Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 2, pp. 269 ~ 278

A Study on Improving the predict accuracy rate of Hybrid Model Technique Using Error Pattern Modeling : Using Logistic Regression and Discriminant Analysis

Yong-Jun Cho[1] · Joon Hur[2]

Abstract

This paper presents the new hybrid data mining technique using error pattern, modeling of improving classification accuracy. The proposed method improves classification accuracy by combining two different supervised learning methods. The main algorithm generates error pattern modeling between the two supervised learning methods(ex: Neural Networks, Decision Tree, Logistic Regression and so on.) The Proposed modeling method has been applied to the simulation of 10,000 data sets generated by Normal and exponential random distribution. The simulation results show that the performance of proposed method is superior to the existing methods like Logistic regression and Discriminant analysis.

Keywords : Combined Model, Error pattern modelling, Hybrid Error Modeling, Voting

1. 서론

독립변수와 목적변수의 관계를 통해 각종 인과성과 예측을 위한 방법으로 통계학은 여러 가지 분석 방법을 제공한다. 회귀분석, 로지스틱 회귀분석, 판별분석 등과 같은 분석 방법들의 궁극적인 목표는 인과 관계에 대한 모형의 적합성을 높여, 향후 새로운 독립변수들이 주어졌을 때, 목적변수의 값에 대한 예측 정확도를 높이는 것이다.

전통적인 통계적인 개념이 아닌 근래의 AI(Artificial Intelligence) 기반의 데이터 마이닝에서도, 위의 언급된 분석들과 유사한 신경망분석이나 의사결정나무분석 등을 수

1) 서울 송파구 오금로 62 수협 수산경제연구원, 수석연구원
E-mail : cyj66@chol.com
2) 서울 강남구 역삼동 701-2 삼정개발빌딩 3층 SPSS Korea, 컨설팅팀장,
E-mail : hoh@spss.co.kr

행하게 되는데, 이 때 또한 목표는 가장 적합한 모형을 만들어, 예측력 및 오차율을 줄이는 것이 된다.

이러한 예측력을 높이고, 오차율을 줄이기 위한 방법은 매우 다양하다. 예를 들어 Michie 등(1994)은 어떤 상황에서 어떤 기법이 가장 적합한 것인가에 대한 연구를 수행하였으며, Dougherty 등(1995)은 연속형의 데이터를 범주형으로의 데이터 변환을 통하여 학습능력 및 처리 속도를 향상 시켜 전체적인 예측율의 향상에 대한 연구를 수행하였다. 이와 같은 예측율을 향상시키기 위한 여러 방법 중 하나가 바로 단일 알고리즘의 방법이 아닌 2개 이상의 방법을 혼용하거나 데이터의 분할을 통한 가중치를 변경하여 단일 알고리즘이 가지는 모형의 한계를 극복하고자 하는 방법이다. 이러한 방법 중 대표적인 것이 Hybrid 방법과 Combined 방법이다. Hybrid 모형과 Combined 모형의 차이는 일반적으로 전혀 다른 분석 기법을 융합하여 사용하는 경우를 Hybrid 모형이라고 하고, 동일 기법 내에서 여러 개의 데이터 집합이나 가중치를 변경하여, 여러 개의 모형을 만들어 이를 결합하는 것을 Combined 모형이라고 정의할 수 있다.

이런 Hybrid 방법과 Combined 방법을 통해서 기존의 단일 알고리즘보다 더 좋은 예측율을 가져오거나, 오차율을 줄이고자 하는 연구가 기존에 국내외에서 다양한 형태로 전개되어져 왔다.(참고문헌 1, 2, 4, 5, 6, 9, 10, 11, 13, 14 등이 대표적이다.) 그리고 Hansen과 Salaman(1990)은 이런 Hybrid 방법이나 Combined 방법이 기법들 간에 오차율이 독립적으로 분포할 때 가장 유용한 방법이라는 구체적인 연구를 수행하였다.

또한 현재 가장 많이 사용되는 Hybrid 방법 및 Combined 방법으로는 Voting, Bagging, Boosting 등이 있고, 이들 방법은 현재 통계 패키지 및 데이터 마이닝 패키지에 알고리즘이 내장되어, 보편적으로 일반적인 사용자들이 손쉽게 할 수 있도록 되어져 있다.

보편적으로 알려져 있는 방법들 이외에 여러 방법을 Hybrid 하는 방법이 제시되어져 있다. 그 중 허준, 김종우(2005)는 오차의 패턴을 또 하나의 모형으로 만들어, 이를 이용하여 2개 이상의 여러 분석 방법 중 더 좋은 방법을 판별하고, 이를 적용하여 최종 예측값을 제공하는 오차 패턴 모형화 Hybrid 방법을 제안하였다. 해당 연구에서는 데이터 마이닝의 보편적인 방법인 신경망과 의사결정나무 알고리즘인 C5.0을 이용하여, 오차 패턴 모형을 개발하고, 실제 데이터 10개의 사례를 이용하여 그 성능이 단일 알고리즘 또는 일반적인 Voting이나 Boosting 방법보다 더 효율적인 것을 증명하였다. 또한 저자들은 이렇게 만들어진 Hybrid 모형의 수행시간이 단일 알고리즘을 사용할 때 보다 다소 증가하지만, 여러 개의 알고리즘 중 가장 효율적인 알고리즘을 선택하여 최적의 효율을 낼 수 있는 방법임으로 카드 사기나 의료 데이터 등 정확도에 매우 민감한 분야에서는 크게 유효할 것으로 판단된다.

본 논문에서는 허준, 김종우(2005)가 제시한 오차 패턴 모형화(해당 논문에서는 모델링이란 표현을 사용하였다.)를 통한 Hybrid 방법이 제안된 C5.0과 신경망 분석이 아닌 전통적 통계방법론인 로지스틱 회귀분석과 판별분석에서도 유효한 성능을 나타내는지 확인하고, 또한 10개의 데이터 사례를 통해서 성능을 증명한 것과 달리 시뮬레이션을 통해 성능향상의 일반화하는 것이 목적이다. 이런 확인을 통해 오차 패턴 모형화를 이용한 Hybrid 방법이 효율적인 것이라는 타당성을 제공하고자 한다.

2. 오차 패턴 모형화를 통한 Hybrid 모형의 이론적 배경

2.1 오차 패턴 모형화를 통한 Hybrid 모형의 개념

데이터 마이닝에서 2개의 서로 다른 알고리즘을 적용한 결과가 동일한 예측정확도를 나타내는 두 모형이 있다고 가정하자. 하지만, 이 두 모형은 동일한 성능을 지닌 모형이라고 볼 수 없다. 왜냐하면, 실제 적용에 있어서 A의 모형을 통해 더 잘 예측되는 레코드(사례)가 있고, B의 모형을 통해 더 잘 예측되는 레코드가 있기 때문이다. 즉, 이 두개의 모형 중 서로 잘 맞추는 경우만을 선택하는 모형이 있다면, 성능이 향상될 것이다. 이러한 모형을 Hybrid 모형이라고 한다.

오차 패턴 모형이란 서로 다른 2개 이상의 알고리즘을 동일한 데이터에 적용하여, 2개 이상의 모델이 서로 다른 결과를 내는 경우만 추출하여, 데이터 집합을 구성하고, 이 데이터 집합을 가지고, 다시 A방법과 B방법이 잘 맞추는 오차 데이터 모형을 생성한 다음, 실제 적용할 데이터 집합에서는 각 사례에 대하여 Hybrid 모형과 같이 A방법과 B방법이 서로 잘 맞추는 사례를 맞추게 하여, 최종적으로는 오분류 및 잘못된 예측이 가장 적은 예측결과를 만들어 내는 방법이다.

2.2 오차 패턴 모형화를 통한 Hybrid 모형의 과정

오차 패턴 모형을 이용하여 Hybrid 모형을 만드는 과정을 정리하면 다음과 같다.

(1) 전체 훈련용(Training) 데이터 집합(set)을 S, y_n은 목적변수, $x_{i,n}$은 독립변수, $i = 1, \cdots, I$는 설명변수의 수, $n = 1, \cdots, N$은 데이터의 레코드 수로 정의한다. 그리고 훈련용 데이터를 통해서 나온 모형을 검정하기 위한 시험용(Test)데이터 집합을 V로 정의한다.

(2) 이 전체 훈련용 데이터 집합 S를 임의추출을 통해 2개로 분리를 한다. 이 2개의 데이터 집합을 다음과 같이 정의한다.

$$S_1 = \{(y_m, x_{i,m}), m = 1, 2, \ldots . M\}, \quad S_2 = \{(y_p, x_{i,p}), p = 1, 2, \ldots . P\}$$

단, $M + P = N$,

(3) 먼저 S_1 데이터 집합을 이용하여, 분석기법 A를 이용하여, 모형화를 수행한다. 이 때 생성된 기법 A의 모형을 $M(A)$라 하자. 또한 분석기법 B를 이용하여 모형화를 수행하고 이에 생성된 모형을 $M(B)$라 하자.

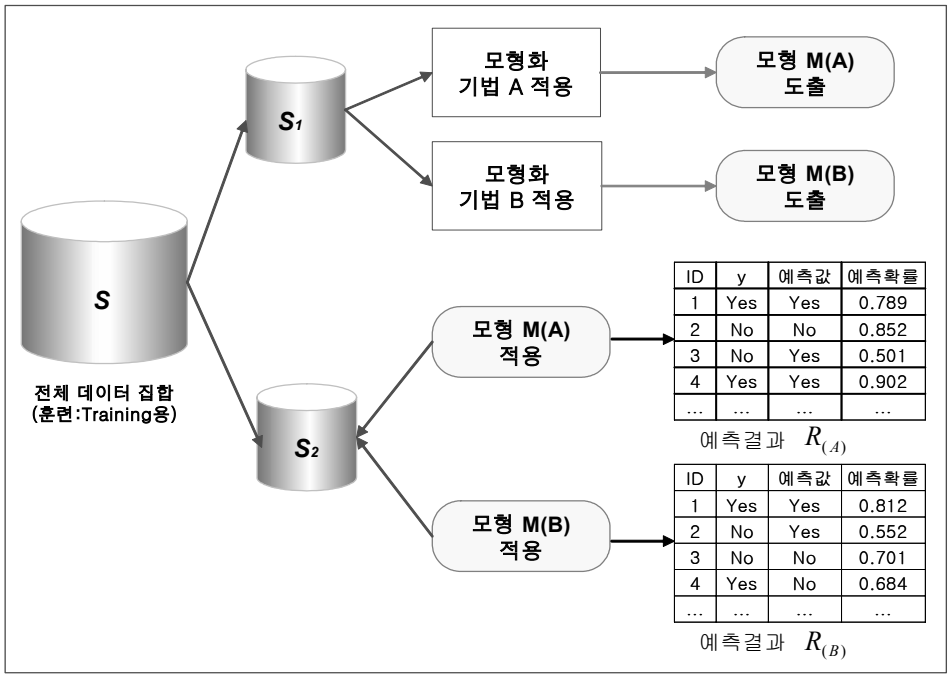
(4) 다음으로 단계 (2)에서 분리한 또 다른 훈련용 데이터 집합인 S_2에 $M(A)$와 $M(B)$를 적용시킨다. 먼저 $M(A)$를 적용시켜서 나온 예측결과를 $R_{(A)}$라고 하자. 마찬가지로 모형 $M(B)$를 적용시켜서 나온 예측결과를 $R_{(B)}$라고 하자. 지금까지의 과정은 다음 <그림 1>과 같다.

(5) 데이터 집합 S_2를 통해서 나온 2개의 결과 값 $R_{(A)}$와 $R_{(B)}$를 서로 비교하여 결과값이 서로 틀린 데이터 집합만을 추출한다. 이렇게 추출해 낸 데이터 집합을 S_D라고 정의한다. 즉, $S_D = set\{R_{(A)} \neq R_{(B)}\}$이다. 이 때 S_D는 K개의 레코드를 지니게 되고

이는 서로 다른 결과값의 개수가 된다. 또한 $k = 1, \cdots, K$이 된다.

(6) 다음 데이터 집합 S_D에서 목적변수 y_k값과 기법 A를 이용하여 생성된 예측 결과치 $R_{(A),k}$와 비교하여 서로 일치하면 T 아니면 F인 새로운 목적변수를 생성한다. 즉, k번째 목적변수의 값과 예측 결과치를 비교하여 서로 일치하면 T 아니면 F를 할당한다. 이렇게 새롭게 파생된 목적변수의 집합을 $T_{(A,S_D)}$ 라고 하자. 반대로 목적변수 y_k 값과 기법 B를 이용하여 생성된 예측 결과치 $R_{(B),k}$와 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적변수를 생성한다. 이렇게 새롭게 파생된 목적변수의 집합을 $T_{(B,S_D)}$라고 하자.

예를 들어 목적변수가 '1' 또는 '2'라고 가정하자. 그리고 ID001인 레코드가 있다고 가정하고 이것의 목적변수는 '1'이라고 하자. 여기에 $M(A)$를 적합한 예측결과는 '1'로 산출되었고, $M(B)$를 적합한 예측결과는 '2'로 산출되었다. 그렇다면 ID001는 S_D 데이터 집합에 포함시키게 된다. 이때 기법 A를 적용한 결과는 원래의 목적변수인 '1'을 정확히 예측하였다. 그렇다면 새로운 목적변수 집합인 $T_{(A,S_D)}$에는 'T'의 값이 입력된다. 반대로 기법 B를 적용한 결과는 원래의 목적변수인 '1'의 값에 대한 예측이 틀렸으므로 새로운 목적변수 집합인 $T_{(B,S_D)}$에는 'F'의 값이 입력되게 된다.



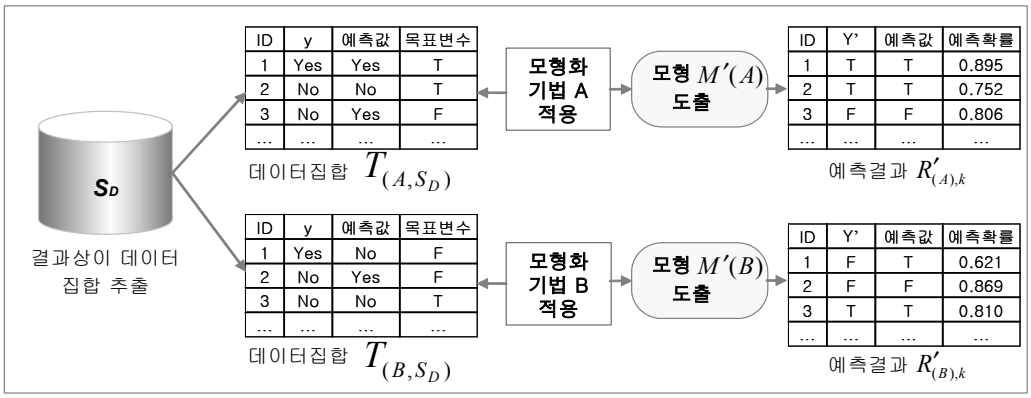
<그림 1> 오차 데이터 모형에서 초기 데이터 분할과 두 기법의 적용

(7) S_D 데이터 집합에서 기존의 목적변수 y_k대신에, 새롭게 만들어진 목적변수의 집합

$T_{(A,S_D)}$와 $T_{(B,S_D)}$를 바탕으로 데이터 집합 $T_{(A,S_D)}$에 모형화 기법 A를 다시 적용하고 이렇게 생성된 모형을 $M'(A)$라고 하자. 또한 $T_{(B,S_D)}$에 모형화 기법 B를 다시 적용하고 이를 통해 생성된 모형을 $M'(B)$라 하자.

(8) 모형 $M'(A)$를 적용시켜서 나온 예측결과를 $R'_{(A),k}$라고 하고 모형 $M'(B)$를 적용시켜서 나온 예측결과를 $R'_{(B),k}$라고 하자. 이를 이해하기 쉽게 정리한 그림이 <그림 2>이다. 예를 들어 (6)의 사례를 바탕으로 ID001은 2개의 데이터 집합 $T_{(A,S_D)}$와 $T_{(B,S_D)}$에 들어가게 된다. 또한 목적변수가 'T' 또는 'F'로 바뀌게 된다. 이를 기법 A를 통해 다시 적용하였을 때, ID001은 집합 $T_{(A,S_D)}$에 속하므로 목적변수의 값은 'T'이고 이를 바탕으로 예측된 결과는 'T' 또는 'F'의 값을 예측하게 될 것이다. 마찬가지로 기법 B를 통해 다시 적용하였을 때, ID001은 집합 $T_{(B,S_D)}$에 속하므로 목적변수의 값은 'F'이고 이를 바탕으로 예측된 결과 역시 'T' 또는 'F'의 값을 예측하게 될 것이다.

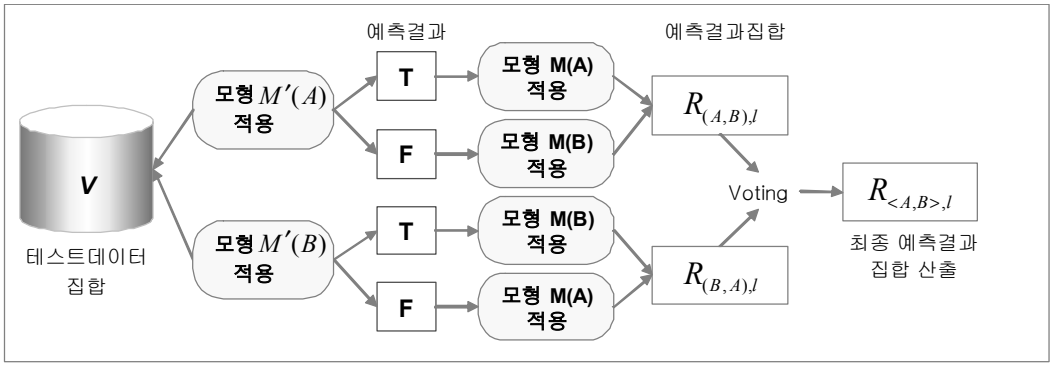
이번 단계에서 만들어진 모형 $M'(A)$와 이에 따라 산출된 예측결과 $R'_{(A),k}$ 그리고 모형 $M'(B)$와 이에 따라 산출된 예측결과 $R'_{(B),k}$의 의미는 2개의 기법 A와 B가 서로 다른 결과를 낸 데이터만 모아둔 데이터 집합 S_D에서, 기법 A와 B가 서로 잘 맞추는 형태의 데이터 패턴을 다시 파악하는 로직이라고 할 수 있으며, 본 오차 패턴 모형화를 통한 Hybrid 모형의 핵심이라고 할 수 있다. 이를 오차 패턴 모형(Error Pattern Model)이라고 정의 한다.



<그림 2> 오차 패턴 모형화의 과정

(9) 다음 이렇게 오차 패턴 모형을 구했으면, 이를 적용한 최종 예측결과를 생성하게 되는데, 이 과정은 Voting이라는 방법을 통해 이루어진다. 예를 들어 시험용 데이터 집합인 V에 모형 $M'(A)$를 적용하고 산출된 예측결과 $R'_{(A),l}$에서 예측값 T로 예측되는 사례에 모형 $M(A)$를 적용하고 F로 예측되는 사례에 모형 $M(B)$를 적용한

다. 이를 통해 산출된 예측결과 $R_{(A,B),l}$이라 하자. 이때 L은 데이터 집합 V의 레코드 개수이고 $l=1,\cdots,L$이 된다. 반대로 모형 $M'(B)$를 적용하고 산출된 예측결과 $R'_{(B),l}$에서 예측값 T로 예측되는 사례에 모형 $M(B)$를 적용하고 F로 예측되는 사례에 모형 $M(A)$를 적용한다. 이를 통해 산출된 예측결과 $R_{(B,A),l}$이라 하자. $R_{(A,B),l}$와 $R_{(B,A),l}$에는 예측결과에 대한 각각의 예측 확률값이 생성되어 있다. 이 확률값을 비교하여 확률값이 높은 예측결과를 선택하여 최종 예측결과를 산출한다. 이때 산출된 최종 예측결과를 $R_{<A,B>,l}$이라 하고 이를 통해 예측 및 분류를 수행한다. 오차모형을 통한 최종 결과산출 과정은 다음 <그림 3>과 같다.



<그림 3> 오차모형을 통한 최종 결과 도출과정

3. 시뮬레이션 결과 비교

3.1 시뮬레이션 개요

위에서 제시한 오차 패턴 모형화를 통한 Hybrid 모형의 예측 정확도와 단일 알고리즘과의 예측 정확도를 비교 검증하기 위한 시뮬레이션에 몇 가지 제약을 두었다. 첫째, 실험 데이터의 목적변수는 이분형(binary)형태이다. 둘째, 독립변수의 수는 4개로 한정하였다. 셋째, 이 4개의 독립변수 $x_{1,}\cdots,x_4$를 난수 생성을 통하여 산출한다. 이때, 목적변수의 범주 '0'은 정규분포($N(0,1)$)의 특성을 지닌 독립변수들로 가정하고 범주 '1'은 지수분포 ($exp(1,1)$)의 특성을 지닌 독립변수들로 가정한다. 넷째, 실험에 사용된 알고리즘은 로지스틱 회귀분석과 판별분석 2가지만을 사용하였다. 다섯째, 정확도 향상을 위한 조정(예를 들어 로지스틱 회귀분석의 'cut ratio=0.5'로 고정)을 하지 않는다.

시뮬레이션 과정은 위의 세 번째 조건에 따라 목적변수의 범주가 '0'인 경우를 20만 레코드(case), '1'인 경우 20만 레코드, 총 40만 레코드의 데이터를 생성하여, 이를 S_1 데이터 셋(set)(각각 10만 레코드, 총 20만 레코드)으로 할당한 후, 로지스틱 회귀분석

과 판별분석을 적용하여 모형을 생성한다. 생성된 모형을 같은 S_2 데이터 셋(각각 10만 레코드, 총 20만 레코드)에 적용하여 상호 예측 결과가 다른 데이터를 추출한다. 이를 오차 패턴 모형화를 통한 Hybrid 모형을 적용하여 모형을 생성한다. 또한 로지스틱 회귀분석과 판별분석을 통한 모형도 각각 생성한다. 이렇게 생성된 3개의 모형을 가지고, 어떤 모형의 예측정확도가 안정적으로 높은지를 파악하기 위하여 훈련용 데이터 셋을 동일한 조건으로 10,000개를 생성하고 각 훈련용 데이터 셋의 레코드는 100개('0'이 50, '1'이 50)로 한다. 10,000번의 시뮬레이션 결과를 통해 어떤 알고리즘이 더 예측 결과 높은지를 측정하였다. 여기서 훈련용 데이터 셋을 40만 레코드로 한 것은 대용량 데이터를 통해 충분히 모집단의 특성을 반영할 수 있도록 하기 위함이고, 10,000번의 시뮬레이션은 예측 결과치가 충분히 일반화할 수 있을 정도의 결과를 산출하기 위하여 설정한 수치이다.

즉, 본 연구의 시뮬레이션은 서로 다른 특성(분포)을 지니고 있는 목적변수의 특징을 어떤 알고리즘이 안정적이고 정확하게 분류·예측 할 수 있는지를 입증하고자 하였다.

3.2 시뮬레이션 결과

3개의 알고리즘이 적용된 예측 정확도는 예측 오차율을 계산하여 결과를 제시하도록 한다. 시뮬레이션 결과는 <표 1>에서 평균 오차율(%)과 이에 대한 표준편차를 제시하였다. 오차 패턴 모형화를 통한 Hybrid 모형은 단일모형에 비하여 오차율이 낮고, 표준편차 역시 적은 것으로 나타났다. 또한 <표 2>에서는 산출된 예측오차율에 차이가 있는지에 대한 대응표본 T-검정을 실시하였다. 오차 패턴 모형화를 통한 예측 오차율이 로지스틱 회귀분석, 판별분석보다 낮은 것을 알 수 있다. 이를 통해 오차 패턴 모형화를 통한 Hybrid 모형이 예측 정확도가 높다는 것을 나타내고 있다.

<표 1> 3개 모형별 오차율 결과

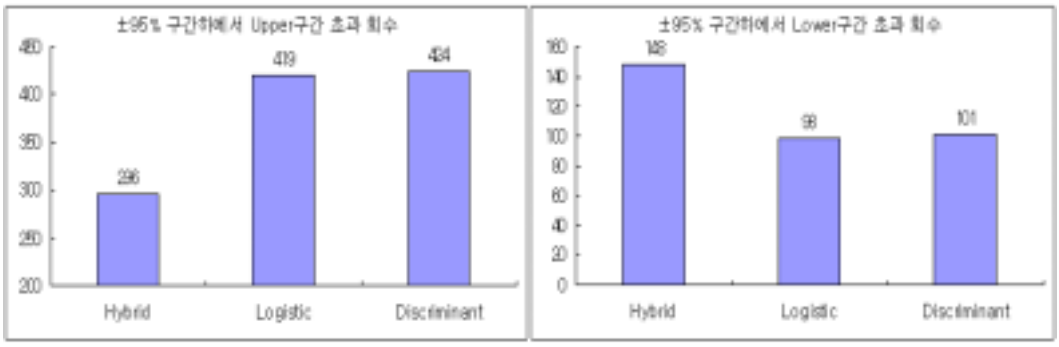
	횟수	Min(%)	Max(%)	Mean(%)	Std Dev.
오차 패턴 모형화 Hybrid 모형	10000	1.0	29.0	14.56	3.516
로지스틱 회귀분석 모형	10000	3.0	29.0	15.10	3.568
판별분석 모형	10000	3.0	29.0	15.10	3.573
total	30000	1.0	29.0	14.92	3.561

<표 2> 각 모형별 대응표본 T-검정 결과

대응모형		대응차평균	평균의 표준오차	t-value	P-value(양쪽)
대응 1	Hybrid-로지스틱	-0.5383	0.007341573	-73.3222	0.00000
대응 2	Hybrid-판별분석	-0.5323	0.007586925	-70.1602	0.00000
대응 3	로지스틱-판별분석	+0.006	0.002082507	2.881143	0.003971

<그림 4>는 3개 모형의 오차율을 모두 통합하여 산출된 전체 오차율에서 ±95% 구간을 설정하고 이 구간을 초과하는 회수에 대한 도표이다. 상위구간을 초과하는 횟수가 많다면 예측 오차율이 다른 모형에 비해 높다는 것으로 모형의 정확도가 떨어지는

것을 의미한다. 반대로 하위구간을 초과하는 횟수가 많다면 모형의 정확도가 다른 모형에 비해 높다는 것을 의미한다. 그림 (A)에서 오차 패턴 모형화를 통한 결과가 다른 모형에 비해 적게 나타났고 그림 (B)에서는 다른 모형에 비해 많이 나타남을 보여주고 있다. 이는 오차 패턴 모형화를 통한 Hybrid모형이 다른 모형에 비해 안정적으로 높은 예측 정확도를 산출한다는 것을 의미한다.

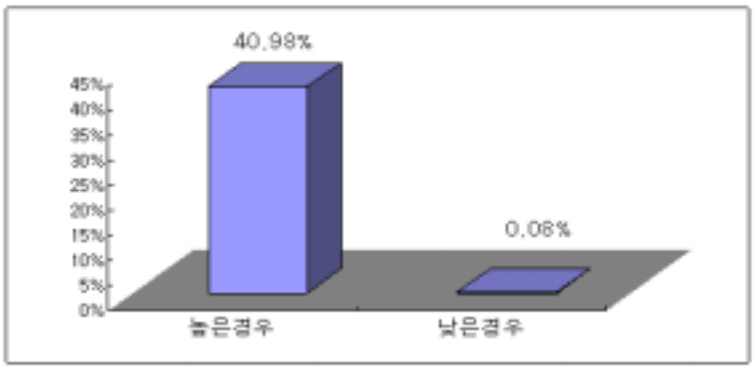


(A) 각 모형별 예측오차율이 높은 상위구간을 초과하는 회수

(B) 각 모형별 예측오차율이 낮은 하위구간을 초과하는 회수

<그림 4> 전체 오차율 ±95% 구간을 초과하는 회수

<그림 5>는 오차 패턴 모형화를 통한 Hybrid 모형의 예측 정확도와 단일 알고리즘 모형의 예측 정확도를 각 데이터 셋마다 비교한 결과이다. 오차 패턴 모형화를 통한 Hybrid 모형의 경우가 단일 알고리즘 모형 중 높은 정확도를 나타내는 모형보다 예측 정확도가 떨어지는 경우는 0.08%이고, 예측 정확도의 상승은 약 40.98%정도였으며, 로지스틱 회귀 모형과 판별분석 모형과 동일한 경우가 58.94%임을 나타낸다.



<그림 5> 오차 모형화를 통한 Hybrid 모형과 단일 알고리즘과의 예측 정확도 비교

이러한 결과를 바탕으로 오차 패턴 모형화를 통한 Hybrid 모형이 단일모형에 비해 안정적이면서, 예측 정확도가 높은 모형이라고 할 수 있다.

4. 결론 및 추후 연구과제

본 연구에서는 오차 패턴 모형화를 통한 Hybrid 모형이 기존의 단일 분석 알고리즘을 사용하여 분석(모형화)했을 때보다 얼마만큼 더 많은 정확도의 향상을 가지고 오는지에 대하여 실험하였다.

결론적으로 제한된 시뮬레이션을 통한 실험이었지만, 오차모형을 통한 Hybrid 모형이 단일 알고리즘보다 안정적이고 예측 정확도가 최소한 같거나 높은 것으로 나타났다. 이는 오차 패턴 모형화를 통한 Hybrid 모형이 안정성 및 예측정확도를 향상 시키는 것으로 볼 수 있다. 물론, 서로 다른 특성(분포)을 지니고 있는 이분형 목적변수로 가정하여 이에 대한 독립변수를 난수 생성을 통해 생성하여 실험하였으나, 실제 적용에 있어서 이분형 목적변수를 사용하는 경우가 상당히 많고, 목적변수를 분류·예측하는 것이 독립변수의 특징을 통해 이루어지는 것인 만큼 시뮬레이션을 통한 실험의 의의는 크다고 할 수 있다. 또한 허준, 김종우(2005)가 실제 사례를 바탕으로 C5.0과 신경망을 통한 같은 방식의 실험에 있어서도 예측정확도 및 안정성이 높은 것으로 제시하였다. 이의 결과를 종합하여 볼 때, 오차 패턴 모형화를 통한 Hybrid방식이 일반적으로 이분형 목적변수를 가지는 데이터 집합에서 안정적 예측정확도를 산출하는 것으로 판단된다.

본 시뮬레이션의 기본 제약으로 종속변수가 이분형 변수에 대해서만 실험을 하였고, 적합 알고리즘의 경우에 로지스틱 회귀분석과 판별분석에 한정하였다. 따라서 추후 연구과제로 다범주형 목적변수를 대상으로 데이터 마이닝에서 활용도가 큰 의사결정나무모형, 신경망모형, 로지스틱 회귀모형 등을 모두 결합하여 적합하고 2가지 모형일 때보다 더 안정적 예측정확도를 상승하는지에 대한 연구를 추후 연구과제로 남기기로 한다.

결론적으로 본 연구는 오차 패턴 모형화를 통한 Hybrid 모형이 안정성과 예측정확도를 만족시키는 적합한 솔루션일 수 있을 것이라고 제안한다.

참고문헌

1. 강문식, 이상용(2002), 데이터 마이닝을 위한 경쟁학습모델과 BP알고리즘을 결합한 하이브리드 신경망, *정보기술과 데이터베이스 저널*, 제9권 2호, pp. 1-16.
2. 이극노, 이홍철(2003), 이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘 연구, *한국지능정보시스템학회지*, 제9권 1호, pp. 139-155.
3. 허준, 김종우(2005), 오차 패턴 모델링을 이용한 Hybrid 데이터 마이닝 기법, *한국경영과학회지*, 제 30권 4호, pp. 27-44.
4. Carvalho, D. R and Fretias, A. A. (2004), Hybrid Decision Tree/Genetic Algorithm Method for Data Mining, *Information Science*, Vol. 163, No. 1, pp. 13-35.

5. Coenen, F. G. S., Swinnen, K. V. and Wets, G. (2000), The improvement of response modeling: Combining rule-induction and case-based reasoning, *Expert Systems with Application*, Vol. 18, pp. 307-313.
6. Conversano, C., Siciliano, R. and Mola, F. (2002), Generalized Additive Multi-mixture Model for Data Mining, *Computational Statistics & Data Analysis*, Vol. 38, No. 4, pp. 487-500.
7. Dougherty, J.,Kohavi, R., and Sahami, M. (1995), *Supervised and unsupervised discretization of continuous features*. In Proc. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, pp. 194-202.
8. Hansen, L. K. and Salaman, P. (1990), Neural Networks ensembles, *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp. 993-1001.
9. Indurkhya, N. and Weiss, S. M. (1998), Estimating Performance Gains for Voted Decision Trees, *Intelligence Data Analysis*, Vol. 2, No. 1, pp. 303-310.
10. Kuncheva, L .I, Bezdek. C. and Shutton, M. A. (1998), On Combining Multiple Classifiers by Fuzzy Templates, *International Conference on Artificial Neural Networks IEEE*, pp. 193-197.
11. Lin, F. Y and McClean, S. (2001), A Data Mining Approach to the Prediction of Corporate Failure, *Knowledge-Based Systems*, Vol. 14, No. 3, pp. 189-195.
12. Michie, D., Spiegelhalter, D. J. and Taylor, C. (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood.
13. Suh, E. H., Noh, K. C and Suh, C. K. (1999), Customer list segmentation using the combined response model, *Expert Systems with Applications*, vol. 17, pp. 89-97.
14. Zhou, Z-H., Wu, J. and Tang, W. (2002), Ensembling Neural Networks: Many could be better than all, Artificial Intelligence*, Elsevier*, Vol. 137, pp.239-263.

[2006년 1월 접수, 2006년 3월 채택]