Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 481~488

Efficient Training Data Construction Scheme for Prediction of Transferring Students

Ji Young Lee[1] · Gyu Moon Song[2] · Tae Yoon Kim[3]

Abstract

Kim et al.(2003) studied a prediction model for students likely to transfer. In their study they claim that a training data construction scheme is better than other schemes, which trains neural network on the data from the year right before prediction year. One problem with their claim is that it is based on rather high prediction error rate. In this paper we establish a more sound comparison for various training data construction schemes and check validity of their claim. It turns out that the favored scheme has sufficient advantages over other schemes.

Keywords : 전공이탈, 학습데이터 구성기법, 신경망, 예측모형

1. 서 론

2000년 이후 예상되는 학생 인적자원의 감소에 따른 수요자 중심의 교육제도 개혁은 대학내(혹은 대학간)전공 사이에서 학생 이동을 일상화시켰다. 즉 전공 선택의 자유 확대는 한번 선택한 전공도 대학간 편입학 혹은 대학내 전부/전과 제도를 통해 다른 전공으로 바꾸는 것을 용이하게 하였다. 이로 인해 해가 거듭될수록 지방 사립대학을 비롯한 많은 대학들은 순수 자연과학 및 인문과학과 관련된 전공들의 학생수가 줄어드는 심각한 위기를 경험하게 되었다. 이러한 외부 여건의 악화와 제도 변화에 따라 대부분의 대학들은 이탈 가능한 학생들을 사전 파악하여 지도하는 문제에 지대한 관심을 갖게 되었다. 박철용과 송규문(2002), 김태윤, 이지영, 송규문(2003)은 이러한 문제에 대한 통계적 접근에 관심을 갖고 이탈 가능한 학생들의 예측을 위한 통계

1) Department of Statistics, Keimyung University, Taegu 704-701, Korea
E-Mail: frehop07@jinri.kmu.ac.kr
2) Professor, Department of Statistics, Keimyung University, Taegu 704-701, Korea
E-mail: kms252@kmu.ac.kr
3) Professor, Department of Statistics, Keimyung University, Taegu 704-701, Korea
E-Mail: tykim@kmu.ac.kr

적 분석 및 예측모형을 구축하고자 하였다. 즉 박철용과 송규문(2002)은 의사결정나무를 사용하여 대구소재 한 사립대학 내 전부/전과 데이터를 분석하여 이탈 가능성을 판단할 수 있는 주요 변수들을 제시하였으며 김태윤 등(2003)은 그 외에 입력변수로 사용될 수 있는 변수변환들을 제시하고 학습용 데이터의 다양한 구성기법에 대해 연구하였다. 특히 김태윤 등(2003)은 세 가지 유형의 학습 데이터 구성기법들을 고려하여 비교한 결과 "예측 대상의 바로 전년도 데이터만을 학습용 데이터로 사용"하는 구성기법이 가장 적절한 구성 기법임을 밝혔다. 그러나 그들의 분석 결과를 살펴보면 전반적으로 예측 오분류율들이 상당히 높은 것을 알 수 있으며(표 2.2 참조), 이는 비교 결과에 대한 신뢰도를 어느 정도 훼손시킬 수 있다고 판단된다. 본 연구에서는 이러한 높은 오분류율을 개선시키는 신경망 구조 조정 작업을 통해 김태윤 등(2003)의 결론 (예측 대상 바로 전년도 데이터만을 학습용 데이터로 사용하는 구성기법이 가장 적절한 기법이다)이 계속 유효한지 확인한다. 본 연구의 결과는 실제 사용 가능한 전공 이탈학생의 예측모형을 개발하는데 큰 도움이 될 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서는 김태윤 등(2003)의 결과의 문제점을 살펴보고 3장에서는 주어진 학습 데이터에 대한 신경망 선택 및 조정과정을 살펴보며 4장에서는 이들 신경망을 사용하여 여러 학습 데이터 구성 기법 중 가장 최적이라고 판단되는 구성기법을 논한다.

2. 배 경

김태윤 등(2003)이 사용한 자료는 대구 소재 어느 사립대학의 인문학부와 자연과학부의 1995년부터 2002년까지 입학한 학생들의 신상자료와 학적자료이다. 인문학부의 4개 전공과 자연과학부의 9개 전공으로 구성된 전체 데이터는 학적상태에 따라 재학, 전과, 졸업, 제적으로 구성되며 이 중 전과 및 제적을 전공 이탈로, 졸업을 전공 비이탈로 정의하였다. 분석에 포함된 학생은 모두 2081명이며, 그 중 56.6%인 1177명이 전공을 이탈했으며 나머지 43.4%인 904명이 이탈하지 않았다. 분석을 위하여 설명변수로써 고려된 것은 범주형 변수로 전공, 성별, 주야구분, 입학년도, 교직이수여부, 출신지역 등과 양적 변수로 1학년 1학기부터 2학년 2학기까지 각 학기의 성적과 장학금 수혜율 등이다. 이외에 다음과 같은 성적 자료의 변환된 변수들을 추가로 고려하였다.

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

여기서 성적ij는 i학년 j학기 성적을 뜻한다. 반응변수에 대한 설명력을 알아보는 기초분석으로서 범주형 변수에 대해서는 카이제곱 검정을 시행하고, 양적 변수에 대해서는 두 표본 t-검정을 수행하였는데 입력변수들 중 출신지역을 제외한 모든 설명 변수들이 상당히 유의한 것으로 나타났다. 따라서 출신지역은 본 연구에서 제외되었으며 실제 분석에 사용된 입력변수들은 표 2.1에 나타나있다.

표 2.1. 분석에 사용된 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
목표변수	ch	이탈여부 (이탈=1, 비이탈=0)

김태윤 등(2003)은 이 데이터에 인공 신경망을 이용하여 이탈 가능 학생에 대한 예측모형을 개발하였다. 그 과정에서 학습 데이터 구성문제를 제기, 연구하였는데 이는 현재에서 미래를 예측하고자 할 경우 학습에 사용될 수 있는 데이터는 과거로부터 현재까지의 모든 데이터이지만 그 모든 데이터가 유용하지 않을 수 있기 때문이다. 예를 들어 2003년에서 2004년을 예측하고자 하는 경우 2003년 데이터는 많은 도움이 되는 반면 과거 5년 전인 1998년의 데이터는 별로 도움이 되지 않거나 오히려 예측을 방해할 수 있다. 이는 시간의 흐름에 따라 전공이탈 학생들의 행동 반응 양태가 먼 과거와는 많이 달라 질 수 있기 때문이다. 따라서 과거 데이터 중 어느 시기의 데이터를 학습용 데이터로 사용할 것인가는 김태윤 등(2003)의 중요한 연구 결과 중 하나였다.

그들이 고려하여 비교한 세 가지 유형의 학습 데이터 구성기법은 다음과 같다.

(A) 1997년을 훈련 데이터로 사용한다.

(B) 1995년 이후부터 예측 대상 전년도까지의 모든 데이터를 학습용 데이터로 사용한다.

(C) 예측 대상 바로 전년도 데이터만을 학습용 데이터로 사용한다.

여기서 구성기법 (A)가 고려된 이유는 1997년은 처음으로 학부제가 도입되어 전과 및 전부가 실제로 시작된 년도로써 일종의 기준 년도로써의 역할이 기대되었기 때문이다. 그들의 비교 결과, 구성기법 (C)가 가장 적절한 구성 기법임이 밝혀졌다. 그들의 비교 결과는 "고정된 신경망"을 각 구성 기법에 대해 적용, 비교한 결과였으며 구체적으로 살펴보면 아래 표 2.2와 같다.

표 2.2 각 구성기법의 오분류율 [4)]

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%

4) 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.

여기서 “고정된 신경망”은 신경망 구조를 기준 년도인 1997년도 데이터에 잘 적합되도록 구성된 신경망을 뜻한다. 표 2.2의 각 구성기법들의 비교는 구성기법 (C)가 가장 효율적이라는 나름대로 유용한 결과를 보여 주고 있으나 전반적으로 예측 오분류율들이 상당히 높은 것을 알 수 있다. 이러한 높은 예측 오분류율은 기본적으로 매년 학습 데이터 변화에 상관없이 “고정된 신경망”을 각 구성기법에 적용한 결과인 것으로 판단되며 이는 비교 결과에 대한 신뢰도를 어느 정도 훼손시킬 수 있다고 생각된다. 본 연구에서는 이러한 높은 예측 오분류율을 개선하기 위해 각 학습 데이터가 바뀔 때마다 적용되는 신경망을 그 학습 데이터에 적합한 구조로 조정한 후 구성 기법들을 비교하는 작업을 수행한다. 그 결과 예측 오분류율들이 김태윤 등(2003)에 비해 크게 개선됨을 알 수 있었고 그 경우에도 구성기법 (C)가 여전히 효율적인 구성기법으로 드러났다.

3. 학습데이터에 따른 신경망 조정

본 연구에서 사용된 신경망은 가장 널리 사용되는 다층 퍼셉트론(multilayer perceptron, MLP) 신경망이며 이는 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성된 전방향 신경망 모형이다. 일반적인 MLP 신경망 구조는 그림 3.1과 같다. 신경망 모형 실행에서 가장 널리 알려진 알고리즘은 1960년대 초 Rumenlhart와 McClelland에 의해 개발된 역전파 알고리즘인데 이는 MLP 신경망을 학습시키는 알고리즘으로써 실제 주어진 데이터 값과 신경망 모형을 통해 예측된 값을 비교하면서 학습된 데이터를 반복 처리 학습해 나간다. 신경망 모형에 대한 알고리즘과 학습방법은 Haykins(1999)를 참조하기 바란다. 본 연구에서 사용한 신경망 모형은 역전파 알고리즘을 사용한 MLP 신경망이다.

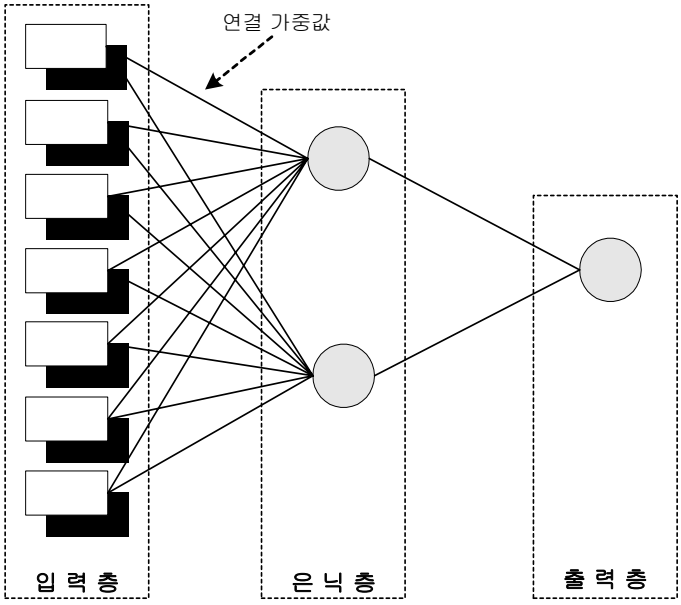

그림 3.1. MLP 신경망 구조

학습용 데이터가 바뀔 때마다 그에 적절한 신경망 구조를 찾기 위해 여러 개의 신경망 구조들을 시도해 보았다. 먼저 학습용 데이터를 분석용 33%, 평가용 67%로 분할하여 분석을 한 후 평가 오분류율이 일정 수준의 작은 값이 되도록 (27-30%) 은닉층 및 은닉노드 개수를 선택한다. 물론 입력변수 및 출력변수의 수는 변함이 없으므로 19개의 입력노드로 구성된 입력층과 1개의 노드로 구성된 출력층을 사용한다. 이 분석에서 사용된 신경망의 일반적인 구조를 살펴보면 그림 3.2와 같다.

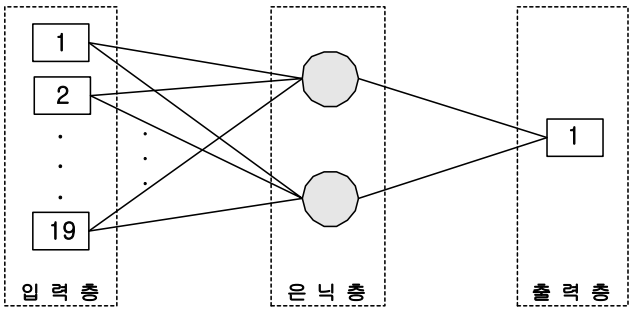

그림 3.2. 신경망 구조

그림 3.2의 은닉층 부분은 경우에 따라 여러개의 노드로 구성된 1개 혹은 2개의 은닉층을 갖게됨을 의미한다. 여기서 각 학습 데이터마다 일정한 수준의 낮은 평가 오분류율을 갖도록 하기 위해서 신경망 모형을 적절히 조절하는 과정, 즉 은닉층의

개수와 노드수를 조정하는 과정은 주로 반복 시행 과정을 이용하였다. 일정한 수준의 낮은 학습 평가 오분류율이라는 목표는 구성기법들을 객관적으로 비교하기 위해서는 반드시 필요한 기준이라고 판단된다. 각 학습 데이터에 대한 구성 기법에 따른 학습 오분류율 결과가 표 3.1에 주어져있다. 표 3.1에 김태윤 등(2003)의 구성기법 (A)는 포함되지 않았는데 그 이유는 구성기법 (A)의 경우 김태윤 등(2003)에 의해 가장 부적절한 구성기법으로 이미 검증되었을 뿐만 아니라 학습 데이터가 1997년으로 고정된 관계로 추가 조정과정이 불필요하기 때문이다.

표 3.1. 구성기법에 따른 오분류율

구성기법	년도	학습 오분류율		예측 오분류율
		분석용	평가용	
B	95년	.	.	.
	96년	0.157	0.260	0.36
	97년	0.243	0.270	0.31
	98년	0.280	0.270	0.52
	99년	0.310	0.320	0.77
C	95년	.	.	.
	96년	0.157	0.260	0.36
	97년	0.272	0.319	0.31
	98년	0.278	0.277	0.48
	99년	0.140	0.310	0.23

4. 결 론

김태윤 등(2003)은 학습 데이터 구성 문제를 제기, 가장 좋은 구성기법으로써 구성기법 (C), 즉 예측 대상 바로 전년도 데이터만을 학습용 데이터로 사용할 것을 제안하였다. 그러나 그들의 결과는 2장에서 논의한 바처럼 너무 높은 예측 오분류율로 인해 신뢰성에 문제가 있을 수 있다. 이를 해결하기 위해 3장의 학습데이터 변화에 따른 “조정된 신경망”들을 사용하여 두 가지 학습 데이터 구성기법 (B)와 (C)간의 비교를 수행한 결과(표 3.1) 다음과 같은 사실을 관찰할 수 있다.

구성기법 (B)를 사용한 경우의 학습 데이터의 분석 및 평가 오분류율은 96년 16%,

26%, 97년 24%, 27%, 98년 28%, 27%, 99년 31%, 32% 등이었으며 구성기법 (C)를 사용한 경우의 오분류율은 96년 16%, 26%, 97년 27%, 31%, 98년 28%, 28%, 99년 14%, 31% 등이었다. 여기서 평가 오분류율들을 살펴보면 일관되게 27~30%이므로 각 구성기법의 예측 오분류율의 비교가 김태윤 등(2003)에 비해 더 합리적이고 객관적으로 수행될 수 있는 여건이 조성됐음을 알 수 있다. 참고로 김태윤 등(2003)은 표 2.2에서 보는 바처럼 고정된 신경망에 대한 일정 범위의 학습 오분류율(12-27%)들을 기준으로 비교를 수행한 결과 예측 오분류율들이 상당히 높게 (구성기법 B: 31 -78%, 구성기법 C: 41-59%) 관찰되었다. 본 연구의 결과, 구성기법 (B)의 예측 오분류율의 범위는 마찬가지로 높게 나타났으나 (31-77%) 구성기법 (C)의 경우 예측 오분류율의 범위가 23-48%로 크게 개선됨을 알 수 있었다. 이러한 결과를 통해 구성기법 (B)보다 구성기법 (C)가 해가 거듭되면서 예측 오분류율이 점점 줄어든다는 것을 확인할 수 있다. 즉 구성 (B)와 같이 과거의 데이터를 모두 누적하여 사용하는 것보다는 학습 데이터를 계속 추가하여 바꿔가면서 사용하는 구성 (C)가 바람직한 결과를 가져다 줄 수 있다는 김태윤 등(2003)의 결론은 여전히 유효한 것으로 보인다. 다시 말하여 구성기법 (B)의 경우 누적 년도의 수가 증가하면 예측효율이 기하급수적으로 저하되어 예측효율을 안정적으로 관리하기가 불가능해 보이지만 (과거의 누적 데이터들이 예측을 방해할 수도 있는 것처럼 보인다) 구성기법 (C)의 경우 예측효율을 어느 정도 안정적인 수준에서 관리할 수 있을 것으로 판단된다. 한 가지 흥미로운 사실은 99년의 경우 구성기법(C)를 사용할 때 예측 오류율이 급격히 감소함을 알 수 있는데 이러한 현상에 대해 좀더 연구가 필요한 것으로 판단된다.

참고문헌

1. 박철용, 송규문 (2002). Analysis of students leaving their majors using decision tree, Journal of the Korean Data & Information Science Society, vol 12, 157-166.
2. 김태윤, 이지영, 송규문 (2003). Neural network analysis of transferring students, Journal of the Korean Data & Information Science Society, vol 14, 11-21.
3. Azoff, M.E. (1994). *Neural Network Time Series Forecasting of Financial Markets.* John Wiley and Sons, New York.
4. Haykins, S. (1999). *Neural Networks; a comprehensive foundation.* Prentice Hall, New Jersey.

[2003년 5월 접수, 2003년 7월 채택]