데이터 마이닝을 이용한 신인성검사 판정 연구

- 복무적합도검사를 중심으로 -

박영길*, 인호**, 김능회**, 이정빈**
* 고려대학교 소프트웨어공학과
** 고려대학교 컴퓨터학과
e-mail : 0-gill@hanmail.net

A Study on Assessment of Personality Test using Data Mining

YoungGill Park *, Hoh Peter In**, Nunghoe Kim**, Jungbin Lee**
* Dept. of Software Engineering, Korea University
** Dept. of Computer Science and Engineering , Korea University

요 약

복무적합도 검사는 정신질환이나 사고가능성이 있는 병사를 감별하고, 입대 후 적응문제로 조기 전역할 수 있는 집단을 예측하는 신인성검사 중 하나로, 현재 군에서 징병 및 입영단계에 실시하는 인성검사이다. 이는 전체 검사대상자를 상대로 정신과적 문제 식별을 위한 개별면담이 불가능하기 때문에 위 검사를 통해 대상자를 효율적으로 선별하기 위함이다. 본 연구는 데이터 마이닝을 통해 복무적합도 검사의 판정을 예측 할 수 있을지 확인하고자 하였다. 이를 위해 데이터 마이닝의 기법 중 회귀분석의 로지스틱 회귀분석 기법이 복무적합도검사 판정에 우수한 성능을 보임을 확인하였고, 로지스틱 회귀분석의 추정된 회귀계수를 이용하여 만든 반응확률에 대한 예측 모형식은 높은 정분류율을 보였고 평가 결과 통계적으로 의미가 있음을 증명하였다. 따라서 본 연구 결과를 활용하면 소수의 문항으로 복무적합도 검사 이전의 선별용 검사 개발이나 자가 진단용 검사 개발로 활용이 가능 할 것으로 기대한다.

1. 서론

군에서는 군 복무 부적합자들의 선별을 위해 1995 년부터 사용해 오던 KMPI 를 신뢰도 부족 등으로 인해 2007 년부터 군인성검사(MPI)를 실시 해 왔다. 하지만 군 생활 과정에서 부적응이나 다양한 정신 병리에 따라 발생하는 GP 총기 난사사건부터 해병대 총기사고 등 일련의 사건사고는 군 복무 부적응자의 예측정확률을 높이고 판정결과의 신뢰성 향상과 검사의 실용성과 편의성의 제고를 요구했다. 따라서 군인성검사(MPI)를 기본으로 신인성검사(NEWMPI)가 한국국방연구원(KIDA : Korea Institute for Defense Analysis)에서 개발되었고, 그 중 선병용 검사로 복무적합도검사가 병무청, 육군훈련소, 입소대등에서 2010 년부터 실시되어 왔다[1].

복무적합도 검사는 징병 및 입영단계의 전체 검사대상자를 상대로 정신과적 문제 식별을 위한 1:1 의 개별면담이 불가능 하기 때문에 이 검사를 통해 대상자를 효율적으로 선별하기 위하여 실시한다. 이 검사는 총 183 문항으로 구성되어 응답시간은 20 ~ 50 분 정도로 개인별 차이가 크다. 또한 병무청에서는 인성검사를 위한 전산실이 마련되어 있으나 한번에 실시할 수 있는 인원이 제한되어 있고, 육군훈련소등 입소대에서는 한 기수에 들어오는 인원이 많아 지필로 인성검사를 실시하고 있다. 인성검사를 실시하는데 시간적, 공간적 제약이 있다면 타당한 검사결과를 기대하기 어렵다. 이와 관련하여 응답시간을 줄이기 위해 문항을 축소하여 복무적합도 검사 이전의 검사로서 활용한다면, 인성검사 실시 대상자를 선별 할 수 있어 안정된 환경에서 검사를 실시할 수 있으므로 검사결과의 질적 향상을 꾀 할 수 있다[2].

본 연구는 데이터 마이닝을 통해 복무적합도검사 판정을 예측 할 수 있음을 확인하고, 확인된 마이닝 기법을 이용해 판정에 영향을 미치는 문항을 선별하여 더욱 효율적인 검사환경을 만들기 위한 기초정보를 제공하고자 한다.

2. 연구방법

본 연구는 183 문항으로 구성되어 있는 복무적합도 검사를 데이터 마이닝을 통해 판정을 예측 할 수 있을지를 확인하고자 하였다. 데이터 마이닝의 기법 중 분류와 예측을 하는데 있어서 효과적이며, 적용결과에 의해 규칙을 명확하게 나타낼 수 있는 기법을 사용 할 필요가 있었다. 이를 위해 의사나무결정분석과 회귀분석을 적용 하여 우수한 성능을 보이는 분석법을 선택하고자 하였다. 의사나무결정분석 기법 중에서는 가장 오래되어 널리 사용되고 있고, 범주형 변

수를 사용하는 CHAID(Chi-squared Automatic Interaction Detection) 알고리즘을 선택하였다. 그리고 회귀분석 기법 중에서는 종속변수가 이진형(binary type)이며 여러 개의 입력변수들을 고려한 로지스틱 다중회귀모형(Logistic Regression)을 사용하였다[3].

각 모형을 비교하여 우수한 모형을 선택하고 그 모형에서 가장 큰 정분류율을 나타내는 예측 모형식을 선택하였다. 최종적으로 만들어진 모형식을 평가하고 포함된 설명변수의 각 문항번호를 확인하였다.

본 연구에서 사용한 데이터는 2012 년도 상반기 복무적합도검사 판별식으로 판정된 자료 중에서 복무적합으로 판정 된 8000 건과 복무부적합으로 판정 된 2000 건의 자료로서 183 문항에 대한 “예(1)” 또는 “아니오(2)”의 응답결과 만으로 구성되어 있다. 이 데이터는 각 문항에 대한 응답 중 “무응답(0)”이 포함된 자료와 “사고예측-관심” 또는 “재검사” 등의 판정 자료는 배제하였다. 따라서 실제로 판별되는 복무적합 검사의 판정 예측과 복무부적합 판정 예측을 위해 사용된 데이터 현황은 <표 1>과 같다[2].

<표 1> 복무적합도 판정과 예측에 사용 될 데이터 분류

판정 종류	양호	사고예측		정밀진단	재검사	무응답
		관심	위험			
사용 구분	복무 적합	X	복무 부적합	복무 부적합	X	X

분석을 위해 183 문항의 응답을 183 개의 변수로 구분하였고(v1, v2, …, v183) 각 응답 건 별 판정 결과(assessment)를 “적합”과 “부적합”으로 추가하였다. 자료 처리를 위해서 SAS v9.2 를 사용하였고 데이터 마이닝 분석을 위하여 Enterprise Miner v4.3(이하 E-Miner)을 통해 의사결정나무분석과 로지스틱 회귀분석을 사용하였다.

3. 연구결과

우선 인성검사 판정을 위한 데이터 마이닝 기법 중 우수한 성능을 보이는 기법을 선택하기 위하여 복무적합 판정 데이터 8,000 건과 복무부적합 판정 데이터 2,000 건을 합해 10,000 건의 데이터 테이블을 생성하였고 이를 E-Miner 의 Input Data Source 노드에서 지정하였다.

Data Partition 노드에서는 분석용(Train) 데이터로 70%, 평가용(Validation) 데이터로 30%를 지정하였다. 이와 같은 설정은 데이터 테이블에서 무작위로(Random Seed) 선택된 7,000 건의 데이터로 모형을 생성하여 3,000 건의 데이터로 모형을 검증하기 위함이다.

의사결정나무 분석은 분리 기준으로 카이제곱통계량(Chi-square statistics 유의수준 0.20)과 엔트로피 지수(Entropy index) 및 지니 지수(Gini index)를 사용해 비교하기 위하여 세가지로 구성하였다. 로지스틱 회귀분석은 변수선택 방법으로 단계적 선택법(Stepwise)을 선택하였고 모형선택의 기준으로는 Validation Error(오차율이 가장 작은 모형선택 기준)를 선택하였다.

모형의 성능을 비교하기 위하여 네 가지 모형의 리프트 도표(Lift Chart)를 구성한 결과 의사결정나무 모형의 분리기준을 달리 한 세가지 모형보다 로지스틱 회귀모형의 성능이 우수하다는 것을 확인 할 수 있었다.

복무적합도 검사의 판정예측에 적합 한 기법인 로지스틱 회귀모형의 결과로 Output 탭을 확인하면 [그림 1]과 같은 결과를 볼 수 있다.

The DMREG Procedure

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
v58	1	46.5896	<.0001
v64	1	44.8533	<.0001
v78	1	32.2510	<.0001
v84	1	34.3632	<.0001
v9	1	38.3699	<.0001
v93	1	9.2404	0.0024

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-18.7308	1.9869	88.87	<.0001	0.000
v107	1	-1.2642	0.2571	24.17	<.0001	0.282
v118	1	-1.7015	0.3193	28.40	<.0001	0.182
v119	1	-2.3751	0.4348	29.84	<.0001	0.093
v120	1	2.4735	0.3726	44.08	<.0001	11.864
v124	1	-1.9327	0.2858	45.73	<.0001	0.145
v135	1	-0.5030	0.2187	5.29	0.0214	0.605
v17	1	-1.9324	0.3315	33.99	<.0001	0.145
v21	1	-2.4226	0.3984	36.97	<.0001	0.089
v31	1	-1.9267	0.2975	41.95	<.0001	0.146
v33	1	-2.5590	0.4073	39.47	<.0001	0.077
v36	1	-2.7181	0.4654	34.11	<.0001	0.066
v45	1	-2.0807	0.3685	31.88	<.0001	0.125
v58	1	-1.7540	0.2570	46.59	<.0001	0.173
v64	1	-2.5401	0.3793	44.85	<.0001	0.079
v78	1	-1.5011	0.2643	32.25	<.0001	0.223
v84	1	-2.1866	0.3730	34.36	<.0001	0.112
v9	1	-1.8650	0.3011	38.37	<.0001	0.155
v93	1	-0.8670	0.2852	9.24	0.0024	0.420

(그림 1) 로지스틱 회귀모형의 결과

[그림 1]에서 알 수 있듯이 인성검사 판정에 대해 58 번, 64 번, 78 번, 84 번, 9 번, 93 번 문항이 유의한 것으로 나타났지만 107 번, 118 번, 119 번, 120 번, 124 번, 135 번, 17 번, 21 번, 31 번, 33 번, 36 번, 45 번 문항은 통계적으로 유의하지 않지만 판정 예측을 설명하기에 타당한 변수이기 때문에 변수로 채택되었다.

따라서 추정된 회귀계수를 이용하여 반응확률에 대한 예측 모형식을 다음과 같이 만들 수 있다

$$Logit(\hat{p}) = -18.7308 - 1.2642*v107 - 1.7015*v118 - 2.3751*v119 + 2.4735*v120 - 1.9327*v124 - 0.5030*v135 - 1.9324*v17 - 2.4226*v21 - 1.9267*v31 - 2.5590*v33 - 2.7181*v36 - 2.0807*v45 - 1.7540*v58 - 2.5401*v64 - 1.5011*v78 - 2.1866*v84 - 1.8650*v9 - 0.8670*v93$$

.

또한 [그림 2]와 같이 오분류표(정오분류행렬)를 확인 할 수 있다. [그림 2]에서 전체 관찰치중 오 분류가 차지하는 비율을 계산해 보면 (13+20) / 7000 = 0.005 이다. 따라서 제대로 분류된 비율은 1 – 0.005 = 0.995 이므로 정 분류율은 99.5%라고 할 수 있다.

FREQ 프로시저

테이블 : F_assessment * I_assessment

F_assessment(From: assessment)
I_assessment(Into: assessment)

빈도 / 백분율 / 행 백분율 / 칼럼 백분율	부적합	적합	총합
부적합	1395	13	1408
	19.93	0.19	20.11
	99.08	0.92	
	98.59	0.23	
적합	20	5572	5592
	0.29	79.60	79.89
	0.36	99.64	
	1.41	99.77	
총합	1415	5585	7000
	20.21	79.79	100.00

(그림 2) 회귀모형 결과의 정오분류행렬(Confusion Matrix)

다음은 로지스틱 회귀분석을 통해 얻은 모형을 평가하기 위하여 Threshold-based 도표를 이용해 각 분류기준값 별로 사후확률에 근거한 오분류행렬을 확인하였다. 대화식 이익도표(Interactive Profit Chart)를 이용하여 <표 2>와 같이 직접 가중값을 주었다.

<표 2> 이익행렬(Profit Matrix)

		분류범주 (Predicted Target)	
		적합	부적합
실제범주 (Actual Target)	적합	1	0
	부적합	0	20

<표 2>와 같이 부적합을 부적합으로 판정했을 때의 가중값을 더 높이 주는 것은 이 논문의 목적이 기존의 복무적합도 검사를 대체하는 것이 아니라 복무적합도 검사 대상자를 선별하기 위함이다. 그러므로 부적합자를 적합으로 예측하는 수치를 최소화 하고자 하였다. 이 이익행렬을 적용하여 분류기준값(Threshold)에 따라 이익을 나타내는 그래프를 확인하였고 분류기준값이 95% 일 때 부적합자를 적합으로 판정하는 수치가 제일 낮았다. 분류기준값이 95%일 때의 정오분류행렬은 [그림 3]와 같다.

검증 데이터셋의 정오분류행렬인 [그림 3]에서 오분류율은 (4+34) / 3000 = 0.013 이다. 따라서 정분류율은 98.7%가 된다. 회귀모형에 비해 정분류율은 0.8%의 낮아 졌지만, 부적합을 부적합으로 판정하는 백분율은 99.32%로 모형에 비해 0.24%가 높음을 확인할 수 있다. 또한 민감도(sensitivity)는 (실제적합, 예측적합)인 관찰치의 빈도 / 실제적합인 관찰치의 빈도로 계산 되는데 2374 / 2408 = 0.986 이다. 그리고 특이도(specificity)는 (실제부적합, 예측부적합)인 관찰치의 빈도 / 실제부적합인 관찰치의 빈도로 계산 되고 588 / 592 = 0.993 이다. 이는 각각의 범주에 대한 분류 능력이 상당히 높음을 보여주는 수치이다.

---------- thresh=95 ----------

FREQ 프로시저

테이블 : actual * predict

actual predict

빈도 / 백분율 / 행 백분율 / 칼럼 백분율	부적합	적합	총합
부적합	588	4	592
	19.60	0.13	19.73
	99.32	0.68	
	94.53	0.17	
적합	34	2374	2408
	1.13	79.13	80.27
	1.41	98.59	
	5.47	99.83	
총합	622	2378	3000
	20.73	79.27	100.00

Model Name: StepReg 147
Target = ASSESSMENT

(그림 3) 검증데이터의 정오분류행렬

이상의 결과로 로지스틱 다중회귀분석의 결과로 107 번, 118 번, 119 번, 120 번, 124 번, 135 번, 17 번, 21 번, 31 번, 33 번, 36 번, 45 번, 58 번, 64 번, 78 번, 84 번, 9 번, 93 번의 총 18 개 문항이 결과 예측에 타당한 변수로 확인되었다. 이는 복무적합도 183 문항 중 약 1/10 에 해당하는 문항으로 검사 응답시간을 크게 줄일 수 있는 문항 개수다. 또한 회귀분석의 결과로 얻어낸 모형은 [그림 3]과 같이 분류기준값을 95%일 때 정분류율 98.7%로, 복무적합도 검사판정과 매우 유사한 예측 치를 보였다. 또한 부적합자를 부적합자로 예측하여 복무적합도 검사 대상자로 선별하는 백분율도 99.32%로, 이는 복무적합도 검사 대상자를 선별하는데 매우 유의한 수치임을 알 수 있다.

복무적합도검사는 크게 반응왜곡 척도, 임상 척도, 사고 관련 척도로 구성되어있다. 반응왜곡 척도는 검사결과의 타당도를 결정하는 척도이며, 임상 척도와 사고 관련 척도는 검사자의 부적합 판정을 결정하는 중요 척도가 된다. 임상 척도와 사고 관련 척도의 구성은 <표 3>에서 확인 할 수 있고, 각 척도 별 문항의 합계가 복무적합도 183 문항과 선별 된 18 문항 보다 큰 것은 각 척도 별로 구분되는 문항이 중복되기 때문이다.

<표 3> 복무적합도 검사의 척도 구성 및 결과

분류		하위척도명	구성 문항수	선별 문항수	비율 (%)
임상척도	정신병 척도	정신분열 척도	17	3	17.6
		편집증 척도	10	2	20
	신경증 척도	불안 척도	13	4	30.8
		우울 척도	32	10	31.3
		신체화 척도	17	2	11.8
		성격장애 척도	25	3	12
사고관련 척도		군탈 척도	13	2	15.4
		적응문제 척도	23	2	8.7
		행동지체 척도	19	3	15.8
		행동화 척도	24	3	12.5
합계			193	34	17.6

본 연구의 결과로 확인 된 18 개의 문항은 임상척도의 6 개 척도와 사고관련척도의 4 가지 척도에 각각 적게는 2 문항 많게는 10 문항에 해당한다. 선별된 문항은 한 척도에만 의존하지 않고 고르게 분포되어 있음을 보여주고 있으며, 이는 선별 된 문항이 기존 검사에서 선별 하고자 하는 부적합의 요인을 충분히 반영하고 있음을 보여준다.

또한 현대인의 정신건강에 큰 영향을 미치는 요인으로 불안과 우울을 꼽을 수 있는데, 연구 결과로 선별된 문항이 불안 척도와 우울 척도에서 높은 반영비율을 구성하는 것은 이러한 양상을 잘 반영하는 것으로 볼 수 있다.

4. 결론 및 향후 과제

본 연구의 목적은 183 문항으로 이루어진 복무적합도 검사의 판정이 어떠한 데이터 마이닝 기법을 사용했을 때 우수한 기법임을 확인하는 것이며, 다음으로는 선택된 데이터 마이닝 기법으로 복무적합도 판정모형을 생성하여 인성검사 판정을 예측하였을 때 이를 가장 잘 설명해주는 문항들을 도출하고자 하는 것이다. 따라서 본 연구는 복무적합도 검사의 문항 축소 가능성을 알아보고자 한 연구라고 할 수 있다.

먼저, 종속변수가 이진형이며 여러 개의 입력변수들을 고려할 수 있고 설명가능 한 모형식을 도출해 낼 수 있는 데이터 마이닝 기법 중 로지스틱 다중회귀분석이 복무적합도 판정예측에 가장 우수한 기법임을 확인 하였다. 그리고 로지스틱 다중회귀분석을 통해 얻은 18 개의 문항 구성으로 신뢰성 있는 모형을 구성할 수 있음을 확인 했다. 이는 복무적합도 검사에 비해 문항이 적어 판별 과정이 너무 단순하다는 단점이 있을 수 있지만 이것은 복무적합도 183 문항에 비해 간단하고 경제성이 있다는 장점이 될 수 있다.

본 연구에서는 복무적합도 검사의 판정과 크게 일치하는 모형식을 도출하였지만 이를 이용해 실제 병무청에서 부적합 판정을 받은 인원들과 조기 전역한 병사들의 데이터를 판별하여 비교 해 보지 못하였다. 추후 연구에서는 실제 복무부적합자(병무청에서 부적합 판정을 받은 집단, 조기전역자, 사고자들의 인성검사 데이터)와 복무적합자(만기제대자의 인성검사 데이터)를 데이터 마이닝 기법을 이용 한 판정을 연구할 필요가 있다.

참고문헌

[1] 최광현 외 (2007). “새로운 군 인성검사 개발”. 한국국방연구원 연구보고서

[2] 최광현 외 (2011). “2011 년 인성검사 및 병영생활 적응연구”. 한국국방연구원 연구보고서

[3] 강현철, 한상태, 최종후, 김은석, 김미경(2001). “SAS Enterprise Miner 4.0 을 이용한 데이터마이닝 - 방법론 및 활용”. 자유아카데미

[4] 이진호 (2006). “로지스틱 회귀분석을 이용한 장교 진급률 및 임의전역률 추정에 관한 연구”. 고려대학교 학위논문

[5] Bing Liu (2006). “Web Data Mining”. Springer Verlag

[6] Hosmer, D. W. and Lemeshow, S. (2000) “Applied Logistic Regression, Second Edition”. Wiley, New York

[7] SAS Manual.” SAS Enterprise Miner 4.3 Reference.”