

오차 패턴 모델링을 이용한 Hybrid 데이터 마이닝 기법*

허 준** · 김종우***

A Hybrid Data Mining Technique Using Error Pattern Modeling*

Joon Hur** · Jong Woo Kim***

■ Abstract ■

This paper presents a new hybrid data mining technique using error pattern modeling to improve classification accuracy when the data type of a target variable is binary. The proposed method increases prediction accuracy by combining two different supervised learning methods. That is, the algorithm extracts a subset of training cases that are predicted inconsistently by both methods, and models error patterns from the cases. Based on the error pattern model, the predictions of two different methods are merged to generate final prediction. The proposed method has been tested using practical 10 data sets. The analysis results show that the performance of proposed method is superior to the existing methods such as artificial neural networks and decision tree induction.

Keyword : Supervised Learning, Hybrid Model, Combined Model, Voting, Error Pattern Modeling

논문접수일 : 2005년 1월 17일 논문게재확정일 : 2005년 9월 6일

* 이 논문은 2004년 한양대학교 일반연구비 지원으로 연구되었음.

** SPSS Korea (주)데이터솔루션

*** 한양대학교 경영대학 경영학부(교신저자)

1. 서 론

데이터 마이닝(Data Mining)의 가장 알려진 모델 중 하나는 지도학습(Supervised Learning)기법이다. 대표적으로 MLP(Multi-Layer Perceptron), RBF(Radial Basis Function) 알고리즘을 이용한 신경망 분석(Neural Networks)이나 C4.5, C5.0, CART, CHAID 알고리즘을 이용한 의사결정나무 분석(Decision Tree Induction)이 대표적인 지도학습의 분석 알고리즘이다. 지도학습 기법에는 이외에도 다른 분석 기법들이 존재하지만 위에서 열거한 분석 알고리즘들이 보편적으로 가장 많이 사용되거나 잘 알려진 기법이다. 이런 지도학습의 활용에 있어서 중요한 질문 중에 하나는 “예측을 또는 정확도가 가장 높은 기법이 무엇인가?”라는 것이다. 데이터의 유형에 따라서 또는 각종 상황에 따라서 가장 좋은 기법이 어떤 것인지에 대한 몇몇 연구된 사례가 있기는 하지만[23], 절대적으로 최선의 방법이 일반적으로 존재하지는 않는다. 즉, 신경망 분석 기법이든 의사결정나무 분석 기법이든 모두 자체적으로 장점과 단점을 가지고 있다. 이렇게 분석기법이 장점과 단점을 가지고 있다는 것은 다시 말하면 특정한 한 개의 기법만으로는 어느 일정한 수준 이상으로 예측 정확도를 높이는 데(또는 오차율을 줄이는 데) 한계가 있다는 것을 의미한다.

이런 단일 알고리즘의 한계를 극복하기 위한 것이 Hybrid 모델이다. Hybrid 모델은 서로 다른 알고리즘을 통합하여, 좀 더 나은 예측 정확도를 기대하는 방법이다. 즉, 한 개의 단일 알고리즘으로는 예측율(또는 정확도)을 향상시키는데 분명한 한계를 가지기 때문에 다른 접근을 하는 새로운 알고리즘을 이용하여 예측율을 높이겠다는 것이 Hybrid 모델의 목적이다. 또한 유사한 개념 중 Combined 모델이 있다. 이것은 일반적으로 데이터 마이닝의 지도학습에서는 데이터를 크게 훈련용(Training) 데이터와 시험용(Test) 데이터로 나누는데¹⁾ 이 중

훈련용(Training) 데이터의 데이터 분할, 가중치 조정 등을 통해 동일한 알고리즘을 다수 적용하는 기법을 의미하며, 앙상블(Ensemble) 기법이라고도 부른다. 이런 Hybrid 모델과 Combined 모델은 특히 오차율이 기법들 간의 독립적인 경우 보다 유용한 모델링 방법으로 알려져 있다[17].

본 논문에서는 목적 변수가 이진 변수(binary variable) 형태인 경우에, 예측율을 높이는 새로운 Hybrid 모델을 제시한다. 먼저 2장에서는 기존의 Hybrid 모델과 Combined 모델을 이용한 예측율과 정확도의 성능 개선과 관련된 연구들을 정리한다. 또한 Combined 모델 중 Voting 개념, Brieman이 제시한 Bagging[10], 그리고 Freund와 Schapire가 제시한 Boosting[14]방법과 이들의 관계에 대하여 설명한다. 다음 3장에서는 본 논문에서 제시하는 오차패턴 모델 기반의 Hybrid 알고리즘을 제시한다. 본 논문에서 제시하는 Hybrid 알고리즘에서는 훈련용 데이터 집합을 2개로 분리하고 서로 다른 데이터마이닝 기법을 통해 나온 결과의 오차 부분만 다시 추출하여 2차적으로 다시 데이터마이닝 기법을 이용하여 각각 사례별로 더 잘 맞추는 기법을 판별하고, 이에 대한 오차패턴 모델을 만들어 예측율을 높이는 새로운 Hybrid 알고리즘을 제안하고자 한다. 3장에서는 본 방법의 장점 및 시사점을 정리하고, 마지막 4장에서는 여러 가지 데이터 집합을 사례로 하여, 새로운 오차패턴 모델이 들어간 Hybrid 알고리즘의 예측력 및 정확도를 기존의 분석 기법과 비교하여 살펴보도록 한다. 그리고 5장에서는 결론을 제시한다.

2. 관련연구

2.1 Hybrid 기법 또는 Combined 방법을 이용한 성능 향상

지도학습 형태의 데이터 마이닝은 목적변수(Target Variable)가 있고, 정답을 가진 훈련용 데이터를 이용하여, 패턴을 인식한 다음 그것을 검정

1) 그 외 유효검정용(Validation)데이터 등 3가지로 나누는 경우도 있다.

용 또는 실제 예측할 데이터에 응용하여 그 예측력 및 정확도를 알아내는 과정을 의미한다. 이러한 지도학습 데이터 마이닝에는 신경망과 의사결정나무 분석(Decision Tree) 그리고 선형, 로지스틱 회귀분석 등이 대표적으로 사용된다. 그 외에도 많이 사용되는 데이터 마이닝 기법들로는 군집분석을 위한 K-평균 알고리즘과 SOM(Self-Organized Map), 그리고 연관성 분석 및 유전자 알고리즘(Genetic algorithm) 등이 있다.

성능향상을 위해서 앞에서 제시된 여러 가지 데이터 마이닝 방법을 혼합한 Hybrid 모델 및 Combined 모델도 개발되었다. 이런 Hybrid 모델 및 Combined 모델에 대한 연구로는 먼저 어떤 한 분석 기법을 이용하여, 주요한 설명 변수를 추출하거나, 1차적으로 가지치기를 한 다음 다른 분석을 이용하여, 최종 모델을 만들어내는 경우가 있다. 예를 들어, Coenen 등[12]은 C5.0 기법을 통해서 먼저 주요 분류자(Classifier)를 발견하고, 다음 사례기반 추론(Case-based reasoning) 기법을 이용하여 분석하는 Hybrid 모델을 제시하여, Direct Mail에서 응답률 향상 분석에 적용하였다. Carvalho 등[11]은 의사결정나무 기법(C4.5)을 이용하여, 데이터를 분리한 다음, 유전자 알고리즘(Genetic algorithm)을 이용하여 최종 규칙을 찾아내는 Hybrid 방법을 제시하였다. 또한 Renpu and Zheng-ou[21]은 1984년 Pawlak이 제안한 Rough set 이론을 통해서, 불확실한 데이터의 정보를 어느 정도 보완한 다음 신경망을 이용하여, 최종적인 분류 규칙의 효율을 향상시키는 방법을 제시하였다. 또한 이극노와 이홍철[5]은 먼저 의사결정나무 분석 기법인 C4.5를 통해서 주요한 설명변수를 도출해 내고, 다음 이를 이용하여 신경망 분석을 하는 방식으로 2개의 서로 다른 모델을 결합하여, 이를 이동통신 고객 분류에 적용하여 기존의 모델보다 더 좋은 성과를 나타냄을 보였다. 강문식, 이상용[1]은 경쟁학습 모델과 신경망의 Back-propagation 알고리즘을 Hybrid한 HACAB²⁾을 제안하였는데, 이 연구에서는 먼저 경쟁학습 모델에서 가중치를 도출해 내고, 이를 신경

망의 Back-propagation 알고리즘에 적용하여, 평균 8% 정도의 패턴 분류율의 향상을 이루었다.

앞에서 제시한 Hybrid 모델들이 주로 단계별로 2개 이상의 모델을 적용하여, 결과를 향상시키려는 연구들이라면, 2개의 모델 결과의 혼합을 통해서, 새로운 예측값을 만들어 효율을 높이는 Hybrid 모델에 대한 연구도 진행이 되었다. Lin and Salley[22]는 일반적인 다변량 통계분석 기법과 인공지능 기법을 결합하여 도산 가능성이 높은 기업을 예측하는 연구를 수행하였으며, Conversano 등[13]은 여러 개의 통계분석 방법(회귀분석, 판별분석, 비모수 통계방법, CART 등)을 통해서 나온 모수들을 결합하여 다시 새로운 혼합 모델(Mixture Model)을 만들어 성능을 향상시키는 연구를 수행하였다. 또한 Versace 등[28]은 금융 산업의 예측 모델에서 신경망과 유전자 알고리즘을 결합한 모델의 유용성에 대하여 연구를 수행하였다. 국내에서도 김진성[2]은 지도학습 기법인 퍼지 신경망과 비지도학습기법인 연관성 분석의 패턴을 이용한 Hybrid 메커니즘에 대한 연구를 수행하였다.

이와 같이 전혀 다른 2개의 방법을 결합하는 Hybrid 방법 이외에도, 하나의 방법을 다양하게 결합한 Combined 모델에 대한 연구도 있었다. Zhou 등[29]은 신경망 자체의 여러 결합모델이 왜 단독으로 사용된 신경망 모델보다 좋은 지를 분산의 편의(bias) 감소를 통해서 증명하였으며, Indurkha와 Sholom[19]은 의사결정나무 기법을 여러 번 재 샘플링(Sampling)한 다음 이를 결합하는 Voting 방법을 이용하여, 의사결정나무의 최종적인 노드들의 이익(Gain) 추정값의 향상을 연구하였다.

이러한 Hybrid 모델이나 Combined 모델이 기본적으로 단일 모델을 사용한 경우에 비해 예측율이 나 분류율이 어느 정도 향상되는지에 대한 연구도 진행되어왔다. Kuncheva 등[20]이 Hybrid 모델을 이용하여, 정확도가 향상된 사례를 연구하였으며,

2) Hybrid Algorithm Combining a Competition Learning Model and BP Algorithm.

Suh 등[27]은 기법 간에 서로 상관성이 낮은 경우의 Combined 모델이 더욱 더 성능 향상 정도가 크다는 실증 연구 결과를 RFM과 신경망, 로지스틱 회귀모델 등을 결합하여 연구한 바 있다. 이군희[4]는 전체적인 앙상블 방법에 대하여 모델의 성능 향상을 위해서 여러 모델의 의견을 종합하여, 최종 결론을 내리는 모델이 기존 모델보다 우수하다고 보편적으로 얘기할 수 있지만, 이론적인 근거는 아직 제시하지 못하고 있다고 언급하였다.

이 외에 Hybrid 모델과 Combined 모델에 대하여 다른 시각으로 접근한 연구로서 Hsu 등[18]과 같이 데이터에서 범주형 변수의 경우에는 연관성 분석을 한 다음 최적의 트리(Tree) 모델을 찾아내고, 연속형의 경우 바로 트리 모델을 이용하여 다음 단계에서 이들을 혼합한 Hybrid 모델로 학생들의 학습능력 향상에 적용한 연구를 수행하기도 하였다. 또한 Grzymala-Busse[16]와 같이 연속형 데이터를 군집분석 등을 이용하여, 이산화하여 규칙의 분류율을 높이는 또 다른 시각에서의 방법을 사용하기도 하였으며, 이재식, 이진천[6]은 본 논문에서 제안하고자 하는 모델과 유사하게 입력자료 판별에 의한 성능 개선 방안에 대하여 연구를 하였다. 이재식, 이진천[6]의 논문에서는 판별모델, 기본모델, 지원모델이라는 개념을 사용하여, 특정한 분석기법으로 특정 사례에 기본모델을 적용하는 것이 좋을지 지원모델을 적용하는 것이 좋을지를 판별하고, 판별한 결과에 따라 모델을 선택하여 적용하는 방법을 사용하여, 일반 모델의 결과보다 6~7%의 정확도 향상을 가져오는 결과를 제시하였다.

2.2 예측 로직과 Predictor

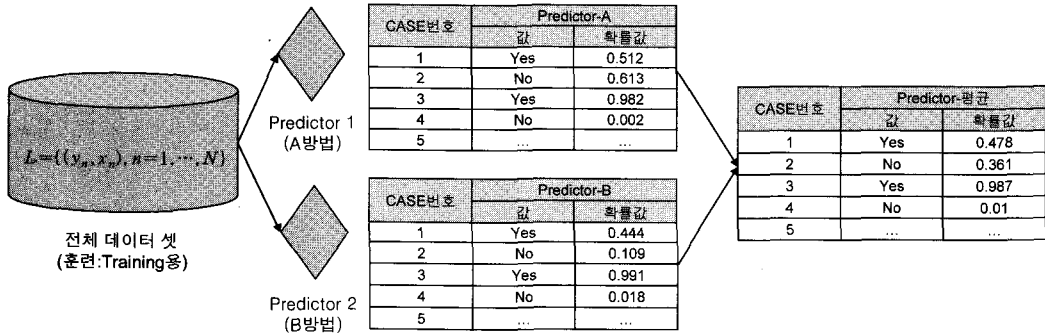
먼저 본 논문에서 사용할 주요 개념과 표기법을 제시하도록 한다. 본 논문에서 Predictor는 지도학습 형태의 데이터마이닝 기법에 의해서 생성된 예측 로직(logic)으로 정의한다. 즉, 데이터 마이닝의 지도학습 모델에서 먼저 훈련용 데이터 집합을 이용하여, 신경망 분석이나 의사결정나무 분석을 수

행하면 일종의 예측 로직이 나오게 된다. 예를 들어 설명 변수가 성별(sex)과 소득(income)이고, 목적 변수가 이탈유무(CHURN)인 경우 신경망 분석은 훈련용 데이터를 통해서 다음과 같은 형태의 예측 로직을 생성할 수 있을 것이다.

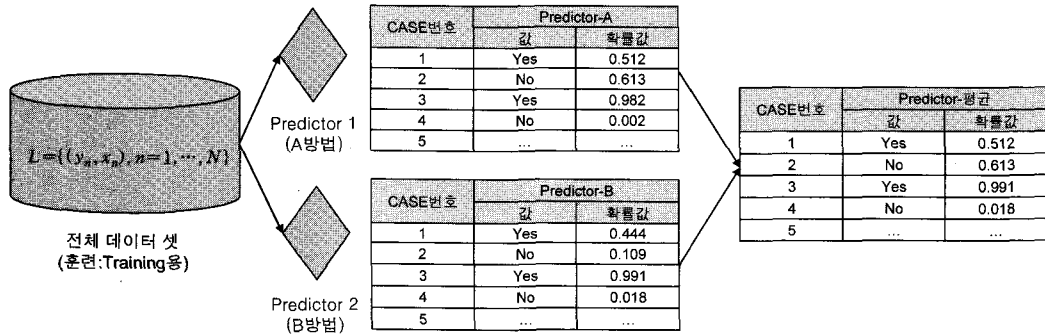
이탈유무 예상 확률값

$$= 0.2 \times \text{성별의 표준 변환값} + 0.3 \times \text{소득수준의 표준 변환값(계수 값은 임의로 정한 것으로 특별한 의미는 없음)}$$

이 경우 성별의 표준 변환값이 남자이면 0.5이고, 여자이면 1인 경우, 새로운 사례가 남성이면서 소득수준의 표준 변환 값이 2인 경우 이탈 예상 확률 값은 $0.2 \times 0.5 + 2 \times 0.3 = 0.1 + 0.6 = 0.7$, 즉 70%의 이탈 가능성을 가지게 된다. 이 이탈확률을 계산하기 위해서 사용된 위 식을 이 신경망 모델의 Predictor라고 한다. 또한 의사결정나무 분석의 경우 만들어진 예측 로직에서 첫 번째 규칙은 성별 “남자” 그리고 소득이 “100만원 이하”이면 “이탈”, 두 번째 규칙은 성별 “여자” 그리고 소득이 “500만원 이하”면 “유지” 등과 같은 여러 개의 규칙이 생성되었다고 하면, 새로운 사례가 오는 경우, 이 규칙들에 적용시켜 새로운 사례의 이탈 유무를 예측하게 된다. 이와 같이 지도학습기법에서, 예측값을 산출하기 위한 로직(logic)인 예측 로직을 본 논문에서는 Predictor라고 부르기로 하고, 보다 엄밀하게 다음과 같이 정의한다. Predictor ϕ 가 데이터마이닝 기법 A 와 훈련용 데이터 집합 L 을 사용하여 생성되었다면, Predictor ϕ 는 시험용 사례의 설명변수 벡터 집합에서 정의되고, 예측값과 이 값을 갖을 확률값의 튜플을 반환하는 함수로 정의되고, $\phi_A(\cdot|L)$ 로 표시한다. 즉 i 번째 사례의 설명변수 벡터 x_i 에 대한 Predictor값은 $\phi_A(x_i|L)$ 라고 표기되고, 이는 결과값은 $(P_i, \text{Prob}(P_i))$ 로 P_i 는 목적 변수에 대한 예측값이며, $\text{Prob}(P_i)$ 는 이 결과값이 나올 확률 값이다(단, P_i 는 binary이다.).



[그림 1] 두 기법의 확률 값의 평균을 이용한 Voting



[그림 2] 두 기법 중 더 확률 값이 더 높은 것을 선택하는 Voting

2.3 Combined 모형 중 Voting

Voting은 그 의미 자체가 곧 Hybrid 모델 또는 Combined 모형을 대표한다고 할 정도로 매우 보편적인 방법이다[9]. Voting의 기본 개념은 매우 간단하다. 예를 들어, 데이터 집합 L 에 대하여, A라는 데이터마이닝 알고리즘(또는 기법)을 통해서 나온 Predictor를 $\phi_A(\cdot|L)$ 라고 하고, B라는 데이터마이닝 알고리즘을 통해서 나온 Predictor를 $\phi_B(\cdot|L)$ 라고 한다면, 첫 번째 Voting 방법은 이들의 확률 값의 평균을 내서 최종 Predictor를 만드는 것이고([그림 1] 참조), 둘째는 A와 B방법의 Predictor 확률 값 중 더 큰 값을 선택하여 이를 조합하여 이것을 최종 Predictor로 사용하는 방법([그림 2] 참조)이다.

[그림 1]과 같이 평균을 이용한 Voting은 일반적으로 목적 변수가 연속형인 경우 많이 사용을 하

고, [그림 2]와 같이 더 좋은 예측 확률 값을 선택하는 경우는 주로 목적 변수가 범주형인 경우에 사용된다. 또한 위의 설명에서 데이터마이닝 기법 A방법과 B방법은 서로 다른 분석 기법(예를 들어 하나는 신경망, 다른 하나는 의사결정나무 분석)으로 의미하였는데, 다음의 Bagging과 Boosting의 경우에는 내부적으로 데이터를 분리하여 같은 분석 기법을 Voting 방법으로 결합하기도 하므로 반드시 A방법과 B방법이 완전히 다른 분석 알고리즘만을 의미하지는 않을 수도 있다.4)

3) [그림 1]의 확률값은 결과 값(Yes, No)에 대한 확률값을 의미함. 즉, (Yes, 0.444)인 것은 Yes가 될 확률이 44.4%라는 의미이다.
4) 실질적으로 Bagging과 Boosting에서는 일반적으로 의사결정나무 분석 기법(ex : C5.0, CART, CHAID) 등 한 가지 기법을 다양하게 나누어 Voting을 하는 경우가 대다수이다.

2.4 Combined 모형 중 Bagging(Bootstrap aggregating)과 Boosting

Bagging을 간단히 정의하면, 지도학습에서 주어진 훈련용 데이터를 복원추출(bootstrapping : resampling randomly with replacement)하여 여러 개의 훈련용 데이터 집합으로 만들고, 각각의 데이터 집합에 대하여 모델링을 하여 이를 결합하는 방법이라고 할 수 있다[10]. 물론 Bagging은, 이렇게 단순하게 사용하는 것이 아니라 Bootstrap 표본추출(Sampling)을 하면서, Poisson(1)분포를 요구하기 때문에[7], 이러한 점들을 모두 감안하여 표본 추출을 하게 되고, 또한 최종 결과인 Predictor를 통하여, 계산할 때 오분류 비용까지 감안하여 하는 경우가 있게 된다. Bagging에 대한 보다 자세한 내용은 여러 참조 논문이나 문헌을 참고하도록 한다. Brieman[10]은 CART 알고리즘에 $N=50$ 에 대한 Bagging을 적용하여서, 57%였던 정확도를 94%까지 끌어올렸다고 보고하고 있다.

Boosting은 지도학습 모델 등에서 학습이 잘 되지 않는 패턴에 가중치를 두어 단계별로 샘플링하고 이를 다음 단계에서 반영시켜, 최종적으로는 이렇게 표본 추출된 데이터를 다시 Voting 형태로 결합하는 기법이다. Boosting은 원래 데이터 집합으로부터 학습 데이터에 포함될 원 데이터의 패턴이 일부 특정 패턴으로 축소 또는 강조되는 방법이라고 할 수 있다[3]. 일반적으로 Boosting 기법은 C5.0(또는 C4.5), CART, CHAID 등 의사결정나무 분석 기법에서 주로 활용이 되어진다. Boosting의 탄생은 Schapire[25]가 매우 약한 학습 로직(weak learner)을 강하게 향상시키는 방법을 연구하면서 개발되었다. 현재 Boosting은 매우 연구가 활발하게 진행되는 분야이며[15], 최근에 가장 정리가 된 Boosting 방법은 Ada Boosting(**Adaptative Boosting**)이다[14]. 또한 현재 Boosting 방법을 다중 클래스(Multi-Class)문제에도 적용하기 위한 연구도 진척되고 있다[26]. Boosting의 핵심인 가중치를 두는 단계의 수는 일반적으로 특정하게 정해진 것은 없지만, R.

Quinlan이 Bagging, Boosting 그리고 C4.5를 비교한 논문에서는 10회를 기준으로 비교하여[24], 일반 데이터마이닝 패키지 등에서는 10회를 기본 설정으로 사용하고 있다[8]. 또한 Freund와 Schapire[14]는 오차율이 0.5보다 작다는 조건 하에서 일반적인 분류 모델을 통해서 얻은 예측력이 Boosting을 통해서 얻은 결과의 예측력보다 같거나 작다는 것을 증명하여, Boosting 방법의 우수성을 입증하기도 했다[4].

정리하면, Bagging과 Boosting 모두 궁극적으로 예측력 및 정확도의 향상을 위해서 하나의 데이터 집합으로부터 한 가지 로직만을 추출하는 것이 아닌 다양한 로직을 추출하여 이를 결합하여, 오분류된 부분을 보강하고, 이를 통한 예측력 및 정확도의 향상을 가져오는 방법이다. 하지만 둘 사이의 차이점은, Bagging은 임의추출을 통한 데이터 집합의 분리와 단순한 Voting 방법을 사용하고, Boosting은 단계적으로 데이터의 가중치를 재조정하고, 가중치를 통한 Voting 방법을 사용한다는 점이다.

2.5 Hybrid 모델과 Combined 모델 그리고 데이터 마이닝

앞서도 언급했다시피 Hybrid 모델과 Combined 모델의 차이는 일반적으로 전혀 다른 분석 기법을 융합하여 사용하는 경우 Hybrid 모델이라고 하고, 동일 기법 내에서 여러 개의 데이터 집합이나 가중치를 변경하여, 여러 개의 모델을 만들어 이를 결합하는 것을 Combined 모델이라고 정의할 수 있다. 두 모델간의 의미는 조금 다르지만 이 두 방법의 공통된 의미와 목적은 Michile and Taylor[23]가 주장한 바와 같이 현존하는 학습 알고리즘들의 경험적 비교결과 각각의 알고리즘은 어떤 선택적인 장점(selective superiority)을 가지고 있다는데서 찾을 수 있을 것이다. 즉, 오분류가 되는 사례에 대하여, 하나의 기법 내 다양한 로직(규칙)을 발견(Combined 모형의 사상)하거나, 완전히 다른 모델의 새로운 알고리즘이 결합하여(Hybrid 모델의 사상), 오

분류를 줄이고 정확도를 높이는 최종 데이터 마이닝 모델을 생성한다는 것이다.

3. 오차패턴 모델링을 이용한 Hybrid 기법

3.1 오차패턴 모델링을 이용한 Hybrid 모델의 개념

본 장에서는 제시하고자 하는 오차패턴 모델을 이용한 Hybrid 모델을 기술하도록 한다. 예를 들어, 지도학습 알고리즘을 이용하는 데이터 마이닝에서 어떤 분석 데이터가 10건이 있고, 어떤 A기법(예를 들어 의사결정나무 분석 기법)을 이용하여 모델을 만들었다고 가정하자. 그리고 그 모델을 시험용 데이터를 이용하여 검정한 결과 정확도가 70%(7건) 그리고 오분류 또는 예측을 잘못된 것이 30%(3건)였다고 가정하고, 다음 똑같은 데이터를 이용하여, 다른 모델링 기법인 B기법(예를 들어 신경망 분석 기법)을 이용하여 똑같이 시험용 데이터를 이용하여 검정한 결과 역시 정확도가 70%(7건) 그리고 예측을 잘못된 것이 30%(3건)이었다고 가정한다. 이렇게 단순한 요약 정보만 있다면, 이 2개의 모델은 동일한 성능을 가졌다고 판단할 수 있다. 하지만 다음 [그림 3]과 같은 경우를 생각해 본다.

[그림 3]에서 보면 A방법 예측과 B방법 예측은 정확도가 동일하게 70% 이지만 서로 동시에 틀린 부분도 있고(번호 8, 10번), A방법이 잘 맞춘 경우(번호 9번), B방법이 더 잘 맞춘 경우(번호 3번)도 있다. 즉, 이것을 다시 말하면 공통적으로 예측을 잘 하는 사례가 있는 반면, 공통적으로 못하는 부분도 있으며, 특정한 방법에 따라 잘 맞출 수 있는 사례와 못 맞추는 사례가 존재한다는 것이다. 이렇게 공통적으로 다 잘 맞춰줄 수 있는 부분을 제외하고 오류가 난 부분 중 각각의 방법이 잘 맞추는 경우만 가지고 오는 Hybrid 기법이 있다면, [그림 3]의 가장 우측의 결과처럼 정확도는 80%로 올라가게 될 것이다.

번호	실제값	A방법예측	B방법예측	Hybrid기법
1	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes
3	Yes	No	Yes	Yes
4	No	No	No	No
5	No	No	No	No
6	Yes	Yes	Yes	Yes
7	Yes	Yes	Yes	Yes
8	Yes	No	No	No
9	No	No	Yes	No
10	No	Yes	Yes	Yes

[그림 3] Hybrid 기법의 예측값 선택 과정

본 연구에서 제시하는 오차패턴 모델이란 서로 다른 2개 이상의 기법을 동일한 데이터에 적용하여, 2개 이상의 모델이 서로 다른 결과를 내는 경우만 추출하여, 데이터 집합을 구성하고, 이 데이터 집합을 가지고, 다시 A방법과 B방법이 잘 맞추는 오차패턴 모델을 생성한 다음, 실제 적용할 데이터 집합에서는 각 사례에 대하여 [그림 3]의 Hybrid 기법 결과와 같이 A방법과 B방법이 서로 잘 맞추는 사례를 맞추게 하여, 최종적으로는 오분류 및 잘못된 예측이 적은 Predictor를 만들어 내는 방법이다.

3.2 오차패턴 모델을 이용한 Hybrid 기법의 과정

구체적으로 오차패턴 모델링을 이용하여 Hybrid 모델을 만드는 과정을 정리하면 다음과 같다.

- (1) 전체 훈련용 데이터 집합을 $L = \{(y_n, x_n), n = 1, 2, \dots, N\}$ 이라고 한다. 여기서 y_n 은 목적 변수를 의미하고, x_n 은 설명 변수 벡터를 의미하며, N 은 데이터의 레코드 수를 의미한다. 또한 데이터 집합 L 에서 목적 변수값을 가진 컬럼을 T_L 로 표현한다. 그리고 훈련용 데이터를 통해서 나온 로직을 검정하기 위한 시험용(Test) 데이터를 $L_t = \{(y_t, x_t), t = 1, 2, \dots, T\}$ 라고 정의하도록 한다. 본 단계는 지도학습 기법에서 일반적으로 하는 것과 같이 전체 데이터를 훈련용 데이터와 시험용 데이터를 나누고, 이를 식으로 정리한 단계이다.

- (2) 전체 훈련용 데이터 집합을 임의 추출방법을 이용하여, 2개로 분리를 한다. 2개의 데이터 집합을 다음과 같이 정의한다.

$$L_1 = \{(y_m, x_m), m = 1, 2, \dots, M\},$$

$$L_2 = \{(y_p, x_p), p = 1, 2, \dots, P\}$$

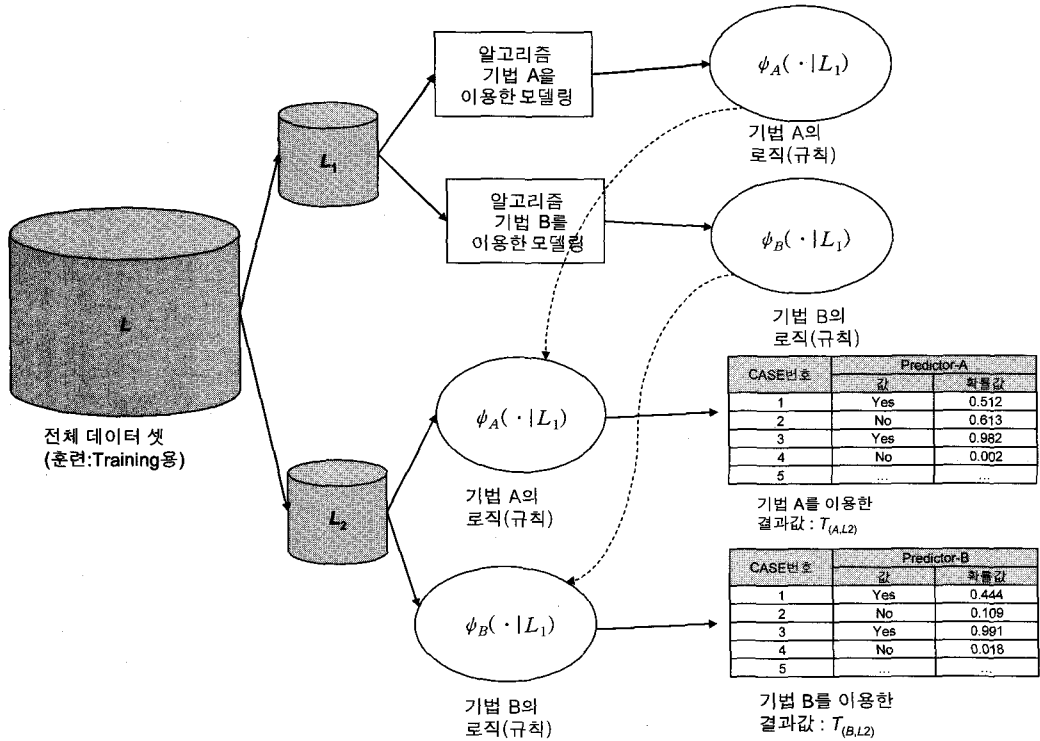
단, $M+P=N$

본 단계는 앞서 (1)에서 나눈 훈련용 데이터를 임의 추출을 통해 2개로 데이터 집합을 분리하는 단계로, 이는 내부적으로 또 하나의 모델인 오차 패턴 모델을 만들기 위하여, 훈련용 데이터를 다시 훈련용과 시험용 데이터로 분할을 하는 것이다.

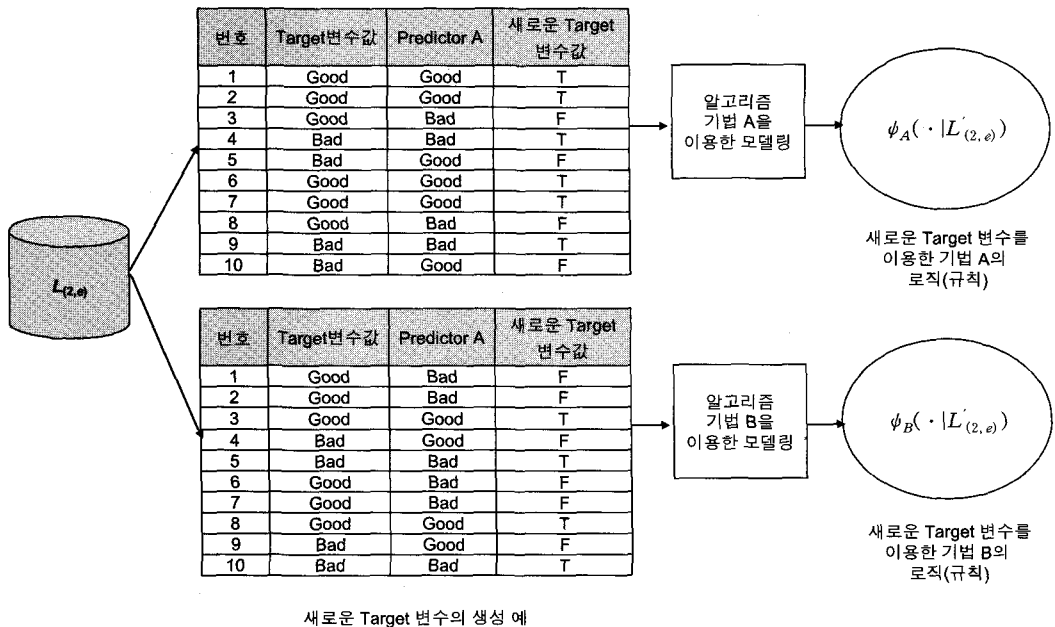
- (3) 먼저 L_1 데이터 집합을 이용하여, 분석 기법 A 를 이용하여, 모델링을 수행한다. 이 때 모델링을 통해서 생성한 기법 A 의 로직, 즉, Predictor를 $\phi_A(\cdot|L_1)$ 이라고 한다. 다음 동일한 데이터 집합에 A 와는 다른 분석 기법 B 를 이용하여 모델링을 수행한다. B 기법을 이용하여 모델링을 수행하여 나오게 되는 로직을 $\phi_B(\cdot|L_1)$ 이라고 한다. 본 단계는 (2) 단계에서 나누어진 2개의 훈련용 데이터 중 하나를 오차패턴 모델링을 만들기 위해 훈련용 데이터로 지정을 하고, 기법을 적용시킨 것이며, 이 단계를 통해서 나온 로직을, 남아 있는 다른 데이터 집합에 적용하여, 결과를 산출하는 것이 다음 단계이다.
- (4) 다음으로 단계 (2)에서 분리한 또 다른 훈련용 데이터 집합인 L_2 에 L_1 데이터를 이용하여 생성된 두 기법의 예측 로직인 Predictor $\phi_A(\cdot|L_1)$ 과 $\phi_B(\cdot|L_1)$ 를 적용시킨다. 먼저 기법 A 를 적용시켜서 나온 예측 결과(이 결과는 하나의 컬럼 형태가 될 것이다.)를 $T_{(A,L_2)}$ 라고 하자. 마찬가지로 기법 B 를 적용시켜서 나온 결과를 $T_{(B,L_2)}$ 이라고 하자. 지금까지의 과정을 그림으로 표현하면 다음 [그림 4]와 같다.
- (5) 데이터 집합 L_2 를 통해서 나온 2개의 결과 값을 서로 비교하여 결과 값이 서로 틀린 데이터

집합만을 추출한다. 이것은 [그림 3]에서 번호 3번, 9번 데이터만 추출하는 것과 동일하다고 할 수 있다. 이렇게 추출해 낸 데이터 집합을 $L_{(2,e)}$ 라고 정의한다. 즉, $L_{(2,e)} = \{(y_i, x_i) | (y_i, x_i) \in L_2 \text{ and } \phi_A(x_i|L_2) \neq \phi_B(x_i|L_2)\}$ 이다. 이렇게 추출된 데이터 집합이 바로 오차 패턴 모델링을 생성하기 위한 사례 데이터 집합 생성에 사용이 되며, 다음 단계는 해당 사례 데이터 집합의 새로운 목적변수를 생성하는 단계이다.

- (6) 다음 데이터 집합 $L_{(2,e)}$ 에서 $x_i \in L_{(2,e)}$ 의 목적 변수 y_i 값과 기법 A 를 이용하여, 생성된 Predictor $\phi_A(x_i|L_1)$ 에 의하여 나온 결과 값과 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적 변수를 생성한다. 이렇게 새롭게 파생된 목적 변수를 $T_{(A,L_{(2,e)})}$ 라고 한다. 다음 반대로 역시 기존의 목적 변수와 기법 B 를 이용하여, 생성된 Predictor $\phi_B(x_i|L_1)$ 의 결과 값과 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적 변수를 생성한다. 이렇게 새롭게 파생된 목적 변수를 $T_{(B,L_{(2,e)})}$ 라고 한다.
- (7) $L_{(2,e)}$ 데이터 집합에서 기존의 목적 변수 $T_{L_{(2,e)}}$ 대신에, 새롭게 만들어진 목적 변수를 대체하여, 이 데이터 집합을 $L'_{(2,e)}$ 라 하고, 이 데이터 집합에 기법 A, B 를 다시 적용한다. 즉, $L_{(2,e)}$ 데이터 집합에서 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$ 로 교체한 다음 다시 모델링 기법 A 를 다시 수행하고, 다시 한 번 역시 기존의 목적 변수 대신에 $T_{(B,L_{(2,e)})}$ 로 교체한 다음 다시 모델링 기법 B 를 수행한다. 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$ 로 해서 모델링 기법 A 를 수행한 후 발생하는 Predictor를 $\phi_A(\cdot|L'_{(2,e)})$ 라고 하고, 마찬가지로 $T_{(B,L_{(2,e)})}$ 를 목적 변수로 해서 모델링 기법 B 를 수행한 후 발생하는 로직을 $\phi_B(\cdot|L'_{(2,e)})$ 라고 한다. 이를 예시적으로 표시한 것이 [그림 5]이다.



[그림 5] 오차 패턴 모델에서 초기 데이터 분할과 두 가지 기법의 적용



새로운 Target 변수의 생성 예

[그림 6] 오차 패턴 모델링의 과정

<표 1> 시험용 데이터에서 오차패턴 모델의 판별 결과와 두 기법의 Predictor 예제

번호	$\psi_A(\cdot L_{(2,e)})$	$\psi_B(\cdot L_{(2,e)})$	$\psi_A(\cdot L_1)$		$\psi_B(\cdot L_1)$	
			값	확률값	값	확률값
1	T	T	Good	0.5	Good	0.4
2	T	F	Bad	0.6	Bad	0.5
3	F	T	Good	0.7	Bad	0.5
4	F	F	Bad	0.8	Bad	0.6

<표 2> 두 기법을 Voting한 결과

번호	$\psi_{(A,B)}(\cdot L)$		$\psi_{(B,A)}(\cdot L)$		$\psi_{(A,B)}(\cdot L)$	
	값	확률값	값	확률값	값	확률값
1	Good	0.5	Good	0.4	Good	0.5
2	Bad	0.6	Bad	0.6	Bad	0.6
3	Bad	0.5	Bad	0.5	Bad	0.5
4	Bad	0.6	Bad	0.8	Bad	0.8

이 단계에서 만들어진 로직 $\psi_A(\cdot | L_{(2,e)})$ 와 $\psi_B(\cdot | L_{(2,e)})$ 의 의미는 2개의 기법 A와 B가 서로 다른 결과를 낸 데이터만 모아둔 $L_{(2,e)}$ 데이터 집합에서, 기법 A와 B가 서로 잘 맞추는 형태의 데이터 패턴을 다시 파악하는 로직이라고 할 수 있으며, 본 Hybrid 오차패턴 모델의 핵심이라고 할 수 있다. 본 논문에서는, 이 로직 $\psi_A(\cdot | L_{(2,e)})$ 와 $\psi_B(\cdot | L_{(2,e)})$ 을 오차패턴 모델(Error Pattern Model) 또는 오차 모델(Error Model)이라고 정의한다.

(8) 다음 이렇게 오차패턴 모델(또는 오차 모델)을 구했으면, 이를 적용한 최종 Predictor를 생성하게 되는데 이 과정은 Voting 방법을 이용한다. 예를 들어서 시험용 데이터 집합인 $L_t = \{(y_t, x_t), t = 1, 2, \dots, T\}$ 에서 먼저 $\psi_A(\cdot | L_{(2,e)})$ 의 로직을 적용하여, 예측값이 T가 되는 사례는 $\psi_A(\cdot | L_1)$ 로직을 이용한 결과 값을 선택하고, 아닌 것은 $\psi_B(\cdot | L_1)$ 로직을 이용한 결과 값을 선택한다. 이렇게 판별하여 생성된 Predictor를 $\psi_{(A,B)}(\cdot | L)$ 라고 정의하고, 다음 반대로

$\psi_B(\cdot | L_{(2,e)})$ 의 로직을 적용하여, 예측값이 T가 나온 사례에는 $\psi_B(\cdot | L_1)$ 로직을 적용한 결과 값을 선택하고 F인 것은 $\psi_A(\cdot | L_1)$ 로직을 적용한 결과 값을 선택한다. 이렇게 조합을 통해서 나온 최종 Predictor를 $\psi_{(B,A)}(\cdot | L)$ 라고 한다. 이들이 각기 다를 수 있으므로 안정적인 Predictor를 생성하기 위해 이 2개의 계산된 Predictor들에게서 확률 값이 큰 쪽을 선택하여 만들어진 최종 Predictor인 $\psi_{(A,B)}(\cdot | L)$ 를 생성하게 되면, 모든 과정이 완료된다. 본 내용을 예제 데이터를 사용하여 표현하면 위의 <표 1>, <표 2>와 같다.

3.3 오차패턴 모델을 이용한 Hybrid 기법의 의미

오차패턴 모델을 이용한 Hybrid 기법을 만들어 내는 가장 큰 목적은 더 높은 예측 정확도를 위해서 다른 기법들 간에 자신의 기법이 더욱 효율적인 데이터 사례에만 해당 기법이 적용되도록 하는 것이다. 이를 위해 오분류 데이터에 대해 데이터 마이닝 모델링을 한 번 더 사용하여, 해당 기법이 잘 맞

추는 데이터는 어떤 데이터인가를 구분하는 패턴을 알아내는 모델링 과정이 추가가 된 것이다. 이 경우 일반적으로 2개의 기법보다는 조금이라도 데이터의 오분류 가능성이 낮아지고, 적어도 두 기법 중 더 나은 기법 정도의 정확도를 제공하는 구조여서, 상당히 안정성이 있다고 논리적으로는 설명할 수 있다. 그러나 그에 비해 모델링하는 단계가 복잡하고, 한 번 더 모델링을 수행함에 따라 수행시간이 더 경과되는 단점도 있을 것이다. 다음 장에서는 실제 여러 데이터 집합에 본 방법을 적용한 실험 결과에 대하여 설명한다.

4. 성능 비교를 위한 실험 결과

4.1 실험 가정

제시한 오차패턴 모델의 성능 비교를 위한 실험에서는 다음과 같은 몇 가지 가정을 두었다.

첫째, 실험 데이터의 목적 변수는 전부 이분형(Binary) 형태이다. 예를 들면 Yes/No, Response/No Response, T/F, Good/Bad 등의 형태이다.

둘째, 본 논문에서 정확도의 비교를 위하여 사용된 기법은 단일 알고리즘 중에서는 신경망의 MLP 방법과 의사결정나무 분석 기법의 C5.0 기법을 선택하였다. 그리고 비교를 위한 Hybrid/Combined 방법으로는 C5.0/신경망의 Voting 방법과, C5.0의 Boosting 방법을 사용하였다. 이들 Hybrid/Combined 방법 중 C5.0의 Boosting 방법은 데이터 마이닝 S/W에서 제공하는 Boosting 방법을 그대로 이용하였으며, Voting의 경우 가장 일반적인 방식인 C5.0과 신경망 중에서 예측 정확도의 확률값이 큰 것을 선택하여 최종적인 결과를 생성하는 방법을 이용하였다.

셋째, 비교되는 4개의 기법은 데이터에 따라 정확도 향상을 위한 각종 option을 변경하지 않는다. 이는 option 변경으로 인해 정확도가 향상되는 부분을 배제하기 위해서이다.

본 연구에서 MLP와 C5.0을 이용한 것은 신경망

과 의사결정나무 분석은 매우 이질적인 기법이므로 이를 통해서 Hybrid 모델의 사상을 잘 나타낼 것이기 때문이다. 본 실험의 수행을 위해서, 알고리즘을 자체적으로 프로그래밍하는 경우 그 신뢰성을 보장할 수 없기 때문에 전 세계적으로 광범위한 활용과 신뢰도를 가진 SPSS Inc.⁵⁾사의 데이터 마이닝 S/W인 클레멘타인 8.1(Clementine 8.1)을 이용하여 분석을 수행하였다.

4.2 실험 데이터의 설명

본 실험에서는 서로 다른 10가지 종류의 데이터 집합을 활용하였으며, 데이터 집합의 간단한 설명은 <표 3>과 같다.

4.3 실험 결과

<표 3>에서 나열한 10개 서로 다른 데이터 집합에 대하여 신경망, C5.0, C5.0의 Boosting, C5.0/신경망의 Voting, 그리고 오차패턴 모델을 이용한 Hybrid 모델을 수행하였을 때의 정확도와 그에 관련된 데이터들을 정리한 것이 <표 4>이다.

<표 4>의 결과를 보면 오차패턴 모델을 이용한 Hybrid 모델의 정확도가 신경망 단독 정확도, C5.0(일반, Boosting) 단독 정확도, Voting보다는 전체적으로 높다는 것을 알 수 있다. 이에 대한 계량적 검증을 위하여, 10개 사례 데이터의 최종 정확도를 이용하여, 대응 이집단 비율 검정(paired two sample test for proportions)을 수행하였다. 즉, Hybrid 모델과 <표 4>에서 비교한 4개의 기법을 각각 대응 검정하였으며, 데이터의 수가 10개로 매우 작으므로, 분포 등과 무관하게 검정할 수 있도록, 2개의 대응 집단 비율 검정방법은 비모수적인 방법인 Wilcoxon의 부호순위 검정을 이용하였다. 검정 결과는 다음 <표 5>와 같다.

결과 <표 5>를 보면, 검정결과가 전부 유의수준

5) SPSS Inc.(<http://www.spss.com>), SPSS Korea(<http://www.spss.co.kr>).

〈표 3〉 실험에 사용된 데이터 집합의 설명

데이터 번호	설명변수 설명	목적 변수	비고
1	연속형 데이터 : 가입기간, 3개월 납부금액, 누적 납부금액, 연체금액 등 13개 필드 범주형 데이터 : 지역, 상품명, 가입, 기업유형 등 5개 필드	가입/해지 여부	실제 유선 및 전용선통신사의 기업고객의 전용망 가입 여부 데이터
2	연속형 데이터 : 연령, 교육기간, 금융소득액, 금융대출액 등 6개 필드 범주형 데이터 : 인종, 결혼상태, 성별, 인접국가, 직업 등 8개 필드	소득 5만달러 이상/ 5만달러 미만	소득 예상 계층을 분류하는 데이터
3	연속형 데이터 : 연령, 개인평균소득, 자녀수 등 3개 필드 범주형 데이터 : 차량 소유여부, 거주지역, 담보여부, 저축여부 등 7개 필드	반응/비반응	미국의 어느 잡지 구독 고객을 대상으로 메일 캠페인에 대한 반응 여부 데이터
4	연속형 데이터 : 연령, 부가서비스요금액, 사용일수 등 3개 필드 범주형 데이터 : 직업구분, 성별, 납입종류, 지역 등 6개 필드	이탈고객/유지고객	실제 이동 통신사의 이탈, 유지 관련 데이터
5	연속형 데이터 : 시외통화시간, 국제전화요금, 시내통화시간 등 7개 필드 범주형 데이터 : 지불방법, 청구형태, 성별, 결혼상태 등 6개 필드	이탈여부(break-out/ keep-up)	실제 영국 유선통신사의 개인전화 가입과 이탈 여부 데이터
6	연속형 데이터 : 주계약료, 특약료 등 10개 필드 범주형 데이터 : 고객등급, 상품코드 등 3개 필드	종신보험가입/ 비가입	한국의 생명보험회사에서 종신보험에 가입여부 데이터
7	연속형 데이터 : 주식선물옵션 3개월 평균, 주식, 선물옵션 누적약정금액 등 11개 필드 범주형 데이터 : 성별, 연령대 등 3개 필드	우수/비우수	한국의 한 증권사에서 고객의 우수등급과 비우수 등급을 분류한 데이터
8	연속형 데이터 : 보석구매여부, 식품구매여부 등 14개 필드 범주형 데이터 : 주문금액, 최근구매기간, 주문횟수 9개 필드	전기전자 제품의 구매여부 (구매/비구매)	한국의 한 홈쇼핑 업체의 자료
9	연속형 데이터 : 식료품 구매금액, 의류 구매금액 등 10개 필드 범주형 데이터 : 식품구매 여부 등 10개 필드	제품의 구매여부 (구매/비구매)	영국의 한 유통점의 특정 상품 구매여부
10	연속형 데이터 : 단선횟수, 주말통화시간 등 31개 필드 범주형 데이터 : 성별, 요금제 등 6개 필드	Churn(유선가입에 대한 이탈/유지)	미국의 한 유선 통신회사의 이탈/유지 여부 데이터

0.05(5%)보다 작은 것을 알 수 있다. 이는 전체적으로 오차 패턴을 이용한 Hybrid 모형이 신경망이나 C5.0 단독모형 그리고 일반적으로 잘 알려진 Voting 방법이나 Boosting 방법보다 효율적임을 나타내고 있다. 그러나 <표 4>에서 개개별의 정확도를 살펴보면 편차의 정도는 상당히 많이 나는 것도 알 수가 있다. 예를 들어 9번과 8번 데이터들의 경우에는 기존의 신경망이나 C5.0 보다 10~20% 이상의 정확도가 향상된 반면, 2번, 3번 데이터들의 경우 기존의 신경망이나 C5.0 그리고 Voting 방법보다 0~0.3% 정확도 편차를 나타내어, 정확도의 향상이 거

의 이루어지지 않은 것을 볼 수 있으며, 심지어 6번 데이터의 경우 매우 미미하기는 하지만 C5.0의 Boosting 방법이 제안하는 Hybrid 모델보다 좋은 예측 결과를 나타내고 있다. 이렇게 데이터마다 편차가 발생하는 원인을 파악하기 위하여, 각 데이터별 불일치비율, 변수의 종류 차이 비율, 그리고 기법과 정확도 간의 편차를 하나의 표로 정리한 것이 <표 6>이다.

<표 6>에서 보면, 특히 8번과 9번 데이터의 경우 다른 데이터와 비교하여, 상당히 높은 정확도의 향상을 보여주고 있다. 이 때 불일치 비율 역

〈표 4〉 실험 결과

데이터 번호	전체 데이터수	시험용 데이터수	C5.0/신경망 불일치 비율	신경망 단독정확도	C5.0 단독정확도 (BOOSTING G)	C5.0 단독정확도 (일반)	VOTING	오차패턴 모델을 이용한 Hybrid 모델 정확도
1	14,490	4,765	1,500 (31.48%)	68.96% T : 3,286 F : 1,479	76.24% T : 3,633 F : 1,132	73.33% T : 3,494 F : 1,271	75.30% T : 3,588 F : 1,177	77.73% True : 3,704 False : 1,061
2	32,561	10,750	1,017 (9.46%)	83.23% T : 8,947 F : 1,803	85.45% T : 9,186 F : 1,564	85.45% T : 9,186 F : 1,564	85.50% T : 9,191 F : 1,559	85.67% True : 9,209 False : 1,541
3	12,000	3,995	227 (5.68%)	91.34% T : 3,649 F : 346	93.97% T : 3,754 F : 241	94.67% T : 3,782 F : 213	94.42% T : 3,772 F : 223	94.67% True : 3,782 False : 213
4	3,284	1,079	258 (23.9%)	66.73% T : 720 F : 359	68.58% T : 740 F : 339	67.10% T : 724 F : 355	68.49% T : 739 F : 340	72.01% True : 777 False : 302
5	4,574	1,539	105 (6.82%)	76.09% T : 1,171 F : 368	76.60% T : 1,179 F : 360	76.54% T : 1,178 F : 361	76.41% T : 1,176 F : 363	76.60% True : 1,179 False : 360
6	100,000	33,013	1,314 (3.98%)	95.14% T : 31,409 F : 1,604	96.60% T : 31,890 F : 1,123	96.01% T : 31,697 F : 1,316	96.18% T : 31,751 F : 1,262	96.22% True : 31,765 False : 1,248
7	228,774	76,243	15,194 (19.93%)	80.02% T : 61,003 F : 15,231	98.87% T : 75,371 F : 863	98.87% T : 75,371 F : 863	98.89% T : 75,384 F : 850	98.90% True : 75,398 False : 836
8	2,181	754	236 (31.29%)	61.01% T : 460 F : 294	63.53% T : 479 F : 275	63.40% T : 478 F : 276	63.40% T : 478 F : 276	71.49% True : 539 False : 215
9	30,000	10,000	4,853 (48.53%)	70.30% T : 7,030 F : 2,970	83.95% T : 8,395 F : 1,605	71.65% T : 7,165 F : 2,835	71.46% T : 7,146 F : 2,854	91.10% True : 9,110 False : 890
10	16,004	5,396	296 (5.48%)	90.77% T : 4,898 F : 498	92.98% T : 5,017 F : 379	93.48% T : 5,044 F : 352	93.50% T : 5,045 F : 351	93.64% True : 5,053 False : 343

〈표 5〉 기법 간 정확도에 대한 대응 이집단 비율 분석결과

대응집단	신경망-Hybrid	C5.0(Boosting)-Hybrid	C5.0(일반)-Hybrid	Voting-Hybrid
근사유의확률	0.005	0.021	0.008	0.005

시 31.29%, 48.53%로 다른 데이터 집합의 경우보다 불일치 비율이 높은 것을 알 수 있다. 반대로 변수의 불일치 비율이 낮은 2, 3, 5번과 같은 데이터의 경우 편차가 거의 나타나지 않는 것을 알 수 있다. 또한 약하기는 하지만 변수의 종류 차이 즉, 범주형과 연속형의 데이터가 균형을 이루는 경우

(비율이 작을 수록 설명변수 종류의 비율적인 균형을 이룬다고 할 수 있다) Hybrid 모형과 다른 기법과의 전체적인 정확도 편차가 나타나는 것을 알 수가 있다. 이런 관계를 좀 더 명확하게 알아보기 위하여 여러 지표들과의 상관분석(Correlation Analysis)을 수행하면 다음 <표 7>과 같다.

<표 6> 데이터별 불일치 비율과 예측 정확도 편차

데이터 번호	불일치 비율	변수 종류 차이 비율 [ABS(연속형변수수 - 범주형변수수)/전체변수수]*100	Hybrid 모형과 최저예측 기법과의 정확도 편차 (A)	Hybrid 모형과 최고예측 기법과의 정확도 편차 (B)	4개 기법의 편차의 평균 (C)
1	31.48%	44.44%	8.77%	1.49%	4.27%
2	9.46%	14.29%	2.44%	0.17%	0.76%
3	5.68%	40.00%	3.33%	0.00%	1.07%
4	23.9%	33.33%	5.28%	3.43%	4.29%
5	6.82%	7.69%	0.51%	0.00%	0.19%
6	3.98%	53.85%	1.08%	-0.38%	0.43%
7	19.93%	57.14%	18.88%	0.01%	4.74%
8	31.29%	21.74%	10.47%	7.96%	8.66%
9	48.53%	0.00%	20.80%	7.15%	16.76%
10	5.48%	67.57%	2.87%	0.14%	0.96%

<표 7> 불일치 비율과의 상관분석 결과

상관분석 수행 구분	상관계수 값
불일치 비율 vs Hybrid 모형과 최저예측 기법과의 편차 상관분석(A)	0.803
불일치 비율 vs Hybrid 모형과 최고예측 기법과의 편차 상관분석(B)	0.825
불일치 비율 vs 4개 기법의 평균 편차(C)	0.935
변수의 종류차 비율 vs 4개 기법의 평균 편차(C)	-0.474
데이터 수와 vs 4개 기법의 평균 편차	-0.152
데이터 변수의 수 vs 4개 기법의 평균 편차	0.334

<표 7>을 보면 본 논문에서 제안하는 오차 패턴 모델을 이용하는 Hybrid 모형의 경우, 단독기법으로 사용했을 경우의 양 기법이 맞출 수 있는 사례가 서로 다른 불일치 비율이 높은 경우에 상관계수가 0.8~0.9 정도로 매우 높아 불일치 비율을 분석 전에 살펴보는 경우에 본 논문에서 제안 Hybrid 방법이 다른 분석 기법보다 좀 더 효율적이라는 것을 알 수가 있으며, 또한 좀 더 많은 경우에서의 연구가 더 필요하지만 연속형이나 범주형의 변수 분포가 균형을 이루는 경우에도 음의 상관성이 0.5에 가까우므로, 설명 변수들이 다양한 종류의 데이터 형

을 가지고 있는 경우 본 오차 패턴 모델링을 이용한 Hybrid 모형이 의미가 있을 수 있다는 것을 보여 주고 있다. 따라서 분석하고자 하는 데이터의 설명변수들이 범주형, 연속형 그리고 이분형 등 여러 형태를 균형있게 포함하는 경우에도 제안하는 오차 패턴 모델링을 활용하는 것을 고려해 볼만 하다. 그 외에, 전체 데이터 수의 많고 적음은 모형의 정확도를 향상시키는데 거의 상관성이 나타나지 않은 것을 알 수 있으며, 데이터의 변수 수가 많은 경우에도 상관성이 작은 것을 알 수가 있다. 향후 이에 대한 연구는 더 많은 사례 및 Simulation 등을 통해서 더 많은 검증 연구가 필요할 것으로 보인다.

5. 결 론

본 연구에서는 데이터 마이닝의 지도학습 모델에서 Hybrid 모델 및 Combined 모델을 이용하는데 있어 기존의 Voting이나 Bagging, Boosting 이외에, 오차 패턴 모델(Error Pattern Modeling)을 이용한 새로운 Hybrid 모델을 제시하고, 제시한 Hybrid 모델이 기존의 단일 분석 알고리즘과 일반적인 Hybrid 방법을 사용하여 분석(모델링)했을 때 보다 얼마만큼 더 많은 정확도의 향상을 가지고 오

는지에 대하여 10개의 데이터 집합을 사용하여 분석하였다. 분석 결과를 요약하면, 전체적으로 본 모델의 정확도는 최소한 같거나 향상을 가져오며, 특히 기법 간의 사례의 정확도 불일치 비율이 높고, 데이터에서 설명 변수의 유형이 범주형과 연속형이 균형을 가지고 있는 경우 향상도가 높은 것으로 나타났다. 따라서 분석하고자 하는 데이터를 탐색해서 위와 같은 조건을 만족하는 데이터라면 본 모델을 사용하여 좀 더 높은 정확도를 가진 분석을 할 수 있을 것으로 사료된다. 또한 모든 모델에는 장단점이 있을 수 있는데, 본 모델을 유용하게 사용할 수 있는 경우로는, 아주 조금이라도 정확도가 더 높은 모델을 필요로 하거나 의미가 있는 경우(예를 들면 신용 카드社 등에서 카드 사기적발과 같이 1건의 피해라도 매우 큰 영향을 주는 경우와 같은) 나 의료 데이터와 같이 데이터를 수집하기 어려워서, 양과 질적으로 부족한 데이터를 이용하여, 데이터의 정확도 향상을 피하고자 하는 경우에 적용하면, 좋은 성과를 올릴 수 있을 것으로 판단된다. 그러나 기법 간의 불일치 비율이 아주 낮거나, 특정한 하나의 기법 정확도가 유난히 높은 경우 등에서는 본 모델의 효과가 기존 모델에 비해서 크게 향상되지 않는 것으로도 나타났다. 이런 경우에는 모델의 수행 시간과 시스템 자원의 절감 차원에서 본 모델을 사용하지 않는 것이 더 나을 수도 있을 것이다. 그러나 지금과 같이 각종 S/W 및 H/W의 성능 향상이 나날이 이루어져, 물리적인 시간을 줄일 수 있는 경우에는 본 모델이 정확도의 향상을 위해서 사용되어질 수도 있을 것이다.

추후 연구방향은 다음과 같다. 본 연구에서는 신경망과 C5.0만을 한정하여 실험하였으나, 타 지도학습 모델들인 로지스틱 회귀분석이나 CHAID, CART 등을 포함하여 검토하는 것이 필요하다. 또한 목적 변수를 이분형(Binary) 데이터로 한정하였는데, 범주가 3개 이상인 범주형 데이터와 연속형으로 확장하는 노력이 필요하다. 아울러서 어떠한 특성을 갖는 데이터 집합에서 더욱 향상도가 높아지는 지에 대한 추가적인 연구 및 다양한 사례들을

이용한 지속적인 검증도 필요하다.

참고 문헌

- [1] 강문식, 이상용, "데이터 마이닝을 위한 경쟁학습모델과 BP알고리즘을 결합한 하이브리드 신경망", 「정보기술과 데이터베이스 저널」, 제9권 2호(2002), pp.1-16.
- [2] 김진성, "연관규칙과 퍼지 인공지능망에 기반한 하이브리드 데이터 마이닝 매커니즘에 대한 연구", 「한국경영과학회/대한산업공학회 2003 춘계 공동학술대회 논문집」, (2003), pp.884-888.
- [3] 신현정, "양상블 학습알고리즘의 일반화 성능비교 : OLA, Bagging, Boosting", 「정보과학회논문지」, 제97호(2000), pp.226-228.
- [4] 이근희, "모형평가와 앙상블을 이용한 데이터 마이닝에 관한 연구", 「서강경영논총」, 제9권 (1998), pp.293-306.
- [5] 이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘 연구", 「한국지능정보시스템학회지」, 제9권, 제1호 (2003), pp.139-155.
- [6] 이재식, 이진천, "입력자료 편비에 의한 데이터 마이닝 성능개선", 「한국지능정보학회학술대회」, (2000), pp.293-303.
- [7] 허명희, "Clementine Stream Prototypes : Part 2", 「SPSS Korea White paper」, (2004), pp.1-7.
- [8] _____, 「Clementine Ver. 8 User's Guide」, SPSS Inc, 2003.
- [9] Ali, K. and M. Pazzani, "Error Reduction through Learning Multiple Descriptions," *Machine Learning*, Vol.24, No.1(1996), pp. 105-112.
- [10] Brieman, L., "Bagging Predictors," *Machine Learning*, Vol.24, No.2(1996), pp.123-140.
- [11] Carvalho, D.R. and A.A. Freitas "Hybrid Decision Tree/Genetic Algorithm Method for Data Mining," *Information Sciences*,

- Vol.163, No.1/3(2004), pp.13-35.
- [12] Coenen, F.G., K.V. Swinnen and G. Wets
 “The Improvement of Response Modeling :
 Combining Rule-induction and Case-based
 Reasoning,” *Expert Systems with Applica-
 tion*, Vol.18, No.4(2000), pp.307-313.
- [13] Conversano, C., R. Siciliano and F. Mola,
 “Generalized Additive Multi-mixture Model
 for Data Mining,” *Computational Statistics &
 Data Analysis*, Vol.38, No.4(2002), pp.487-500.
- [14] Freund, Y. and R.E. Schapire, “Experiments
 with a New Boosting Algorithm,” *Proceed-
 ings of 13th International Conference on
 Machine Learning*, Morgan Kaufmann(1996),
 pp.148-156.
- [15] Gama, Joao Maguel Portela da, *Combining
 Classification Algorithms*, Departamento de
 Ciecia de Computadores Faculdade de Ciecias
 da Universidade do Porto, 1999.
- [16] Grzymala-Busse, J.W., “A Comparison of
 Three Strategies to Rule Induction from
 Data with Numerical Attributes,” *Electro-
 nic Notes in Theoretical Computer Science*,
 Vol.82, No.4(2003), pp.1-9.
- [17] Hansen, L.K. and P. Salaman, “Neural
 Networks Ensembles,” *Transactions on
 Pattern Analysis and Machine Intelligence*,
 Vol.12, No.10(1990), pp.993-1001.
- [18] Hsu, P.L., R. Lai, C.C. Chui, and C.I. Hsu,
 “The Hybrid of Association Rule Algor-
 ithms and Genetic Algorithm for Tree
 Induction : An Example of Predicting the
 Student Course Performance,” *Expert Sys-
 tems with Application*, Vol.25, No.1(2003),
 pp.51-62.
- [19] Indurkha, N. and S.M. Weiss, “Estimating
 Performance Gains for Voted Decision
 Trees,” *Intelligent Data Analysis*, Vol.2,
 No.1/4(1998), pp.303-310.
- [20] Kuncheva, L.I.C. Bezdek, and M.A. Shutton,
 “On Combining Multiple Classifiers by
 Fuzzy Templates,” *International Conference
 on Artificial Neural Networks IEEE*, (1998)
 pp.193-197.
- [21] Li, R. and Z.-O. Wang, “Mining Classifica-
 tion Rules Using Rough Sets and Neural
 Networks,” *European Journal of Opera-
 tional Research*, Vol.157, No.2(2004), pp.
 439-448.
- [22] Lin, F.Y. and S. McClean, “A Data Mining
 Approach to the Prediction of Corporate
 Failure,” *Knowledge-Based Systems*, Vol.
 14, No.3/4(2001), pp.189-195.
- [23] Michie D., D.J. Spiegelhalter, and C. Taylor,
*Machine Learning, Neural and Statistical
 Classification*, Ellis Horwood, 1994.
- [24] Quinlan, R., “Bagging, Boosting and C4.5,”
*Procs. 13th American Association for Arti-
 ficial Intelligence*, AAAI Press, 1996.
- [25] Schapire, R., “The Strength of Weak Learn-
 erbility,” *Machine Learning*, Vol.5, No.2
 (1990), pp.197-227.
- [26] Schapire, R., Y. Freund, P. Bartlett, and
 W.S. Lee, “Boosting the Margin : A New
 Explanation for the Effectiveness of Voting
 Methods,” *Proceedings of the 14th Interna-
 tional Conference on Machine Learning*,
 Morgan Kaufmann, (1998), pp.322-330.
- [27] Suh, E.H., K.C. Noh and C.K. Suh, “Customer
 List Segmentation Using the Combined Re-
 sponse Model,” *Expert Systems with Appli-
 cation*, Vol.17, No.2(1999), pp.89-97.
- [28] Versace, M., R. Bhatt, O. Hinds and M.
 Shiffer, “Predicting the Exchange Traded
 Fund DIA with a Combination of Genetic
 Algorithm and Neural Networks,” *Expert*

Systems with Application, Vol.27, No.3(2004), pp.417-425.
[29] Zhou, Z.-H., J. Wu and W. Tang, "Ensem-

bling Neural Networks : Many Could Be Better Than All," *Artificial Intelligence*, Vol.137, No.1/2(2002), pp.239-263.