Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.1 pp. 11~21

Neural Networks Analysis of Transferring Students[1)]

Tae Yoon Kim[2)] · Ji Young Lee[3)] · Kyu Moon Song[4)]

Abstract

In 1997 a new educational system that allows student to tranfer between and within universities was first introduced. As a result, most colleges of basic arts and sciences face a serious problem since quite a few students there have transferred or seems to want to transfer. In this paper we study a problem of building a forecasting neural network for students who can possibly transfer.

Keywords : 신경망, 예측모형, 전공이탈

1. 서론

1997년 이후 수요자 중심의 교육제도 도입은 대학내(혹은 대학간) 전공 사이에서 학생 이동을 보편화 시켰다. 즉 전공선택의 자유 확대는 한번 선택한 전공도 대학간 편입학 혹은 대학내 전부/전과 제도를 통해 쉽게 다른 전공으로 바꾸는 것을 가능하게 하였다. 이러한 제도 변화에 따라 대부분의 대학들은 이탈 가능한 학생들을 사전 파악하여 지도하는 문제에 많은 관심을 갖게 되었다. 박철용과 송규문(2002)은 이러한 문제에 대한 통계적 접근에 관심을 갖고 이탈 가능한 학생들의 예측을 위한 통계적 모형을 구축하고자 하였다. 그들은 의사결정나무를 사용하여 지방소재 특정 사립대학내 전부/전과 데이터를 분석하여 그 결과를 토대로 이탈 가능성을 판단할 수 있는 주요 변수들을 제시하였다. 본 연구는 이들의 연구 결과를 토대로 i) 그들이 제시한 주요변수들 이외에 입력변수로 사용될 수 있는 변수변환들을 제시하고 ii) 훈련용 데이터의 다양한 구성을 통해 실제 사용 가능한 이탈학생의 예측모형을 개발하고자 한다. 이를 위해 의사결정나무 대신 신경망을 사용하는데 본 연구에서 인공 신경망을 이용하여 모형을 구축하는 데는 몇 가지 이유가 있다. 신경망의 장점은 데이터에 대한 특별한 가정없이 복잡한 데이터에 적합능력이 뛰어나고 입력변수와 출력변수에 범주형

1) The present research has been conducted by the attached research institute Research Grant of Keimyung University in 2001.

2) Professor, Department of Statistics, Keimyung University, Taegu, 704-701, Korea
E-Mail: tykim@kmu.ac.kr

3) Department of Statistics, Keimyung University, Taegu, 704-701, Korea

4) Professor, Department of Statistics, Keimyung University, Taegu, 704-701, Korea

변수나 양적 자료를 사용할 수 있다는 점 등인데 이러한 장점들은 기존 자료에 대한 여러 가지 변수 변환 및 훈련자료의 다양한 구성을 시도하고자 하는 본 연구의 주된 목적에 잘 부합되는 것으로 판단되었기 때문이다 (Azoff (1994) 참조). 본 연구에서 구축되는 모형이 최적의 모형이라 할 수는 없으나 박철용과 송규문(2002)에 의해 제시된 신상자료와 학적자료에 근거한 단순한 의사결정나무보다는 좀 더 의미있고 실제 사용가능 한 모형인 것으로 판단된다.

본 연구의 구성은 다음과 같다. 2절에서는 주어진 자료의 설명 및 변수 변환의 필요성에 대해 설명하고 3절에서는 예측을 목적으로 한 다양한 훈련 데이터의 구성문제에 대해 논의한 후 전공이탈 학생에 대한 신경망 예측모형을 설정하게 된다. 4절에서는 분석과 예측모형의 결과를 요약하고 최종적인 결론을 내린다.

2. 분석대상 자료 및 입력변수

이 연구에서 분석에 사용하는 자료는 기본적으로 박철용과 송규문(2002)에서 사용된 자료이다. 그들은 대구시내 한 대학교의 기초학문분야인 인문학부, 자연과학부에 1995년부터 2002년까지 입학한 학생들의 신상자료와 학적자료를 이용하였다. 인문학부의 4개 전공과 자연과학부의 9개 전공으로 구성되는 전체 데이터는 학적상태에 따라 재학, 전과, 졸업, 제적으로 구성되며 이 중 전과 및 제적을 (전공)이탈로, 졸업을 비이탈로 정의하되 재학은 졸업 때까지 이탈여부가 확인되지 않은 검열자료(censored data)이기 때문에 분석에서 제외하였다. 최종적으로 분석에 포함된 학생은 모두 2081명이며, 그 중 56.6%인 1177명이 이탈했으며 나머지 43.4%인 904명이 이탈하지 않았다.

표 1에 박철용과 송규문(2002)의 분석에 포함된 변수를 요약하였다. 구체적으로 살펴보면 설명변수로서 고려된 것은 범주형 변수로 전공, 성별, 주야구분, 입학년도, 교직이수여부, 출신지역이 있으며 양적 변수로 1학년 1학기부터 2학년 2학기까지 각 학기의 성적과 장학금수혜율등이다.

표 1. 2002년 연구의 입력변수

변 수	변수내용과 코딩
전 공	인문학부: 인문학부와 4개 전공 [1, 2, 3, 4, 5] 자연과학부: 자연과학부와 9개 전공 [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
성 별	남자 = 1, 여자 = 2
주야구분	주간 = 1, 야간 = 2
입학년도	학부제이전 = 1, 학부제이후 = 2
교직이수여부	이수 = 1, 미이수 = 0
출신지역	세 개의 지역: 1, 2, 3
성 적	1학년 1학기부터 2학년 2학기까지의 성적
장학금수혜율	1학년 1학기부터 2학년 2학기까지의 장학금/등록금*100%
이탈여부	이탈 = 1, 비이탈 = 0

이 중 입학년도는 전공 이탈의 분수령이 되는 것으로 보이는 1996년을 기준으로 1996년까지의 학부제 이전과 1997년 이후의 학부제 이후로 구분하였으며, 또한 이탈한 학생들의 성적과 장학금수혜율은 이탈이후 결측값으로 존재하여 그냥 놔두면 반응변수인 이탈여부에 직접 관련된 변수가 되기 때문에 결측값을 이탈 이전의 평균값으로 대체하여 분석에 포함시키도록 하였다. 본 연구에서는 박철용과 송규문(2002)에 의해 고려된 변수들 모두를 포함한 상태에서 다음과 같은 새로운 변수들을 추가하였다. 즉 다음과 같은 성적들의 변수 변환을 고려하였다.

성적 범위 : {max(성적11, ... , 성적22) - min(성적11, ... , 성적22)}
성적 평균값 : (성적11+... + 성적 22)/4
성적 변화율 : (성적12/성적11, 성적21/성적12, ... , 성적22/성적21)

여기서 성적ij 는 i학년 j학기 성적을 뜻한다.

표 2. 추가변수가 포함된 입력변수

범 주	변수이름	변수내용
전 공	ma	인문학부: 인문학부와 4개 전공 [1, 2, 3, 4, 5] 자연과학부: 자연과학부와 9개 전공 [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
성 적	x1	1-1학기 성적
	x2	1-2학기 성적
	x3	2-1학기 성적
	x4	2-2학기 성적
	x5	성적의 범위값
	x6	1-1학기 성적/1-2학기 성적(x1/x2)
	x7	1-2학기 성적/2-1학기 성적(x2/x3)
	x8	2-1학기 성적/2-2학기 성적(x3/x4)
	x9	성적의 평균값
장학금수혜율	x10	1-1학기 장학금수혜율
	x11	1-2학기 장학금수혜율
	x12	2-1학기 장학금수혜율
	x13	2-2학기 장학금수혜율
성 별	se	남자 , 여자
입학년도	en	학부제 이전 , 학부제 이후
주야구분	dn	주간 , 야간
교직이수여부	tm	이수 , 미이수
출신지역	re	세 개의 지역 = 1, 2, 3
목표변수	ch	이탈여부 (이탈=1, 비이탈=0)

이들 새로운 변환 변수들을 포함한 이유는 원하는 새로운 전공으로 이탈을 할 수 있느냐의 여부는 많은 경우 기존 전공에서의 성적에 의해 영향을 받는다고 알려져 있으며 그에 따라 전공 이탈을 준비하는 학생들의 기존 전공내에서 성적 분포 혹은 성적 변화가 전공 이탈을 준비하지 않는 학생들과 구별되는 특성을 보일 것으로 기대되기 때문이다.

반응변수에 대한 설명력을 알아보는 기초분석으로서 범주형 변수에 대해서는 카이제곱 검정을, 양적 변수에 대해서는 2표본 t-검정을 시행해 보았다. 그 결과 범주형 변수 중 출신지역의 유의확률이 0.5333으로 상당히 크게 나타났으며 그 외의 모든 설

명변수는 유의확률이 0.001보다 작은 것으로 나타났다 (표 3,4 참조). 따라서 신경망을 이용한 분석에서 출신지역이 이탈여부에 대한 설명력이 거의 없을 것이기 때문에 제외하는 것이 타당한 것으로 보이나 박철용과 송규문(2002)과의 직접 비교를 위해 이 변수들을 그냥 분석에 남겨두도록 한다.

표 3. 양적 변수에 대한 2표본 t-검정

변 수	t-value	p-value
x1	188.83	<.0001
x2	195.84	<.0001
x3	196.62	<.0001
x4	209.68	<.0001
x5	73.38	<.0001
x6	227.05	<.0001
x7	206.06	<.0001
x8	230.55	<.0001
x9	210.22	<.0001
x10	6.56	<.0001
x11	18.63	<.0001
x12	18.95	<.0001
x13	16.83	<.0001

표 4. 범주형 변수에 대한 χ^2-검정

변 수	χ^2-value	p-value
ma	196.7618	<.0001
se	211.2746	<.0001
en	472.5973	<.0001
dn	33.1650	<.0001
tm	102.6461	<.0001
re	1.2575	0.5333

앞에서 설명한 신상자료와 학적자료에 근거하여 이탈학생을 설명하는 신경망 분석을 시도하였다. 신경망 모형을 구축하는데는 여러 가지 모형이 있지만 가장 널리 사용되는 모형은 다층 퍼셉트론 (multilayer perceptron, MLP) 신경망이며 이는 입력층

(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성된 전방향 신경망 모형이다. MLP 신경망 구조는 그림 1과 같다.

신경망 모형실행에서 가장 널리 알려진 알고리즘은 1960년대초 Rumenlhart와 McClelland에 의해 개발된 역전파 알고리즘인데 이는 MLP 신경망을 학습시키는 알고리즘으로써 실제 알려진 데이터의 값과 신경망 모형을 통해 예측된 값을 비교하면서 훈련된 데이터를 반복 처리 학습해 나간다 (Haykins(1999) 참조). 본 연구에서 사용한 신경망 모형은 역전파 알고리즘을 사용한 MLP 신경망이다.

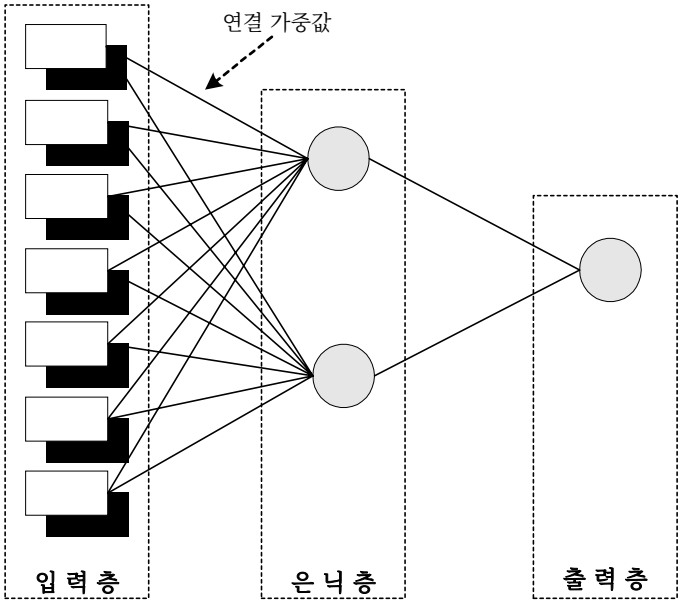

그림 1. MLP 신경망 구조

신경망의 적절한 구조를 찾기 위해 1997년 데이터를 훈련용 데이터로 사용하여 여러 개의 신경망 구조들을 시도해 본 결과 은닉층 1개와 은닉노드 3개로 구성된 신경망 모형이 RMS (root mean square) 오류가 0.317로 가장 작았다. 따라서 입력층은 19개의 입력노드로 구성되어 있고, 은닉층은 1개의 은닉층내에 3개의 노드로 구성되어 있는 신경망 구조를 선택하였다. 출력층은 이탈여부를 판단하는 것이 목적이므로 1개로 구성한다. 신경망의 학습방법은 시그모이드 비선형 활성함수를 이용한 역전파 알고리즘을 사용하였다. 이 분석에서 사용된 신경망 구조를 살펴보면 그림 2와 같다.

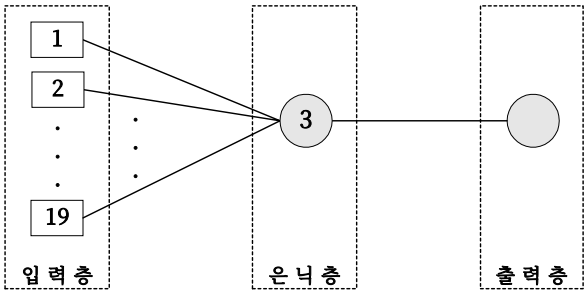

그림 2. 신경망 구조

1997년 데이터를 훈련 데이터로 사용한 주된 이유는 1997년에 처음으로 학부제가 도입되어 전과 및 전부가 실제로 시작된 년도이기 때문이다. 1997년 데이터를 훈련용 데이터로 훈련을 시킨 결과 훈련된 신경망의 오분류율은 12.1%였다. 이러한 오분류율은 박철용과 송규문(2002)의 입력변수들만을 사용한 경우의 오분류율 14.6%보다 작은 것으로써 새로운 변수들을 도입한 신경망 훈련이 어느 정도 의미있는 작업이었음을 뜻한다. 참고로 1997년 데이터는 총 404명으로써 그 중 전공을 이탈한 학생이 141명(34.9%), 비이탈 학생이 263명(65.1%)이다. 1997년에 대해 새로운 입력변수들을 사용했을 때의 구체적인 오분류행렬은 표 5에 주어져 있다.

표 5. 오분류행렬

예측 / 실제	0	1	총합
0	263	0	263
1	49	92	141
총합	312	92	404

신경망 분석의 결과를 의사결정나무 모형을 통해 살펴본 결과 박철용과 송규문(2002)에 의해 밝혀진 대로 성별이 가장 중요한 요인으로 드러났으며 새로 도입된 성적분포 및 성적변화를 나타내는 변수들 중 특히 2학년 1학기에서 2학년 2학기 성적변화가 이탈 학생의 판단에 많은 정보를 주는 것으로 판단되는 데 특히 이 기간의 성적이 많이 개선되지 않은 학생들의 경우 기존의 전공을 떠나는 경향이 뚜렷함을 알 수 있었다 (그림 3 참조).

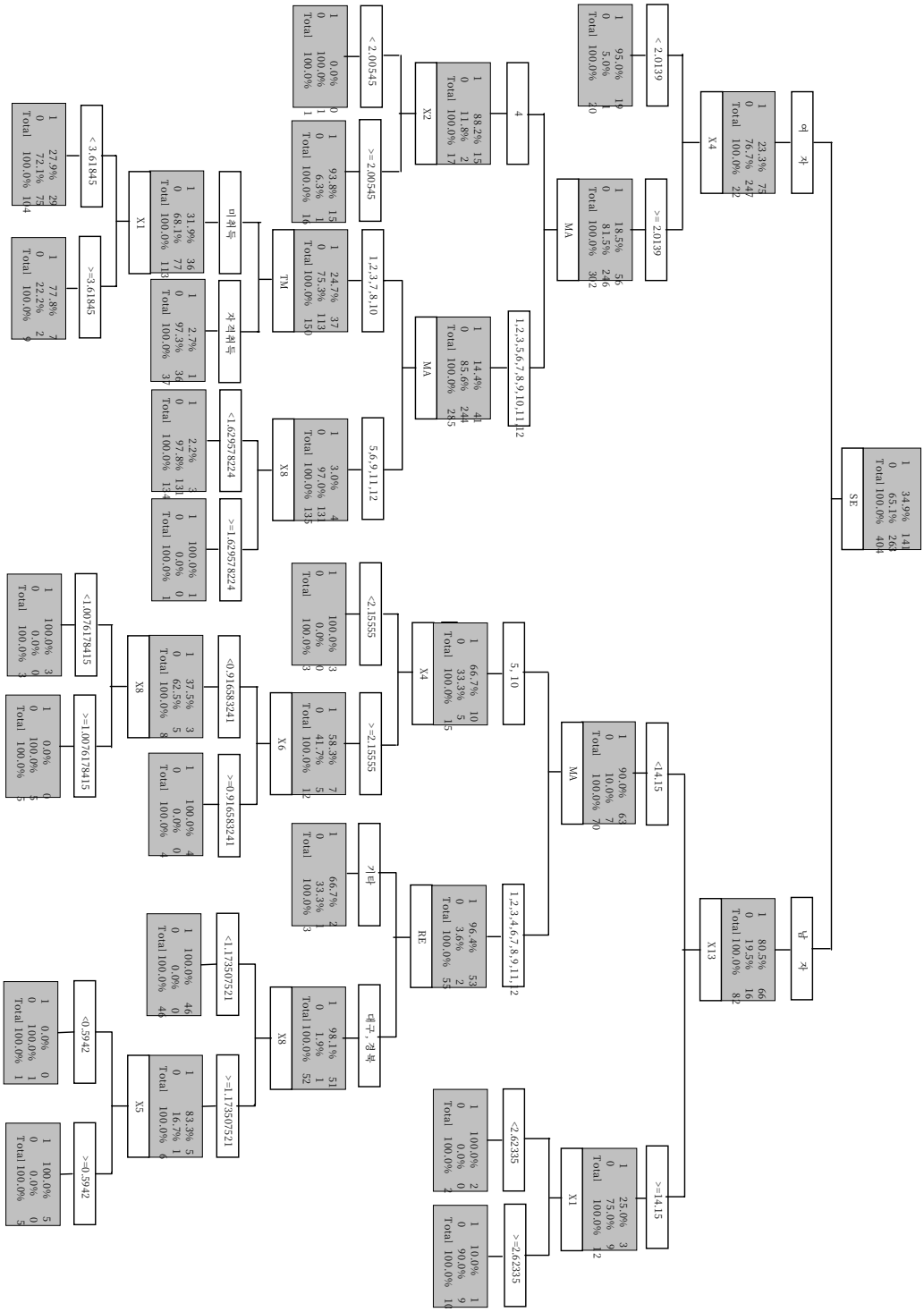

그림 3. 의사결정나무 모형

3. 훈련 데이터 구성과 예측

다가오는 년도의 이탈가능 학생들을 예측하기 위한 모형 구축의 첫 단계는 적절한 훈련 데이터를 통한 신경망 구축이다. 예를 들어 현재 2002년에서 2003년을 예측하고자 할 경우 훈련에 사용될 수 있는 데이터는 과거로부터 2002년까지의 모든 데이터이지만 그 모든 데이터가 유용하지 않을 수 있다. 즉 2002년 데이터는 많은 도움이 되는 반면 과거 5년전인 1997년의 데이터는 별로 도움이 되지 않을 수도 있다. 이는 시간의 흐름에 따라 전공이탈 학생들의 행동 반응 양태가 먼 과거와는 달라 질 수 있기 때문이다. 따라서 과거 데이터중 어느 데이터를 훈련 데이터로 사용할 것인가는 본 연구의 중요한 문제이며 본 연구에서는 이를 위해 훈련 데이터의 적절한 구성을 찾기 위해 세가지 유형의 훈련 데이터를 실험해 보았다. (A) 1997년을 훈련 데이터로 사용한다. (B) 1995년 이후서부터 바로 예측 대상 전년도까지의 모든 데이터를 훈련 데이터로 사용한다. (C) 예측 대상 바로 전년도 데이터만을 훈련 데이터로 사용한다.

각 구성기법을 사용하여 훈련 및 예측시 발생 오분류율을 살펴보면 아래와 같다. (단 여기서 사용된 신경망 구조는 비교 목적상 앞의 그림 2에 주어진 구조로 일률적으로 고정되었다.) 구성 A를 사용한 경우의 훈련 오분류율은 12%이며 예측 오분류율은 95년 40%, 96년 39%, 98년 50%, 99년 82%이다. 구성 B를 사용한 경우의 훈련 및 예측 오분류율은 96년 13%, 41%, 97년 20%, 31%, 98년 24%, 48%, 99년 27%, 78% 등이었으며 구성 C를 사용한 경우의 오분류율은 96년 13%, 41%, 97년 22%, 35%, 98년 12%, 50%, 99년 20%, 59% 등이었다 (표 6 참조).

표 6. 각 구성기법을 사용한 오분류율

구성기법	년도	훈련 오분류율	예측 오분류율
구성 A	95년	.	40%
	96년	.	39%
	97년	12%	.
	98년	.	50%
	99년	.	82%
구성 B	95년	.	.
	96년	13%	41%
	97년	20%	31%
	98년	24%	48%
	99년	27%	78%
구성 C	95년	.	.
	96년	13%	41%
	97년	22%	35%
	98년	12%	50%
	99년	20%	59%

여기서 년도들은 예측대상 연도들을 나타낸다. 구성 A의 경우 97년을 훈련 데이터로 사용한 관계로 97년에 대한 예측 결과가 없으며 구성 B와 C의 경우 95년 데이터의 예측은 95년 이전 데이터의 부재로 인해 불가능하다. 또한 구성 A의 경우 97년 데이터로 97년 이전 예측은 큰 의미가 없으나 다른 구성 방법과의 비교를 위해 실행하였다. 부연할 점은 99년 데이터부터 전공 이탈자에 대한 정보만 제공된 관계로 99년 데이터를 훈련 데이터로 사용할 수 없었으며 이는 99년 이후 구성 B와 구성 C의 실행을 불가능하게 하였다.

전반적으로 표 6의 각 구성기법들의 오분류율들은 상당히 높은 것으로 나타났으나 각 구성기법간의 비교에는 그나름대로 유용한 결과를 보여 주고 있다. 즉 각 구성기법들의 비교를 통해 다음과 같은 사실들을 주목할 수 있다. 구성 A의 경우 년도가 97년에서 멀어짐에 따라 예측 오분류율이 증가하며, 구성 B는 연도가 바뀜에 따라 훈련 오분류율과 예측 오분류율 모두 증가하며, 구성 C는 다른 기법들에 비해 훈련 및 예측 오분류율이 안정적이라고 할 수 있다. 특히 99년 예측의 경우 타 구성 기법들에 비해 구성 C는 훨씬 나은 결과를 보여 주고 있다. 이러한 결과를 토대로 특정 년도를 훈련 데이터로 고정하여 사용하는 것(구성 A)이나 과거의 데이터를 모두 누적하여 사용하는 것(구성 B)보다는 훈련 데이터를 계속 update하여 바꿔가면서 사용하는 것(구

성 C)이 바람직한 결과를 가져다 줄 수 있는 것으로 보인다. 이러한 결과는 전공이탈 제도에 대한 학생들의 반응 양태가 매년 빠르게 변화되고 있다는 사실을 암시하고 있으나 이는 분석 대상기간이 전공 이탈제도가 처음 도입된 시기여서 학생들의 반응 양태가 아직 안정적이지 못하다는 사실과도 연관되어 있는 듯하다.

4. 결론

박철용과 송규문(2002)은 기초학문분야인 인문학부와 자연과학부 학생들의 신상자료와 학적자료에만 근거하여 의사결정나무를 이용하여 전공이탈 학생들의 특징을 분석하였다. 본 연구에서는 이들의 연구 결과를 확장하여 전공이탈가능 학생에 대해 실제로 사용될 수 있는 예측모형을 신경망을 통해 구축하고자 하였다. 이를 위해 입력변수 특히 성적 변수 변환들을 통해 새로운 변수들을 도입하였으며 훈련 데이터 구성 문제에 대한 여러 가지 가능성들을 살펴보고 대안을 제시하였다. 본 연구의 주된 결과는 첫째 학생들의 성적 분포 및 변화에 대한 정보가 이탈 학생 분석 혹은 예측을 위해 유용할 수 있으며 (이는 기존 자료에 대한 변수 변환의 중요성을 보여주고 있다) 훈련자료로써 어느 특정 년도를 사용하거나 과거 모든 자료를 사용하는 접근법은 현재로써는 그다지 효율적이지 못하고 바로 전년도의 자료만을 사용하는 방법은 효율적일 수 있다는 사실을 보였다. 물론 이러한 결론은 연구의 완결을 뜻하는 결론이라기 보다는 이 분야 연구의 시발점으로써의 결론이라 생각한다. 즉 이러한 연구 결과의 타당성을 입증하기 위해서는 좀 더 긴 시간에 걸친 연관 데이터 축적 및 신경망 훈련의 기술적 문제에 대한 연구가 필요한 것으로 생각된다.

참고문헌

1. 박철용, 송규문 (2002). Analysis of students leaving their majors using decision tree, *Journal of the Korean Data & Information Society,* vol 12, 157-166.
2. Azoff, M. E. (1994). *Neural Network Time Series Forecasting of Financial Markets,* John Wiley and Sons, New York.
3. Haykins, S. (1999). *Neural Networks; a comprehensive foundation,* Prentice Hall, New Jersey.

[2002년 12월 접수, 2003년 1월 채택]