System Hybrid," *Proceedings of the IEEE International conference on Neural Networks,* Seattle, Washington, (1991).

Casey, C., McGee, V. and C. Stickney, "Discriminating between reorganized and liquidated firms in bankruptcy," *The Accounting Review*, April (1986), 249～262.

Emery, G. W. and K. O. Cogger, "The measurement of liquidity," *Journal of Accounting Research,* Vol. 20, No. 2(1982), 290～303.

Hand, D. J., Mannila, H., and P. Smyth, *Principles of Data Mining*, Cambridge, MA:MIT Press, 2001.

Hanweak, G. A., "Predicting Bank Failure - Research Papers in Banking and Economics," *Financial Studies Section*, FRB, November (1977).

Hong S.-H. and K.-S. Shin, "Using GA based Input Selection Method for Artificial Neural Network Modeling: Application to Bankruptcy Prediction," *Journal of Intelligence and Information Systems*, Vol.9, No.1(2003), 227～249

Johnson, W. B., "The Cross-Sectional Stability of Financial Ratio Patterns," *Journal of Financial and Quantitative Analysis*, Vol. 14, No. 5(1979), 97～108.

Jung, J. B., "A proposal of new method of recruits classification using a statistical clustering," *Proceedings of the Korean Institute of Industrial Engineers,* (2009), 401～411.

Kang, K. Y., "Effective assignment method to promote recruit's proficiency," *Master's Dissertation*, Korea National Defense University, 2010.

KIDA, "Interpretation of Aptitude Adaptation Degree," 2012.

Kim, S.-W, G.-G. Kim, and B.-K. Yoon, "A Study

on a way to usilize Big data Analytics in the Defense Area," *The Korean Operations Research and Management Science Society*, Vol.39, No.2(2014), 133～134.

Kim, H. S., "A study of recruit's assignment method using AHP and goal programming," *Master's Dissertation*, Korea National Defense University, 1998.

Kim, Y.-S., N.-G. Kim, and S.-R. Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining," *Journal of Intelligence and Information System*, Vol.18, No.2(2012), 143～156.

Lee, E. G., and S. Y. Park, "Emotional & Behavioral problems in children from Broken Families," *Journal of the Korean Home Economics Association*, Vol. 42, No.12(2004), 191～204.

Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

Martin, D., "Early Warning of Bank Failure: A Logit Regression Approach," *Journal of Banking and Finance*, Vol. 1, No. 3(1977), 249～276.

Ok, J.-K. and K.-J. Kim, "Integrated Corporate Bankruptcy Prediction Model using Genetic Algorithms," *Journal of Intelligence and Information System*, Vol.15, No.4(2009), 99～120.

Salton G. and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983.

Yang, W. "Stock price predictin vased on fuzzy logic," *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Vol.3(2007), 19～22.

Witten, I, H., Text Mining, *Practical Handbook of Internet Computing*, CRC Press, 2004.

Abstract

Development of the Accident Prediction Model for Enlisted Men through an Integrated Approach to Datamining and Textmining

Seungjin Yoon* · Suhwan Kim** · Kyungshik Shin***

In this paper, we report what we have observed with regards to a prediction model for the military based on enlisted men's internal(cumulative records) and external data(SNS data). This work is significant in the military's efforts to supervise them. In spite of their effort, many commanders have failed to prevent accidents by their subordinates. One of the important duties of officers' work is to take care of their subordinates in prevention unexpected accidents. However, it is hard to prevent accidents so we must attempt to determine a proper method. Our motivation for presenting this paper is to mate it possible to predict accidents using enlisted men's internal and external data.

The biggest issue facing the military is the occurrence of accidents by enlisted men related to maladjustment and the relaxation of military discipline. The core method of preventing accidents by soldiers is to identify problems and manage them quickly. Commanders predict accidents by interviewing their soldiers and observing their surroundings. It requires considerable time and effort and results in a significant difference depending on the capabilities of the commanders. In this paper, we seek to predict accidents with objective data which can easily be obtained. Recently, records of enlisted men as well as SNS communication between commanders and soldiers, make it possible to predict and prevent accidents.

This paper concerns the application of data mining to identify their interests, predict accidents and make use of internal and external data (SNS). We propose both a topic analysis and decision tree method. The study is conducted in two steps. First, topic analysis is conducted through the SNS of enlisted men. Second, the decision tree method is used to analyze the internal data with the results of the first analysis. The dependent variable for these analysis is the presence of any accidents. In order to analyze their SNS,

* M.S. Candidate, Dept. of Military Operation Research, Korea National Denfense University
** Corresponding Author: Suhwan Kim
Dept. of Military Operation Research, Korea National Denfense University
33 2Jayu-ro, Dukyang-gu, Goyang-si, Kyungki-do 412-170, Korea
Tel: +82-2-300-2174, Fax: +82-2-309-9774, E-mail: ksuhwan@kndu.ac.kr
*** Ewha school of Business, Ewha Womans University

we require tools such as text mining and topic analysis. We used SAS Enterprise Miner 12.1, which provides a text miner module. Our approach for finding their interests is composed of three main phases; collecting, topic analysis, and converting topic analysis results into points for using independent variables. In the first phase, we collect enlisted men's SNS data by commender's ID. After gathering unstructured SNS data, the topic analysis phase extracts issues from them. For simplicity, 5 topics(vacation, friends, stress, training, and sports) are extracted from 20,000 articles. In the third phase, using these 5 topics, we quantify them as personal points. After quantifying their topic, we include these results in independent variables which are composed of 15 internal data sets. Then, we make two decision trees. The first tree is composed of their internal data only. The second tree is composed of their external data(SNS) as well as their internal data. After that, we compare the results of misclassification from SAS E-miner. The first model's misclassification is 12.1%. On the other hand, second model's misclassification is 7.8%. This method predicts accidents with an accuracy of approximately 92%. The gap of the two models is 4.3%. Finally, we test if the difference between them is meaningful or not, using the McNemar test. The result of test is considered relevant.(p-value : 0.0003)

This study has two limitations. First, the results of the experiments cannot be generalized, mainly because the experiment is limited to a small number of enlisted men's data. Additionally, various independent variables used in the decision tree model are used as categorical variables instead of continuous variables. So it suffers a loss of information.

In spite of extensive efforts to provide prediction models for the military, commanders' predictions are accurate only when they have sufficient data about their subordinates. Our proposed methodology can provide support to decision-making in the military. This study is expected to contribute to the prevention of accidents in the military based on scientific analysis of enlisted men and proper management of them.

Key Words : data mining, decision tree, topic analysis, SNS, accident of enlisted men

Received : June 5, 2015 Revised : June 16, 2015 Accepted : June 16, 2015
Type of Submission : Fast Track Corresponding Author : Suhwan Kim

저 자 소 개

윤승진
육군사관학교에서 군사학과 전자공학 학사 학위를 취득하고, 현재 국방대학교 군사운영분석학과 석사과정에 재학 중이다. 현역 육군 대위로 전국 각 지역에서 지휘관 및 참모 역할을 수행하였다. 관련 주요 관심분야는 데이터 마이닝, 소셜 네트워크 분석 등이다.

김수환
현재 국방대학교 군사운영분석학과 교수로 재직 중이다. 육군사관학교를 졸업하고, 한국과학기술원(KAIST)와 미국 Texas A&M 대학교에서 산업공학으로 석사 및 박사 학위를 취득하였다. 주요 연구 분야는 정수계획법, 일정계획법, Small & Big data 분석 등이다.

신경식
현재 이화여자대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고 미국 George Washington University에서 MBA, 한국과학기술원 (KAIST)에서 인공지능, 지식기반 시스템 등 지능형 기법을 경영분야에 적용하는 연구로 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 데이터 마이닝과 비즈니스 인텔리전스, 빅데이터 분석/비즈니스 애널리틱스, 인공지능 응용과 지식공학 등이다.