□ 研 究 論 文 □

情韓國國防經營分析學會誌
第33卷 第2號, 2007. 12. 31.

신병 주특기교육 성취집단 예측모형 개발

(Development of newly recruited privates on-the-job Training Achievements Group Classification Model)

곽 기 효(Kihyo Kwak)*, 서 용 무(Yongmoo Suh)**

초 록

국방부에서 발표한 '국방개혁에 관한 법률'에 따라 2014년까지 현역병들에 대한 복무기간이 단계적으로 단축될 예정이다. 이에 따라 육군에서는 좀 더 효율적인 직무교육 방안의 일환으로 훈련병들에게 '차등제 교육'을 시행하고 있다. 이러한 차등제 교육의 효과를 향상시키기 위해서는 훈련병들의 예상 학업 성취도를 미리 예측하여 성취집단별로 차별화된 교육과정을 거치게 하는 것이 매우 중요하다. 따라서 본 연구에서는 입교 초기에 얻을 수 있는 신병들의 제한된 자료들만을 이용하여 그들의 예상 교육 성취집단을 예측하는 모형을 개발하였다. 본 모형의 목적 변수는 '성취집단'이며 '일반관리 인원' 및 '집중관리 인원'의 두 가지 값을 갖는다. 사용된 기법은 인공신경망(Neural Network) 모형, 의사결정나무(Decision Tree) 모형, SVM 모형, 그리고 Naïve Bayesian 모형 등 4가지 순수 모형과, 각각의 순수 모형을 *k*-means 군집기법과 혼합한 4 가지의 혼합 모형 등 총 8개의 모형의 성능을 비교 분석하였다. 실험 결과 *k*-means 군집기법과 인공신경망 기법을 혼합한 모형이 가장 좋은 예측력을 보이는 것으로 나타났다. 이러한 교육 성취집단 예측 모형은 향후 군에서 이루어지는 다양한 교육 프로그램에 효과적으로 이용될 수 있을 것으로 기대된다.

Abstract

The period of military personnel service will be phased down by 2014 according to 'The law of National Defense Reformation' issued by the Ministry of National Defense. For this reason, the ROK army provides discrimination education to 'newly recruited privates' for more effective individual performance in the on-the-job training. For the training to be more effective, it would be essential to predict the degree of achievements by new privates in the training. Thus, we used data mining techniques to develop a classification model which classifies the new privates into one of two achievements groups, so that different skills of education are applied to each group. The target variable for this model is a binary variable, whose value can be either 'a group of general control' or 'a group of special control'. We developed four pure classification models using Neural Network, Decision Tree, Support Vector Machine and Naïve Bayesian. We also built four hybrid models, each of which combines *k*-means clustering algorithm with one of these four mining technique. Experimental results demonstrated that the highest performance model was the hybrid model of *k*-means and Neural Network. We expect that various military education programs could be supported by these classification models for better educational performance.

Keywords : 데이터마이닝 (Data Mining), *k*-means 군집 기법 (*k*-means clustering), 직무교육 (On-the-job training), 교육성취 (Education achievements)

* 고려대학교 경영학과 석사과정
** 고려대학교 경영학과 교수

1. 서 론

21세기 지식정보사회에 이르러 인적자원개발은 조직전략의 일부분으로서 점차 조직의 핵심적 과제로 떠오르고 있으며, 그 중에서 특히 교육훈련은 조직의 가치창출과 연관되어 조직 목표 달성의 핵심요소로 그 중요성이 더욱 증대되기 시작하였다. 이러한 교육훈련 중에서도 특히 직무에 대한 전문성 향상과 직무몰입을 위해 실시하고 있는 교육 중 하나가 직무교육이다[7]. 육군에서는 이러한 병사 직무교육의 일환으로 신병교육 후, 각 특기 별로 전문적인 지식을 교육하는 '후반기 교육'을 실시하고 있다. 그런데 '국방개혁에 관한 법률'에 따르면 현역병에 대한 복무기간이 2014년까지 단계적으로 단축될 예정이다[1]. 이것은 병사들의 직무 숙련도를 향상 시킬 수 있는 시간이 절대적으로 부족해진다는 것을 의미하는데, 이에 따라 후반기 교육을 통한 직무교육의 중요성이 점차 증대 될 것으로 예상 되고 있다.

현재 후반기 교육을 포함한 신병교육은 '교육훈련 성과 증대 제도'(연합뉴스, 2005.5.30)에 따라 '차등제 교육'으로 실시되고 있다. 차등제 교육이란 훈련병들을 두 집단(일반관리 인원과 집중관리 인원)으로 구분하여 일반관리 인원에게는 원칙에 따라 강도 높은 훈련을 실시하고, 집중관리 인원에게는 1:1 정밀 지도, 정신교육 강화, 그리고 별도 훈련 프로그램 편성 등의 방법을 적용하여, 전반적인 훈련 효과를 증대시키는 것을 말한다. 본 논문에서는 이러한 차등제 교육 방법의 훈련 효과를 향상시키기 위하여 데이터 마이닝 기법을 이용한 교육성취집단 예측모형을 개발하고자 한다. 이러한 예측은 교육실시 이전에 차등교육의 대상자를 일반 관리 인원과 집중 관리 인원으로 미리 분류함으로써 차등제 교육의 효과를 향상 시킬 수 있을 것으로 기대된다.

지금까지의 교육성취도 예측 모형의 개발에는 통계적 기법을 이용한 연구가 주를 이루고 있었으며 데이터마이닝 기법은 그 적용빈도가 매우 낮았다. 특히 국방관련 데이터에 대한 적용은 그 빈도가 더욱 낮았다. 따라서 본 연구에서는 후반기 교육에 입소하는 신병들의 교육 성취집단의 예측을 위하여 기초속성 자료를 사용한 예측모형을 개발하고자 한다. 특히 예측모형의 개발에는 순수 인공신경망 모형, 순수 의사결정나무 모형, 순수 SVM 모형, 그리고 순수 Naïve Bayesian 모형과, KMNN모형[1], KMDT모형[2], KMSVM모형[3] 그리고 KMNB모형[4] 등 네 가지의 혼합모형, 총 8개 모형을 개발하여 이들의 성능을 비교함으로써 가장 성능이 우수한 모형을 찾고자 하였다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 직무교육과 교육성취도에 관련된 기존의 연구를 살펴보았으며 데이터 마이닝 기법들, 특히 k-means 군집기법과 혼합기법들에 대한 문헌 연구를 수행하였다. 3장에서는 본 연구에서 개발한 분류 예측모형의 개발과정에 대하여 기술하였다. 4장에서는 개발된 분류 예측 모형을 이용한 실험과정과 그 결과에 대해서 기술하였고, 마지막으로 5장에서는 본 연구의 결론과 함께 향후 연구방향을 논의하였다.

2. 문헌연구

2.1 교육 성취도와 관련된 기존연구

교육성취도와 관련된 연구로는 다층모형

1) k-means 군집기법과 인공신경망 (Neural Network)의 혼합모형
2) k-means 군집기법과 의사결정나무 (Decision Tree) 기법의 혼합모형
3) k-means 군집기법과 Support Vector Machine 기법의 혼합모형
4) k-means 군집기법과 Naïve Bayesian 기법의 혼합모형

(Multi-level models)을 이용하여 학업 관련 변수들과 교육성취도와의 연관성을 분석하는 연구가 다수 진행되었다[4, 6, 14]. 교육성취도에 데이터마이닝 기법을 적용하려는 연구는 배재호[3]와 김혜숙 등[2]에 의해 수행되었다. 배재호[3]는 고등학생을 대상으로 1학기 성적과 학원 수강 여부, 그리고 수업시간의 학습태도를 이용하여 2학기 성적을 예측하는 의사결정나무모형을 개발하였으며, 김혜숙 등[2]은 방학 중 학습변수를 추가적으로 포함시키고, 의사결정나무 기법과 연관규칙을 함께 사용한 모형을 개발하였다.

이상에서와 같이, 지금까지의 관련 연구들은 주로 아동 및 청소년기 학생들에 대한 교육성취도에 집중되어 있으며 직무교육에 대한 성취도 연구는 그 수가 매우 적었을 뿐만 아니라 사용된 기법들도 다층 모형과 같은 통계적 기법 또는 단순한 데이터 마이닝 기법의 적용에만 국한되어 있다. 따라서 군 복무기간의 단축이라는 제도의 변화로 더욱 절실해진 직무교육의 효율성을 높이기 위한 작업에, 최신 기법으로서 다른 분야에서 좋은 성능을 보이고 있는 데이터마이닝의 다각적인 접근법을 적극적으로 활용해 보는 것이 필요하다고 하겠다. 기존에 통계적 기법에 비해 데이터 마이닝 기법을 적용하여 얻을 수 있는 기대효과들은 다음과 같다. 첫째, 기존의 주로 사용하던 통계적 기법이 다양한 가설을 설정하는 사용자기반의 분석이라면 데이터 마이닝 기법은 주어진 데이터기반에 분석이라 다양한 가설을 설정하는데 따른 제한사항을 극복할 수 있다. 둘째 데이터의 규모가 오늘날과 같이 기하급수적으로 증가하였을 때, 기계학습을 통한 데이터의 예측력을 높이거나 패턴을 분석하는데 있어 빠르게 분석할 수 있다.

2.2 분류모형(classification model) 및 군집분석(clustering analysis)

분류란 클래스가 알려진 많은 개체들에 대한 정보가 주어진 상황에서 새로운 개체가 기존의 어떤 클래스에 속할 것인가를 예측하는 것을 말한다[22]. 이러한 분류 모형의 개발에 관한 연구에서는 대체적으로 인공신경망, 의사결정나무, 그리고 SVM 등의 인공지능 기반 기법과 선형회귀, 그리고 로지스틱(Logistic)회기 등 통계적 기반 기법을 이용하거나 이들의 혼합 기법을 이용하였다[13, 17, 21].

k-means 군집 기법은 비계층적 군집 분석의 대표적인 기법으로, 생성하려고 하는 군집의 개수, 즉 k가 입력데이터와 함께 주어지면 k개의 군집을 생성하게 된다[2, 9, 12]. Hanifi 등[15]은 아날로그 변조신호를 군집화 하는데 있어, k-means 군집기법과 fuzzy C-means, mountain, subtractive 기법 등 3 가지 다른 군집기법을 사용하여 성능비교를 수행하였으며, 이 중 fuzzy C-means 기법이 가장 성능이 좋은 것으로 나타났다.

분류 모형에 관한 초기의 연구에서는 의사결정나무, 인공 신경망, SVM, Rough set theory 등의 다양한 기법 중에서 하나의 기법만 사용하는 방법들이 주로 사용 되었으나, 최근에는 이들 기법과 사례기반 추론, 유전자 알고리즘, 또는 군집 분석 등 서로 다른 기법들과의 혼합 모형을 활용하고 있는 연구가 활발하게 진행되고 있다[10, 11, 20].

Nan-Chen 등[16]은 신용등급을 분류하는데 있어 k-means 군집기법과 인공 신경망 기법을 함께 사용한 혼합기법으로 모형의 성능을 개선시키고자 하였다. Kim 등[18]은 온라인시장에서 개별소비자에 적합한 추천시스템을 개발하기 위하여, k-means 군집기법과 유전자 알고리즘(GA, Genetic Algorithm)[5]을 혼합한

5) 두 부모 유전자로부터 자손 유전자를 생성하고 변이시켜 보다 나은 형질을 가진 유전자를 보존 및 진화시키는 알고리즘

모형을 개발하여 순수 k-means 군집기법, SOM(Self-Organization Map) 군집기법의 성능과 비교하였다. 그 결과 GA k-means 혼합기법이 가장 성능이 좋은 것으로 나타났다. 또한, Kuo 등[19]은 SOM 군집기법과 k-means 군집기법을 함께 사용한 혼합기법의 성능을 순수 SOM 군집기법, two-stage method 군집기법의 성능과 비교하여, 그들이 제안한 혼합기법이 가장 좋은 성능을 내는 것을 보여주었다.

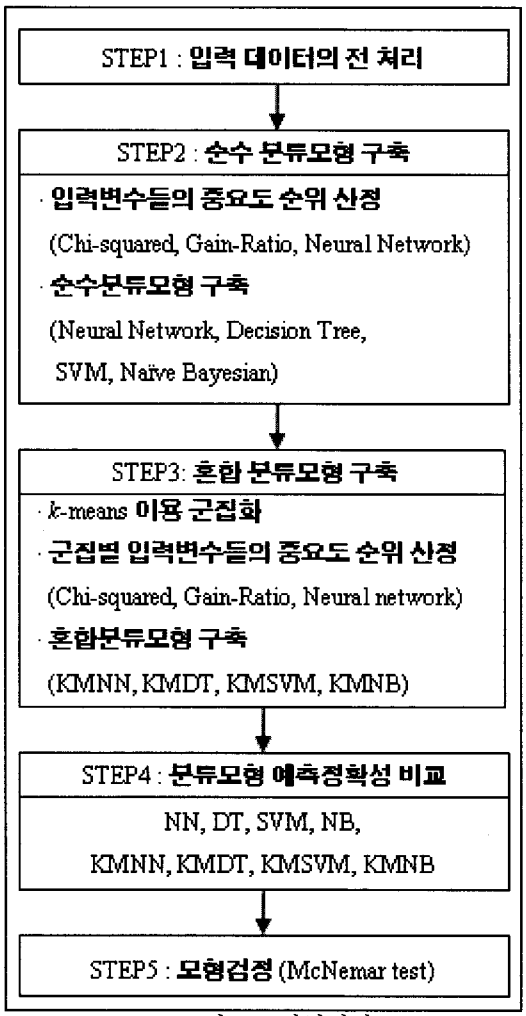

그림 1: 실험절차

3. 데이터 및 실험모형

3.1 실험데이터

실험에 사용된 데이터는 2003년 1년간 육군 훈련소에서 특정 특기에 대한 후반기 교육을 받는 훈련병을 대상으로 수집하였다. 4개 입영 기수에서 총 669명에 해당하는 자료를 수집하였다. 원천 자료의 변수는 총 69개이며 이들은 신체 상태, 지적(知的) 상태, 가정 환경, 그리고 인적(人的) 성향을 나타내는 병사들의 기초속성과 관련된 변수들로 구성되어 있다. 본 연구는 〈그림1〉에서 보인 순서대로 진행하였다.

3.2 실험 절차

3.2.1 입력 데이터의 전 처리 (STEP 1)

원천 변수 중에서 그 값들이 지나치게 하나의 값으로 치중되어 있는 변수들은 실험에서 제외시켰으며, 세 개의 파생변수를 생성하여 모형에 사용하였다. 첫 번째 파생변수는 실질적인 훈련 성취도에 영향을 미칠 것으로 예상되는 '신장'과 '체중'변수를 통합시킨'비만도(BMI)'[6]변수이다. 두 번째와 세 번째 파생변수는 출신환경에 의한 영향요소를 알아보기 위해 '주소'변수에서 추출한 '광역권 주소지'[7]와 도시 및 촌락'[8] 변수이다. 또한 해석의 편의를 위하여 수치형 변수인 지능지수[9]와 나이 변수는 범주형 변수로 변환하였다. 이러한 전처리 과정을 통해 총 15개의 후보 입력변수군이 만들어졌다 (〈표1〉참조).

목표변수의 형태는 처음에 성적을 나타내는 수치형 변수였으나, 4개 기수간의 상대적 차이

6) BodyMassIndex: 체중/(신장)^2 *100, 비만도를 측정하는 지수
7) 현재 주민등록상의 주소지에서 광역권 단위로 추출
8) 도시(행정구역: 구,동) / 촌락(행정구역: 읍,면,리)
9) 최우수(IQ 125 이상, 평균 상위 5%), 우수(IQ 115 이상, 평균 상위 15%), 보통(IQ 115 미만) / http://www.mensakorea.org

표 1: 입력변수의 특성

변수 명	변수 값	통계적 특성
생활정도	상, 중, 하	상:14 (5.19%), 중: 219 (81.1%), 하: 37 (13.7%)
시각결함	D1(안경착용), D2(시각장애(색맹, 색약)), D1D2(안경착용 및 시각장애), N(안경 미착용 및 시각장애가 아닌 자)	D1: 110 (40.74%), D2: 11 (4.07%), D1D2: 21 (7.78%), N: 128 (47.40%)
광역권 주소지	경기, 강원, 충청, 전라, 경상, 제주	경기:145 (53.7%), 강원: 16 (5.93%), 충청: 17 (6.3%), 전라: 39 (14.44%) 경상: 49 (18.15%), 제주: (1.48%)
신체등급	1등급, 2등급, 3등급	1등급: 176 (65.19%), 2등급: 81 (30%), 3등급: 13 (4.81%)
학력	4년제, 2년제, 고졸 이하	4년제: 166(61.48%), 2년제: 73(27.94%), 고졸 이하: 31 (11.48%)
부모	Y(양친), N(편모,편부,고아,계부,계모)	Y: 237 (87.78%), N: 33 (12.22%)
입영구분	징집, 모병	징집: 134 (49.63%), 모병: 136 (50.37%)
신체결함	Y(신체결함이 있음), N(신체결함이 없음)	Y: 22 (8.15%), N: 248 (91.85%)
비만도	저체중, 정상, 과체중	저체중: 23 (8.52%), 정상: 201 (76.67%), 과체중: 46 (17.94%)
혈액형	A형, B형, O형, AB형	A: 99 (36.67%), B: 78 (28.89%), O: 64 (23.7%), AB: 29 (10.74%)
자격면허	Q0, Q1, Q5 (자격증 등급에 따른 분류)	Q0: 218 (80.74%), Q1: 39 (14.44%), Q5: 13 (4.81%)
지능지수	최우수(125 초과), 우수(115~125), 보통(115 미만)	최우수: 63 (23.33%), 우수: 85 (31.48%), 보통: 122 (45.19%)
인성검사	적격, 부적격	적격: 242 (89.63), 부적격: 28 (10.37%)
나이	1(23세 이상), 2(21~22세), 3(20세 이하)	1: 27 (10%), 2: 14 (5.19%), 3: 229 (84.81%)
도시 및 촌락	도시(구, 동), 촌락(읍, 면, 리)	도시:221 (81.85%), 촌락: 49 (18.15%)

를 고려하여 기수 간 평균 점수를 기준으로 정규화 시킨 후, 이를 다시 이진형으로 전환시켰다. 이진형으로 전환할 때에는 육군 군사교육평가 기준[10]에 따라 성적 값의 상위 80%를 '일반관리 인원'으로, 하위 20%를 '집중관리 인원'으로 정의하였다. 이렇게 목표변수를 이분한 이유는 신병교육 및 후반기 교육의 중점은 소수의 정예 직무기술 습득자를 양산하는 것이 아니라, 대상 병사들에게 기본적인 특기 직무기술을 신속하게 습득하게 하는 것에 있기 때문이다. 이를 위해서는 성적 저조 예상자를 미리 예측하여 이들을 집중 관리함으로써 일반 수준으로 향상

10) 상(서열 40%이내),중상(41~80%),중(80% 미만)

시키는 교육이 매우 중요하기 때문에 본 연구에서는 두 집단을 나타내는 이진형의 목표변수를 생성하여 사용하였다.

총 669개의 인스턴스(병사) 중 20%에 해당하는 135건(집중관리인원)과 나머지 80%에서 무작위 추출한 135건(일반관리인원)을 합쳐 목표변수 값의 개수가 동일한 비율을 갖는 총 270건의 인스턴스와 15개 입력변수를 최종 실험에 사용하였다.

3.2.2 순수 분류모형 (STEP 2)

3.2.2.1 입력변수들의 중요도 순위 산정

모형의 성능을 높이면서도 좀 더 경제적인 모형을 만들기 위해서는 목표변수를 예측하는데 좀 더 중요한 역할을 하는 변수만을 선택적으로 사용하는 것이 매우 중요하다. 왜냐하면 목표변수와 관련이 적은 입력 변수가 모형에 투입된 경우에 오히려 모형의 성능을 떨어뜨릴 수 있기 때문이다[12]. 따라서 목표변수와 관련도가 높은 변수를 찾아내고 이들 중 최적 개수만을 입력변수로 사용해야 한다. 또한 본 연구에서 사용된 자료는 입대 초기에 얻을 수 있는 자료를 모두 사용하고 있기 때문에 이들 중에서는 목표변수의 예측과 관련성이 매우 낮은 변수들이 포함되어 있을 것이라 예상되었다. 따라서 후보 입력변수들의 목표변수 예측과의 중요도를 산정하여 목표변수의 예측에 관련성이 높을 것으로 기대되는 변수들을 찾고자 하였다.

입력변수의 중요도 산정 과정은 다음과 같다. 입력변수들의 목표변수와의 관련도를 3가지 기법을 이용하여 순위를 산정하였다. 3가지 기법으로는 카이제곱(Chi-square) 값, Gain Ratio 값, 그리고 신경망 모형의 민감도 분석 결과를 이용하였으며 각 기법들이 산정한 가중치의 평균 순위를 최종 변수의 중요도 순위로 사용하였다. 이들의 결과를 함께 고려하여 순위가 높은 입력변수부터 차례로 하나씩 누적 입력시키며 모형의 성능 변화를 관찰하였다. 이러한 관찰을 통하여 최적의 입력 변수의 수를 선정할 수 있었다[11].

〈표2〉는 입력 변수들의 순위를 보여주고 있는데, 파생 변수 중 하나인 '광역권 주소지' 는 중요도 순위가 3위로서 목표변수에 대한 중요도가 다른 변수들에 비해서 상대적으로 높은 것으로 볼 수 있다. 또한, 개인의 지적 상태(학력, 지능지수, 자격면허) 보다는 신체상태(시각결함, 신체등급, 신체결함)나 가정환경(생활정도, 광역권 주소지, 부모) 요소가 목표변수에 영향을 보다 많이 미치는 것으로 나타났다. 이것은 일반기업에서의 직무교육과는 달리 군(軍)이라는 특수한 상황에서의 교육은 학력, 지능지수 등 개인의 지적 상태 보다는 신체 상태나 가정환경에 따른 개인차가 교육의 성과에 더욱 중요한 요인이 될 수 있다는 가능성을 보여준다.

표2: 중요도 순위 목록

순위	변수 명	순위	변수 명
1	생활정도	9	비만도
2	시각결함	10	혈액형
3	광역권 주소지	11	자격면허
4	신체등급	12	지능지수
5	학력	13	인성검사
6	부모	14	나이
7	입영구분	15	도시 및 촌락
8	신체결함		

3.2.2.2 순수분류모형 구축

위에서 언급한데로 순수분류 모형으로 4 가지의 모형을 구축하였다. 4가지의 순수 모형으로는 인공신경망 모형, 의사결정나무 모형, SVM

11) 변수선정기법을 filter방식으로 사용하여 입력변수들의 중요도순위를 산정한 후 wrapper방식으로 변수를 선정한 것임

모형, 그리고 Naïve Bayesian 모형을 개발하였다.

인공신경망 모형의 경우, Multi-layer Perceptron 알고리즘을 사용하였다. 주요 파라미터 값으로 은닉층(hidden layer)의 노드 수는 8개, 학습률(learning rate)은 0.3, momentum은 0.2, 그리고 훈련횟수는 500회로 설정하였다.

의사결정나무모형의 경우, C4.5 알고리즘을 사용하였다. 주요 파라미터 값으로 가지치기 강도는 0.25, 그리고 자식마디 최소 레코드 수는 2개로 설정하였다.

SVM모형의 경우, polynomial kernel을 사용하였으며, 각각의 파라미터 값으로는, gamma는 0.01, 그리고 c 값은 1.0 으로 설정하였다.

3.2.3 혼합 분류모형 (STEP3)

3.2.3.1 *k-means* 이용 군집화

혼합 분류모형의 흐름은 다음과 같다. 먼저 *k*-means 군집기법을 이용하여 3개의 군집[11]으로 인스턴스(병사)들을 분류해 놓은 후, 새롭게 예측될 병사가 입력되면 해당 병사가 생성된 세 가지 군집 중 어느 군집에 속할지를 예측하게 된다. 해당 병사가 속하게 될 군집이 예측되어 할당되면 해당 군집에서만 다시 학업 성취도를 예측하게 된다. 이렇게 군집화한 목적은 잠재적으로 내제되어 있는 패턴을 기계학습에 의해 비슷한 특성을 가진 병사들로 분류하여 예측모형의 정확도를 향상시키고자 하는데 있다.

이를 위하여 먼저 병사가 속하게 될 군집을 예측하는 모형을 4가지 기법(인공 신경망, 의사결정나무, SVM 그리고 Naïve Bayesian)을 이용하여 개발하고 그 성능을 비교하였다. 개발된 각 모형의 정확도는 98.52%, 97.03%, 97.78%, 그리고 98.15%로 나타났다. 모형의 생성결과 최종적으로 가장 높은 정확도를 기록한 인공신경망 모형(98.52%)을 해당 군집을 예측하는 데 사용하였다.

3.2.3.2 군집별 입력 변수들의 중요도 순위 산정

전술한 바와 같이, 본 연구에서는 *k*-means 군집기법을 이용하여 인스턴스들을 3개 군집으로 분류하여 각 군집별로 분류모형을 구축하였다. 따라서 혼합모형을 만들기에 앞서 각 군집별로 입력변수들의 중요도 순위를 다시 산정하였다. 각 군집에서 입력변수들의 중요도 순위는 각 군집마다 조금씩 차이를 보이고 있다. 이것은 각 군집에 속한 개인의 특성에 따라 조금씩 중요도 순위의 결과가 다를 수 있음을 말해준다. 각 군집에 대한 입력변수들의 중요도 순위는 〈표3〉과 같은데, 군집 3의 경우 전체 데이터로 산정한 결과와 유사한 결과를 보여주고 있다. 이것은 군집 3에 속해 있는 개인들의 특성이 전체 데이터의 특성과 가장 유사했기 때문이라고 추정해 볼 수 있다.

반면, 군집 1의 경우는 전체 데이터나 군집 2, 3에 비하여 가정 환경 (광역권 주소지, 생활정도) 요소가 성취집단을 결정짓는 중요한 변수로 나타나는 것을 볼 수 있다. 또한 군집 2는 다른 군집에 비해 신체 상태 (시각결함, 신체결함) 변수가 성취집단에 중요한 영향을 주는 것을 확인할 수 있다. 이처럼 교육의 결과에 영향을 미치는 요소들이 각 군집별로 차이를 나타내는 것으로 보아 특성이 유사한 집단별로 차별화된 교육 시스템을 적용해야 교육 효과를 극대화시킬 수 있음을 확인시켜 준다.

11) Clementine의 two-step model을 이용한 결과 3개의 군집으로 나누는 것이 가장 적당하다고 판단되었음.

표3: 군집별 변수 중요도 순위 목록

순 위	군집1(94개)	군집2(130개)	군집3(46개)
1	광역권주소지	생활 정도	생활 정도
2	생활 정도	시각 결함	시각 결함
3	학 력	신체 결함	광역권 주소지
4	시각결함	학력	학력
5	신체등급	지능지수	나이
6	자격면허	자격면허	지능지수
7	혈액형	광역권주소지	신체등급
8	비만도	혈액형	비만도
9	지능지수	비만도	자격면허
10	도시 및 촌락	인성검사	혈액형
11	부모	부모	도시 및 촌락
12	인성검사	도시 및 촌락	신체결함
13	신체결함	신체등급	입영구분
14	나이	나이	부모
15	입영구분	입영구분	인성검사

3.2.3.3 혼합분류모형 구축

위와 같이 3개의 군집으로 나누어 주는 분류모형을 먼저 적용 시킨 후, 그 다음에 네 가지 기법을 적용하여 각 군집별로 가장 예측율이 높은 모형을 찾았다. 4 가지 혼합 모형(KMNN, KMDT, KMSVM과 KMNB)은 다음과 같이 개발하였다.

3.2.3.1 절에서와 같이 k-means 군집기법을 이용하여 세 군집으로 군집화한 후, 다시 네 가지의 기법(인공신경망, 의사결정나무, SVM 그리고 NaïveBayesian)을 이용하여 분류 모형을 개발하였다. 전술한 바와 같이 새로운 병사가 예측되기 위하여 시스템에 입력되면 해당 병사가 속하게 될 군집을 먼저 예측한 후, 예측결과 할당 된 군집에서 구축된 분류모형에 입력되어 최종 분류가 이루어지는 과정을 거치게 된다. 최종적으로 생성된 모형의 정확도는 각각의 인스턴스 수[12] 와 군집예측모형의 정확도[13]를 고려하여 산출하였다. 정확도 산출과정을 예를 들어 설명하면, 〈표5〉에 나와 있는 KMNN모형의 정확도는 실험결과, 군집1(C1)의 가장 높은 정확도인 77.66%, 군집2(C2)의 73.08%, 그리고 군집3(C3)의 80.43%에서 각각의 인스턴스 수를 곱한 후 총 인스턴스 수로 나누어 계산하였다 ((77.66% × 94 + 73.08% × 130 + 75.53% × 46) / 270 = 75.93%). 그 후, 군집예측모형의 정확도 98.52%를 곱하여 최종 정확도를 산출하였다(75.93% × 98.52% = 74.80%).

4. 실험결과 및 해석

모형의 개발에는 'Weka V3.4.10'와 'Clementine V10.1'을 사용하였고, 분류자의 정확도를 예측하기 위하여 10-fold cross validation을 실시하였다[21].

순수 모형에 대한 실험 결과는 〈표4〉와 같다. 실험결과는 모형의 성능을 나타내는 백분율로 표시하였다. 〈표4〉에서 볼 수 있는 바와 같이, 인공신경망 모형의 경우는 변수의 수가 2개일 때, 70.37%의 가장 높은 정확도를 보였고, 의사결정나무는 변수의 수가 8개일 때, SVM모형의 경우는 변수의 수가 4개일 때, Naive Bayesian의 경우는 변수의 수가 10개 일 때, 71. 48%의 가장 높은 정확도를 보이고 있다. 이 실험을 통해 비선형(non-linear) 기법들인 인공신경망이나 SVM 모형의 경우에는 다른 모형들에 비해 좀 더 적은 입력 변수를 갖고도 최고 정확도를 보이는 경제적인 모형임을 확인할 수 있었다.

혼합 모형에 대한 실험 결과는 〈표5〉와 같다. 〈표5〉에 따르면, KMNN모형은 군집 1, 군집

12) 군집1: 94개, 군집2 : 130개, 군집3: 46개

13) 인공신경망의 정확도: 98.52%, 의사결정나무의 정확도: 97.03%, SVM의 정확도: 97.78%, NB의 정확도: 98.15%

표4: 순수 모형 분류 예측 정확도 (%)

순위	변수명	인공신경망 (NN)	의사결정나무(DT)	SVM	Naïve Bayesian
1	생활정도	62.22	63.70	63.7	63.7
2	시각결함	70.37 *	70.37	70.37	70.37
3	광역권 주소지	69.63	69.63	69.26	68.89
4	신체등급	70.37	69.26	71.48 *	71.11
5	학력	68.15	68.52	71.48	69.26
6	부모	69.26	68.89	71.48	69.63
7	입영구분	66.67	68.89	71.48	68.89
8	신체결함	66.67	71.48 *	71.48	69.26
9	비만도	67.78	71.48	71.48	68.89
10	혈액형	60.37	71.48	71.48	71.48 *
.		…	…	…	…
15	도시및촌락	64.44	69.63	70	67.04

*: 각 모형의 최고 정확도

표5: 혼합 모형 분류 예측 정확도 (%)

순위	KMNN			KMDT			KMSVM			KMNB		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	57.45	68.46	71.74	56.38	68.46	69.57	57.45	68.46	71.74	57.45	68.46	69.57
2	75.53	70.77	73.91	74.47	68.46	73.91	75.53*	70	69.57	65.96	69.23	69.57
3	77.66*	73.08*	80.43*	75.53*	69.23	82.61*	75.53	70.77*	67.39	70.21	70	76.09*
4	72.34	68.46	80.43	73.4	68.46	82.61	74.47	70	73.91*	73.4*	71.54*	73.91
5	67.02	70	80.43	70.21	72.31	82.61	72.34	70.77	76.09	69.15	70.77	73.91
6	63.83	66.92	76.09	69.15	72.31	80.43	73.4	70.77	69.57	69.15	69.23	65.22
7	65.96	63.98	73.91	71.28	70.77	80.43	70.21	70	65.22	64.89	68.46	67.39
8	63.83	64.62	73.91	71.28	73.08*	80.43	65.96	70.77	67.39	68.09	66.92	63.04
.	…	…	…	…	…	…	…	…	…	…	…	…
15	67	64.6	52.2	74.5	73.08	82.6	61.7	66.9	60.9	62.8	65.4	52.2
	정확도: 74.80			정확도: 74.44			정확도: 72.25			정확도: 71.88		

*: 각 모형의 군집별 최고 정확도

표6: 모형간 정확도 비교 (%)

	KMNN	KMDT	KMSVM	KMNB	NN	DT	SVM	NB
정확도	74.80	73.32	71.70	71.61	70.37	71.48	71.48	71.11

표7: 맥니마 검정 결과

	KMNN	KMDT	KMSVM	KMNB	NN	DT	SVM
NB	***20.55******	***27.77******	***21.41 *****	18.46***	**4.57****	3.86**	1.00
KMNN		2.27	0.11	0.00	***26.68*****	***10.29*****	18.96***
KMDT			1.60	0.93	***34.13*****	***14.75*****	25.83***
KMSVM				0.04	***26.73*****	***10.25*****	19.80***
KMNB					***26.68*****	***9.60*****	17.66***
NN						***15.21*****	***6.23****
DT							2.00

* 유의수준 90%, ** 유의수준 95%, *** 유의수준 99%

2, 그리고 군집 3에서 모두 3가지의 변수를 사용하여 74.80%의 가장 높은 정확도를 보이고 있다. KMDT모형은 군집 1, 군집 3에서 변수의 수가 3개일 때, 그리고 군집 2에서 8개의 변수를 사용하였을 때, 73.32%의 가장 높은 정확도를 보였다. KMSVM모형은 군집 1, 군집 2 그리고 군집 3에서 변수의 수를 각각 2개, 3개, 4개를 사용하였을 때, 72.25%의 가장 높은 정확도를 보였다. 마지막으로 KMNB모형은 군집 1, 군집 2에서 변수의 수가 4개 일 때, 그리고 군집 3에서 3개의 변수를 사용하였을 때, 71.88%의 가장 높은 정확도를 보였다.

정확도 산출 결과, KMDT 모형의 군집 2[14]를 제외한 모든 경우, 4개 이하의 변수만을 사용하였을 때, 가장 높은 결과가 산출된 것을 확인 할 수 있다. 이것은 본 논문에서 제안한 혼합모형이 기존의 순수모형에 비해 비교적 적은 수의 변수만을 가지고도 높은 정확도를 보이는 경제적인 모형을 구축할 수 있는 방법임을 확인시켜 준다고 할 수 있다.

순수 모형 및 혼합 모형 모두에 대한 실험결과는 〈표6〉과 같다. 〈표6〉에서 볼 수 있는 바와 같이, KMNN 모형이 모든 모형과 비교했을 때 74.8%로 가장 좋은 정확도를 보였다. 각 모형별로 가장 좋은 결과를 나타낼 때 사용된 입력변수의 수도 각각 다름을 알 수 있다. 또한 *k*-means 군집기법과 혼합한 모형들이 순수한 모형들보다 정확도가 1~4% 정도 향상되었음을 확인 할 수 있다. 이것은 군집분석의 특성이 예측모형에 반영되어 비슷한 특성을 가진 병사들의 군집을 미리 분류해내 줌으로써 교육성취집단의 예측력을 보다 향상시킨 것으로 판단된다.

마지막으로, 8가지 모형들의 예측 결과 간에 통계적으로 유의한 차이가 있는지를 검정하기 위해 맥니마 검정(McNemar test)을 실시 하였다[8]. 〈표 7〉은 맥니마 검정의 결과를 보여준다. 〈표7〉에서 볼 수 있는 바와 같이, 본 연구에서 제안한 *k*-means 혼합모형의 경우, 순수 모형들인 인공신경망, 의사결정나무, SVM, Naive Bayesian 모형들과 유의수준 99% 또는 95%에서 그 차이가 유의함을 나타내고 있다. 이것을 통해 본 연구에서 제안한 혼합모형의 결과가 기존의 순수모형의 결과와 통계적으로 유의한 차이를 갖는다는 것을 알 수 있다. 하지만, SVM 모형과 NaïveBayesian 모형간, 그리고 SVM 모형과 의사결정나무 모형간에는 유의한 차이가 없는 것으로 나타났다.

14) KMDT 모형의 군집 2에서는 8개의 입력변수에서 가장 높은 결과가 산출되었음.

5. 결론

본 연구에서는 입대 초기에 얻을 수 있는 신병들의 제한된 기초속성 데이터를 이용하여 예상 교육성취집단을 예측하는 모형을 제시하였다. 실험 결과 KMNN 혼합모형이 다른 모형이나 4가지 순수모형보다 예측성능이 좋은 것을 관찰할 수 있었다.

본 연구의 의의는 다음과 같이 몇 가지로 요약할 수 있다. 첫째, 기존 교육성취도 분석에 잘 쓰이지 않는 데이터 마이닝 기법을 적용해 보았다. 둘째, 입대 초기에 확보할 수 있는 신병들의 극히 제한된 기초 속성만을 가지고 75% 수준의 예측율을 얻을 수 있었다. 셋째, 파생변수들을 생성하여 해석의 편의를 높이고 모형의 정확도를 향상 시켰으며 변수들의 중요도 산정을 통해 교육성취도에 중요한 역할을 하는 파생변수를 찾았다. 넷째, 다양한 기법을 이용하여 모형을 개발함으로써 예측력율 향상 시키는 다양한 시도를 경험적으로 수행하였다. 마지막으로 본 연구는 군 조직을 비롯한 다양한 조직체 내에서 직무교육에 대한 성과를 높일 수 있도록 미리 조직원의 성취도 예측을 시도한 것으로, 이러한 예측은 조직원의 맞춤식교육을 확립하는데 있어 중요한 참고자료로 활용될 수 있을 것으로 기대된다.

본 연구의 한계점 및 차후 연구방향은 다음과 같다. 첫째, 다양한 특기에 대한 수집과정상에서의 신병개인정보문제, 군사보안문제 등의 제한사항으로 한 개의 특기 신병만을 수집하여 실험을 하였다. 따라서 좀 더 다양한 특기의 데이터를 모형에 적용시킬 필요가 있다. 둘째, 데이터의 양에 있어서도 보다 많은 데이터를 수정하여 모형의 신뢰도를 향상 시킬 필요가 있다. 마지막으로 본 연구에서는 간부를 대상으로 하는 군사교육평가 기준을 그대로 사용하였으나, 병 교육훈련 실정에 맞게 좀 더 구체적이며 합리적인 기준을 정할 필요가 있다.

참 고 문 헌

[1] 국방부, 국방개혁에 관한 법률 시행령, 2007.

[2] 김혜숙, 문양세, 김진호, 노웅기, "데이터 마이닝을 사용한 방학 중 학습방법과 학업성취도의 관계 분석," 정보과학회논문지: 소프트웨어 및 응용, 제 34권, pp 40-51, 2007.

[3] 배재호, "데이터 마이닝을 이용한 학업성취도 분석,"경희대학교 교육대학원 석사학위논문, 2001.

[4] 오성삼, 구병두, "메타분석을 통한 한국형 학업성위 관련변인의 탐색," 교육학연구, 제 39권, pp. 99-122, 1999.

[5] 육군본부, 육군규정, 2007.

[6] 차지혜, "영어과 학업성취도에 영향을 미치는 배경변수에 대한 다차원적 분석," 이화여자대학교 대학원 석사학위논문, 2001.

[7] 하영자, "공무원의 온라인 직무교육에서 자기효능감과 자기조절학습 수행력이 만족도와 성취도에 미치는 영향", 한국사이버교육학회, e-learning 학술연구, 제4권 제1호, pp. 31~63, 2005.

[8] 허명회, 비교연구를 위한 통계적 방법론. 경기, 자유아카데미, 2005.

[9] Anderberg, R., Cluster analysis for applications, New York, MA: Academic Press, 1973.

[10] Carvalho, D.R., and Freitas, A.A., "A hybrid decision tree/genetic algorithm method for data mining," Information Sciences, Vol. 163, pp 13-35, 2004.

[11] Chang, P.C., Lai, C.-Y., and Lai,

K.R., "A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting," Decision Support Systems, Vol. 42, pp. 1715-1729, 2006.

[12] Dash, M., and Liu, H., "Feature Selection for Classification," Intelligent Data Analysis, Vol. 1, pp. 131-156, 1997.

[13] Delen, D., Glenn W., and Amit K., "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, Vol. 34, pp. 113-127, 2005.

[14] Fuller, B., "What school factors raise achievement in the third word," Review of Educational Research, Vol. 57, pp. 255-273, 1987.

[15] Guldemir, H. and Abdulkadir S., "Comparison of clustering algorithms for analog modulation classification," Expert Systems with Applications, Vol. 30, pp. 642-649, 2006.

[16] Hsieh, N.C., "Hybrid mining approach in the design of credit scoring models," Expert Systems with Applications, Vol. 28, pp. 655-665, 2005.

[17] Hung, S.Y., David, C.Y., and Wang, H.-Y., "Applying data mining to telecom churn management," Expert Systems with Application, Vol.31, pp. 515-524, 2006.

[18] Kim, K.-J., and Ahn, H., "A recommender system using GA K-means clustering in an online shopping market," Forthcoming, 2007.

[19] Kuo, R.J., Ho, L.M., and Hu, C.M., "Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation," Computers & Operation Research, Vol. 29, pp. 1475-1493, 2002.

[20] Min, S.H., Lee, J., and Han, I., "Hybrid genetic algorithms and support vector machines for bankruptcy prediction," Expert Systems with Applications, Vol. 31, pp 652-660, 2006.

[21] Ryu, Y.U., Chandrasekaran, R., and Jacob, V.S., "Breast cancer prediction using the isotonic separation technique," European Journal of Operation Research, Vol. 181, pp. 842-854, 2007.

[22] Witten, I.H., and Frank, E., DATA MINING: Practical Machine Learning Tools and Techniques. San Francisco, MA: Morgan Kaufmann, 2005.

저자소개

곽 기 효 (E-mail: khkwak@korea.ac.kr)

2002 한국외국어대학교 경영정보학과 졸업(학사)
현재 고려대학교 경영학과 석사과정
관심분야 데이터 마이닝, 데이터 웨어하우스, Business intelligence, 군 관련 데이터분석

서 용 무 (E-mail: ymsuh@korea.ac.kr)

1978 서울대학교 수학교육학과 졸업(학사)
1980 한국과학기술원 전산학과 졸업(석사)
1989 미국 University of Texas at Austin 전산학과 졸업(석사)
1992 미국 University of Texas at Austin 경영정보학과 졸업(박사)
현재 고려대학교 경영학과 교수
관심분야 Web-based organizational computing, Ontology, Data warehouse, Data mining