혼합 데이터 마이닝 기법인 불일치 패턴 모델의 특성 연구

허 준* · 김 종 우**

Characteristics on Inconsistency Pattern Modeling as Hybrid Data Mining Techniques

Joon Hur* · Jong Woo Kim**

Abstract

PM (Inconsistency Pattern Modeling) is a hybrid supervised learning technique using the inconsistence pattern of input variables in mining data sets. The IPM tries to improve prediction accuracy by combining more than two different supervised learning methods. The previous related studies have shown that the IPM was superior to the single usage of an existing supervised learning methods such as neural networks, decision tree induction, logistic regression and so on, and it was also superior to the existing combined model methods such as Bagging, Boosting, and Stacking.

The objectives of this paper is explore the characteristics of the IPM. To understand characteristics of the IPM, three experiments were performed. In these experiments, there are high performance improvements when the prediction inconsistency ratio between two different supervised learning techniques is high and the distance among supervised learning methods on MDS (Multi-Dimensional Scaling) map is long.

Keywords : Inconsistency Pattern Model, Supervised Learning, Data Mining, Hybrid Model, Combined Model

논문접수일 : 2008년 03월 06일 논문게재확정일 : 2008년 03월 14일

* 주저자, SPSS Korea (주)데이타솔루션, e-mail : hoh@spss.co.kr
** 교신저자, 한양대학교 경영대학 경영학부 교수, (133-791) 서울시 성동구 행당동 17, Tel : 02-2220-1067, Fax : 02-2220-1169
e-mail : kjw@hanyang.ac.kr

1. 서 론

1.1 논문의 배경

현재 많은 데이터 분석 및 데이터로부터 유용한 패턴 발견에 활용이 되는 데이터 마이닝[Fayyad, 1996]은 초기에 데이터 마이닝이 시작된 이래 현재까지 미래 지향적인 방향으로 진화와 발전을 해오고 있다[Han and Kamber, 2001]. 데이터 마이닝에서 가장 일반적으로 활용되는 지도학습 기법(Supervised Learning) 또한 많은 발전을 해왔다고 할 수 있다. 특히 지도학습 기법에서 가장 많이 발전되고, 일반적으로 관심이 있는 사항은 기법을 통한 사례들(cases)의 정확도 또는 예측력의 향상이라고 할 수 있다. 이러한 예측 및 정확도의 향상을 위해서 다양한 방법을 시도하고 있는데, 그 중 가장 대표적인 방법이 여러 기존 기법들을 다양하게 결합이나 혼합시킨 결합(Combined) 모델과 혼합(Hybrid) 모델이다[허준, 김종우, 2005]. 이 중 결합 기법을 앙상블(ensemble) 기법이라고도 한다. 대표적인 결합 모델 및 혼합 모델 방법으로는 Bagging [Breiman, 1996], Boosting[Freund and Schapire, 1996] 그리고 Stacking[Wolpert, 1992]등이 있으며, 최근에도 Random Forests[Breiman, 2001] 등과 같은 새로운 결합 기법들이 소개 되고 있다. 다음 제 2장의 관련 연구에서도 언급을 하겠지만, Bagging이나 Boosting과 같은 혼합이나 결합 기법을 사용하는 가장 큰 이유는 특히 지도학습 기법에서, 기존 단일 기법만을 이용한 것보다 더욱 더 높은 예측력과 분류 정확도를 보이기 때문이다. 일반적으로 다양한 결합 및 혼합 기법들은 기존 단일 기법보다 더욱 성능이 향상된다고 알려져 있다[이군희, 1998]. 그리고 최근에 허준과 김종우[2005, 2007]가 제안한 불일치 패턴 모델(또는 오차패턴 모델) 역시 다양한 실험을 통해서, 기존의 단일 지도학습 기법은 물론 Bagging, Boosting, 그리고 Stacking 등과 같은 기존 결합 방법보다 더욱 성능이 우수하다고 밝히고 있다. 이렇게 많은 연구에서 결합 모델 및 혼합 모델의 성능이 우수하다는 결론을 도출할 때, 다른 관점에서 가장 중요한 것은 과연 어떤 경우에 혼합 및 결합 모델 기법의 성능이 더 잘 발휘되고, 어떤 기법들을 혼합할 때 더욱 더 성능을 높일 수 있는지에 대하여 정확한 특성을 파악 하는 것이라고 할 수 있다.

실제 데이터 마이닝, 좀 더 좁혀서 말하면 지도학습 기법이 적용되는 수많은 상황들은 무수한 경우의 수를 가지고 있는 사회 및 기업 등에서 나오는 데이터에 기반을 하고 있다. 이는 다시 말해서, 몇 몇 연구에서 나온 성능의 우수성을 입증한 결과만을 가지고 수학공식과 같이 실제 상황에 모든 것을 완벽한 대입을 할 수 없다는 뜻이기도 하다. 따라서 이와 같은 각종 기법의 특성에 대한 연구 또는 원인에 대한 해답을 찾는 연구야 말로, 제안되어지는 다양한 데이터 마이닝의 결합 모델 및 혼합 모델을 더욱 더 효율적으로 활용하는데, 큰 도움을 줄 것이며, 이것이 곧 본 논문의 연구 배경이라고 할 수 있다.

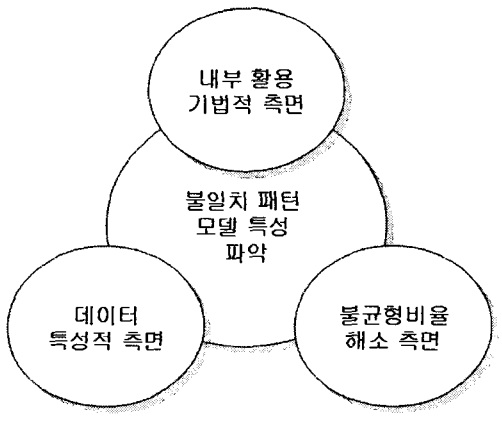
1.2 연구 목적과 논문 구성

본 논문의 목적은 허준과 김종우[2005, 2007]가 제안한 불일치 패턴 모델(또는 오차패턴 모델, 이후 불일치 패턴 모델로 통일하여 사용함)의 특성을 파악하여, 기존 연구에서 이미 증명된 예측력 및 분류 정확도의 향상을 어떠한 경우에 가장 효율적으로 적용할 수 있는지 알고자 하는 것이다.

기존 연구에서는 다양한 기법들 간의 오차율(또는 예측 정확도)을 비교함으로써, 불일치 패턴 모델 성능의 우수성을 검증하였다. 기존 연

구에서 기본적인 성능의 우수성이 실험 데이터를 통해서 증명이 되었다고 할 수 있으며, 효율적인 적용을 위해서, 불일치 패턴 모델의 정확한 특성을 파악하는 것은 데이터 마이닝 기법적인 측면 및 데이터 모델링 측면에서 많은 시사점을 가진다고 할 수 있다.

불일치 패턴 모델의 특성을 파악한다는 것은 다양한 관점에서 접근할 수 있다. 본 논문에서는 <그림 1>과 같이 불일치 패턴 모델의 특성 파악을 위하여, 첫째, 불일치 패턴 모델을 사용하는 데이터 특성적인 측면과, 둘째, 내부 활용의 기법들 측면에서 어떤 특성이 있는지 확인하며, 마지막으로 지도학습 기법에서 가장 민감한 문제 중 하나인 목표 변수의 불균형을 해소한 후 모델 성능의 변화가 있는지 파악하고자 한다. 이를 파악하기 위하여 본 논문의 구성은 1장 서론 부분에 이어서, 제 2장에서는 관련연구를 기술하고, 제 3장에서는 불일치 패턴 모델의 기초적인 알고리즘을 설명한다. 다음 제 4장에서는 실험 설계 과정과 데이터에 대한 간단한 설명을 기술하고, 제 5장에서는 제 4장에서 수립한 계획에 따라 실험 수행을 하고, 그 결과를 기술하였다. 마지막 제 6장에서는 논문의 최종 결론을 도출하였다.



〈그림 1〉 본 논문의 특성 연구 범위

2. 관련 연구

허준과 김종우[2005, 2007]의 연구를 인용하면, 데이터 마이닝 분야에서 혼합(Hybrid) 모델이나 결합(Combined) 모델을 이용하여 좋은 성과를 거둔 연구들은 매우 활발하게 발표되고 있으며, 그 유형별로도 다양하다. 먼저 2개 이상의 상이한 기법을 혼합하여 만든 혼합 모델을 가지고 모델의 각종 성능 개선을 유도한 연구를 살펴보면, Coenen et al.[2000] 등은 C5.0 알고리즘과 사례기반 추론(Case-based reasoning) 기법을 혼합하여 캠페인의 응답율 향상을 꾀하였으며, Carvalho and Alex[2004]의 경우 C5.0과 Genetic 알고리즘을 융합한 새로운 혼합 모델을 제시하였고, Li and Wang[2004]은 인공 신경망 알고리즘에 Pawlak[1995]에 의해 제시된 불명확한 정보를 처리하는데 사용되는 Rough Set 이론을 접목하여 최종 분류 규칙의 정확도를 향상시킨 연구를 수행하였다. Anand, Patrick, Hughes and Bell[1998]은 교차 판매의 문제에서 Rule Induction 기법과 Deviation Detection 기법을 이용하여, 성능이 향상을 한 연구를 수행하였으며, Wong, Lee and Leung[2004]은 CI(Conditional Independent) test와 Search Phase라는 기법으로 구성된 학습 알고리즘을 이용하여, 검정을 위한 탐색 공간을 축소시켜서, Bayesian Networks의 성능을 향상시키는 연구를 수행하였다. 또한 Hsu, Lai, Chui and Hsu[2003]등은 분석할 데이터의 유형(연속형 인지, 범주형 인지)에 따라 다른 기법을 사용하는 Hybrid 모델을 제시하였으며, Lu, Setiono and Liu[1996]는 신경망 기법을 적용할 때 입력변수의 선택 등을 의사결정나무 추론을 이용하였다. 국내에서는 이극노와 이홍철[2003]은 의사결정나무 추론 기법인 C4.5를 이용하여, 주요 설명변수를 도출해내고, 다음 이 변수들을 사용하여, 신경망 분석

을 수행하는 모델을 제시하였으며, 강문식과 이상용[2002]은 경쟁학습 모델과 신경망의 역전파 오류망(Back Propagation) 알고리즘을 결합한 HACAB(Hybrid Algorithm Combining a Competition Learning Model and BP Algorithm)이라는 혼합 모델을 제안하였다.

위에서 언급한 연구 사례들은 주로 혼합 방법들을 순차적으로 사용한 예라고 할 수 있다. 즉, 어느 한 가지 방법을 수행하고 난 다음 다른 방법을 사용하는 것을 의미한다. 이 방법의 경우 기본적으로 학습이나 훈련을 하는 공간을 축소시키거나, 불필요한 변수들을 제거하여, 성능의 향상을 꾀하는 특성들을 가진 모델이라고 할 수 있다. 이와는 달리 혼합 방법 중 하나의 주요 기법 안에 다른 기법이 내재되는 방법을 통해서 각종 성능을 향상시킨 연구들도 있었다. 예를 들어 Chen[2003]은 SOM(Self-organized Map)을 수행함에 있어, Fuzzy 이론을 이용하여 Text의 분류 성능을 향상시킨 연구를 수행하였고, Versace, Bhatt, Hinds and Sheffier[2004]등은 인공 신경망과 유전자 알고리즘(Genetic algorithm)을 결합하여 새로운 모델을 제시하기도 하였는데, 이 연구에서 신경망 로직의 각종 뉴런 값들을 유전자 알고리즘으로 생성과 반복 재 생성시키는 작업을 통해 예측값의 정확도를 향상시켰다. 또 혼합 방법이 다른 기법에 내재 또는 두 기법이 순서성을 가지고 만들어진 모델이 아니라 병렬적으로 활용된 연구도 있는데, Lin and McClean [2001]은 인공 신경망 기법과 다변량 통계분석의 결과를 결합하여, 기업의 부도 예측 능력을 향상시키는 연구를 수행하였으며, 또한 Converno, Roverta, and Francesco[2002]는 여러 개의 분석기법(회귀분석, 판별분석, 비모수 통계방법, C&RT 등)으로부터 결과 모수를 추출하여 결합하는 혼합 모델을 연구하여 분류 정확도를 향상시키었다고 연구하였다. 국내에서는 김진성[2003]이 지도학습 기법인 Fuzzy 신경망에 연관성 분석을 결합한 모델을 연구하기도 하였다.

혼합 모델의 경우 서로 다른 2개 이상의 모델을 결합 또는 혼합하여 모델의 성능 향상을 이루는 것이 보통인데, 이와는 달리 1개의 모델에 다양한 변화를 주어서 이를 결합하는 모델도 있다. 이런 연구로써 Hansen and Salaman [1990]은 여러 개의 신경망 알고리즘을 결합하여 유의한 성능 변화를 보인 연구를 수행하였고, Indurkhya and Weiss[1998]는 의사결정 나무 추론 알고리즘에서 여러 번 표본을 추출하여 이를 결합하여 유의한 성능 개선을 한 연구를 수행하기도 하였다. Webb and Zheng [2004]은 다양한 앙상블(ensemble) 학습은 단일한 하나의 앙상블 기법보다 성능이 우수하다는 것을 증명하기도 하였다. 국내에서도 이재식과 이진천[2000]이 본 논문의 불일치 패턴 모델과 유사하게 판별모델, 지원 모델, 기본 모델이라는 개념을 이용하여, 특정한 분석 기법으로 특정 모델에 기본 모델을 적용할지 지원 모델을 적용할지를 판별하는 모델을 개발하여 일반 모델보다 좋은 성능을 나타내었다고 보고하였다.

이와 같이 기존의 단일 기법보다 효율성이 높은 혼합 모델의 개발과 비교에 관한 연구 외에, 혼합 모델 또는 결합 모델 자체에 대한 연구도 많이 이루어졌다. Kuncheva, Bezdek, and Shutton [1998]은 혼합 모델을 이용하여 예측력이 향상된 사례 자체들을 연구하였고, Suh[1999] 등은 RFM, 로지스틱 회귀분석, 신경망 모델을 가지고 기법 간에 상관성이 낮을수록 혼합 모델의 성능이 더욱 더 좋아진다는 연구를 수행하였다. Zhang and Zhang[2004]은 자신들의 저서에서 모든 데이터 셋에 혼합

모델이 전부 단일 기법 하나보다 우수한 것은 아니지만 특정한 상황에서는 여러 개의 데이터 마이닝 기법이 혼합되어질 필요가 있다는 것을 주장하기도 하였다. 그리고 Zhou, Wu and Tang[2002]은 신경망으로 구성된 앙상블에서 전체 네트워트를 사용하여, 앙상블을 만드는 것보다 선택된 몇 가지 네트워크만을 이용하여 만드는 앙상블 모델이 더욱 더 효율적임을 언급하고, 이 방법을 GASEN(Genetic Algorithm based Selective ENsemble)이라는 이름을 붙이기도 하였다. 또한 Webb and Zheng [2004]은 앙상블 모델을 여러 개 사용하는 전략이 단순한 앙상블 모델보다 더욱 오류(error)를 감소시킬 수 있다는 것에 대한 전반적인 연구를 수행하였다.

3. 불일치 패턴 모델 알고리즘

허준, 김종우[2005]가 제안한 불일치 패턴 모델을 만드는 과정을 다시 인용하여 정리하면 다음과 같다.

먼저 전체 훈련용 데이터 집합을$L = \{(y_n, x_n), = 1, 2, \cdots, N\}$ 이라고 한다. 여기서 y_n은 목적변수를 의미하고, x_n은 설명 변수 벡터를 의미하며, N은 데이터의 레코드 수를 의미한다. 또한 데이터 집합 L에서 목적 변수값을 가진 컬럼을 T_L로 표현한다. 그리고 훈련용 데이터를 통해서 나온 로직을 검정하기 위한 시험용(Test) 데이터를 $L = \{(y_t, x_t), t= 1, 2, \cdots, T\}$ 라고 정의하도록 한다. 다음은 불일치 패턴 모델을 개발하는 과정을 단계별로 설명하고자 한다.

[단계 1] 전체 훈련용 데이터 집합을 임의 추출방법을 이용하여, 2개로 분리를 한다. 2개의 데이터 집합을 다음과 같이 정의한다.

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

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

단, $M + P = N$

[단계 2] 먼저 L_1 데이터 집합을 이용하여, 분석 기법 A를 이용하여, 모델링을 수행한다. 이 때 모델링을 통해서 생성한 기법 A의 로직, 즉, 예측자(Predictor)를 $\varphi_A(x, L_1)$이라고 한다. 다음 동일한 데이터 집합에 A와는 다른 분석 기법 B를 이용하여 모델링을 수행한다. B기법을 이용하여 모델링을 수행하여 나오게 되는 로직을 $\varphi_B(x, L_1)$이라고 한다. 본 단계는 1단계에서 나누어진 2개의 훈련용 데이터 중 하나를 불일치 패턴 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.

[단계 3] 다음으로 1단계에서 분리한 또 다른 훈련용 데이터 집합인 L_2에 L_1 데이터를 이용하여 생성된 두 기법의 예측 로직인 Predictor $\varphi_A(x, L_1)$과 $\varphi_B(x, L_1)$를 적용시킨다. 먼저 기법 A를 적용시켜서 나온 예측 결과(이 결과는 하나의 컬럼(Column) 형태가 될 것이다.)를 $T_{(A, L_2)}$이라고 하자. 마찬가지로 기법 B를 적용시켜서 나온 결과를 $T_{(B, L_2)}$이라고 하자. 데이터 집합 L_2를 통해서 나온 2개의 결과 값을 서로 비교하여 결과 값이

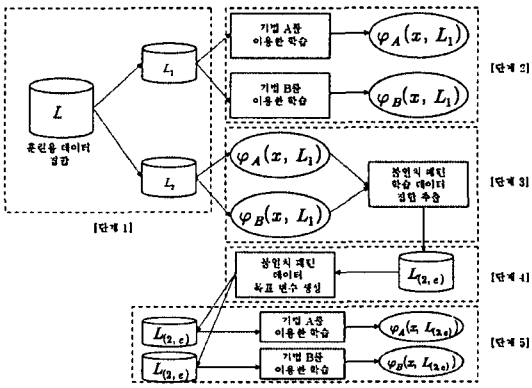
서로 틀린 데이터 집합만을 추출한다. 즉, $L_{(2,e)}=\{(y_i,x_i)|(y_i,x_i)\in L_2$ and $\varphi_A(x_i,L_2)\neq\varphi_B(x_i,L_2)\}$ 이다. 이렇게 추출된 데이터 집합이 바로 불일치 패턴 모델을 생성하기 위한 사례 데이터 집합 생성에 사용이 되며, 다음 단계는 해당 사례 데이터 집합의 새로운 목적변수를 생성하는 단계이다.

[단계 4] 다음 데이터 집합 $L_{(2,e)}$에서 $x_i\in L_{(2,e)}$의 목적 변수 y_i값과 기법 A를 이용하여, 생성된 예측자 $\varphi_A(x_i,L_1)$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 B를 이용하여, 생성된 예측자 $\varphi_B(x_i,L_1)$의 결과값과 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적 변수를 생성한다. 이렇게 새롭게 파생된 목적 변수를 $T_{(B,L_{(2,e)})}$라고 한다.

[단계 5] $L_{(2,e)}$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, B를 다시 적용한다. 즉, $L_{(2,e)}$데이터 집합에서 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$로 교체한 다음 다시 모델링 기법 A를 다시 수행하고, 다시 한 번 역시 기존의 목적 변수 대신에 $T_{(B,L_{(2,e)})}$로 교체한 다음 다시 모델링 기법 B를 수행한다. 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$로 해서 모델링 기법 A를 수행한 후 발생하는 예측자를 $\varphi_A(x,L'_{(2,e)})$라고 하고, 마찬가지로 $T_{(B,L_{(2,e)})}$를 목적 변수로 해서 모델링 기법 B를 수행한 후 발생하는 로직을 $\varphi_B(x,L'_{(2,e)})$라고 한다. 이 단계에서 만들어진 로직 $\varphi_A(x,L'_{(2,e)})$와 $\varphi_B(x,L'_{(2,e)})$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 논문에서 말하고자 하는 불일치 패턴 모델의 핵심이라고 할 수 있다. 본 논문에서는, 이제부터 이 로직 $\varphi_A(x,L'_{(2,e)})$와 $\varphi_B(x,L'_{(2,e)})$을 불일치 패턴 모델(Inconsistent Pattern Model)이라고 정의한다. 이 5단계까지의 과정을 도식화하면 다음의 <그림 2>와 같다.

[단계 6] 다음 이렇게 불일치 패턴 모델을 구했으면, 이를 적용한 최종 예측자를 생성하게 되는데 이 과정은 보팅(Voting) 방법을 이용한다. 예를 들어서 시험용 데이터 집합인 $L_t=\{(y_t,x_t),t=1,2,\cdots,T\}$에서 먼저 $\varphi_A(x,L'_{(2,e)})$의 로직을 적용하여, 예측값이 T가 되는 사례는 $\varphi_A(x,L_1)$ 로직을 이용한 결과값을 선택하고, 아닌 것은 $\varphi_B(x,L_1)$ 로직을 이용한 결과 값을 선택한다. 이렇게 판별하여 생성된 예측자를 $\varphi_{(A,B)}(x,L)$ 라고 정의하고, 다음 반대로 $\varphi_B(x,L'_{(2,e)})$의 로직을 적용하여, 예측값이 T가 나온 사례에는

$\varphi_B(x, L_1)$로직을 적용한 결과 값을 선택하고 F인 것은 $\varphi_A(x, L_1)$ 로직을 적용한 결과 값을 선택한다. 이렇게 조합을 통해서 나온 최종 예측자를 $\varphi_{(B,A)}(x, L)$라고 한다. 이들이 각기 다를 수 있으므로 안정적인 예측자를 생성하기 위해 이 2개의 계산된 예측자들에게서 확률값이 큰 쪽을 선택하여 만들어낸 최종 예측자인 $\varphi_{<A,B>}(x, L)$를 생성하게 되면, 모든 과정이 완료된다. 본 내용을 예제 데이터를 사용하여 표현하면 다음 <표 1>, <표 2>와 같다. 또한 이를 도식화 하면 <그림 3>과 같다.



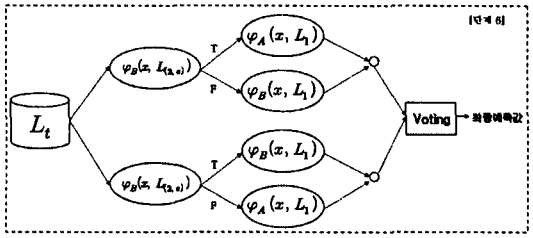
<그림 2> 불일치 패턴 모델의 생성과정 도식화

<표 1> 시험용 데이터에서 불일치 패턴 모델의 판별 결과와 두 기법의 예측자 예제

번호	$\varphi_A(x, L'_{(2,e)})$	$\varphi_B(x, L'_{(2,e)})$	$\varphi_A(x, L_1)$		$\varphi_B(x, L_1)$	
			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 보팅(Voting)한 결과

번호	$\varphi_{(A,B)}(x, L)$		$\varphi_{(B,A)}(x, L)$		$\varphi_{<A,B>}(x, L)$	
	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

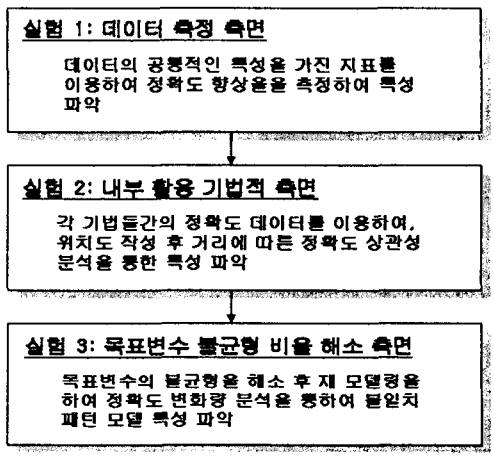


<그림 3> 불일치 패턴 모델의 적용을 위한 최종 보팅(Voting) 과정

4. 실험의 설계

4.1 실험의 기본 설계

본 논문의 목적인 불일치 패턴 모델의 특성을 파악하기 위하여, 3가지의 실험 설계를 계획하였다. 이를 실험을 통하여 수행하고자 하며, 이를 도식화한 것이 <그림 4>와 같다.



<그림 4> 실험의 설계

<그림 2>에서 볼 수 있듯이 불일치 패턴 모델의 성능은 대상이 되는 데이터와 불일치 패턴 모델을 내부적으로 구성하는 지도학습 기법들에 따라서 결정된다고 할 수 있다. 따라서 <그림 4>에서 보여 지는 첫 번째 실험과 두 번째 실험을 통해서, 불일치 패턴 모델이 가지는 가장 큰 특징을 알 수 있다. 세 번째 실험은 첫 번째의 데이터 특성에 따른 모델의 성격을 파악하는 실험에 대한 확장적인 의미를 가지는 실험으로, 지도학습 기법에서 훈련 및 예측의 주 대상이 되는 목표 변수의 불균형을 해소한 후 성능 변화 정도를 실험을 통해 확인함으로써, 불일치 패턴 모델의 특성을 더욱 잘 파악하고자 하는 목적을 가지고 있다. 실험의 편의를 위해서 본격적인 실험을 수행하기 전에 몇 가지 정리할 사항이 있다. 첫 번째로 본 논문에서 단일 기법 및 지도학습 기법으로 사용한 기법으로는 신경망의 MLP(Multi-Layer Perceptron) 기법과 의사결정나무 추론 기법의 C5.0[Quinlan, 1993, 1996], C&RT[Breiman et al., 1983] 그리고 로지스틱 회귀분석을 선정하였다. 실제로는 이들 4가지 기법 이 외에 더 많은 지도학습 기법들이 존재하지만, 이 4개의 기법만을 이용한 이유는 직접 프로그래밍을 하여 발생하는 오류를 없애고, 실험의 신뢰성과 편의성을 위하여 기 검증된 SPSS 社의 Clementine 8.1을 사용하여 분석을 수행하였는데, 활용된 분석 솔루션에서 제공하는 지도학습 기법만을 이용하기 때문이다. 그리고 본 논문에서 적용하는 모든 기법들의 여러 옵션(option)들은 모두 데이터 마이닝 솔루션인 Clementine 에서 제공하는 기본 옵션을 이용하였으며[SPSS Inc., 2003], 통계 검정 및 다차원 분석과 같은 통계분석의 경우 SPSS 14.0.2K를 이용하여 수행을 하였다.

4.2 실험에 사용된 데이터

본 실험에서 활용된 데이터 수는 총 23개 데이터 집합들이다. 데이터 집합의 간단한 구성 내역을 설명하면, 15개의 데이터 집합은 실제 데이터 마이닝 작업을 하면서 활용된 데이터 집합이고, 나머지 8개는 UCI Machine Learning Repository[www.ics.uci.edu]와 Carnegie Mellon 대학의 AutonLab 실험 데이터[www.qutonlab.org], MS-SQL 2000 Server Analysis Service의 예제 데이터 등 누구에게나 공개된 데이터를 이용하였다. 이는 실험을 통해 나온 결과에 대하여 실제 활용성과 신뢰성을 확보하기 위한 구성이라고 할 수 있다. 즉, 데이터 마이닝 자체가 실제 분석 상황에서 많이 활용되기 때문에, 실제 상황의 데이터를 이용하여 활용성을 높이는 목적을 달성하고, 동시에 누구나 다시 재 실험을 할 수 있도록 공개된 데이터를 이용한 실험을 수행하여, 실험 결과의 신뢰성도 높이고자 2가지의 성격을 혼합한 23개의 데이터 집합들이 선택되었다.

5. 실험결과

5.1 데이터 특성에 따른 불일치 패턴 모델 성능 향상 실험 결과

<그림 4>의 실험 설계에 따라서, 첫 번째 실험을 수행하기 위해 데이터의 특성을 정의한 지표들을 살펴보면 다음의 <표 3>과 같다.

<표 3>에서 정의한 지표들은 본 논문에서 사용한 23개의 데이터 집합들이 각기 다양한 성격을 가지기 때문에 이들에게서 공통적으로 특성을 정의하여 줄 수 있는 것만 추출을 한 것이다. 다음의 <표 4>는 23개 데이터 집합별로 데이터 특성 지표들의 값을 정리한 것이다.

〈표 3〉 데이터 특성을 위한 지표의 정의

데이터 특성 지표명	정 의
목표변수 불일치 비율	(A기법과 B기법의 다른 결과를 내는 사례수)/(TEST 데이터 집합 전체 사례수) X 100
목표변수 불균형 비율	ABS((목표변수 중 참 값의 수-목표변수 중 거짓 값의 수)/TEST 데이터 집합 전체 사례수) X 100
설명변수 종류의 차 비율	ABS((설명변수 중 연속형 변수의 수)-(설명변수 중 범주형 변수의 수))/전체 설명 변수의 가짓 수 X 100
데이터의 수	데이터의 전체 사례 수
변수의 수	전체 설명변수의 전체 가짓 수

<표 4>에서 최소 정확도 향상율이라는 것은 불일치 패턴 모델을 수행한 결과 비교하고자 하는 기존의 방법보다 예측율이 더 향상된 정도를 의미 한다. 예를 들어 C5.0과 신경망을 이용하여 개발한 불일치 패턴 모델의 경우 신경망 분석 단일 기법보다 향상된 정도 그리고 단일기법 C5.0보다 정확도가 향상된 비율 2가지의 값 중 최소값을 의미한다. 그 옆의 최대 정확도 향상율은 향상된 비율 2가지 값 중 최대값을 의미한다. 여기의 불일치 비율과 정확도 향상율은 실험에 사용된 불일치 패턴 모델인 신경망/C5.0 불일치 패턴 모델, 신경망/C&RT 불일치 패턴

〈표 4〉 데이터 집합별 데이터 특성 지표값

데이터 번호	불일치패턴 모델목표변수 불일치 비율(%)	목표변수 불균형 비율(%)	설명 변수 의 종류 차 비율(%)	데이터의 수	변수의 수	최소 정확도 향상율(%)	최대 정확도 향상율(%)
1	31.48	18.58	44.44	14,490	18	3.38	4.77
2	9.46	51.98	14.29	32,561	14	0.33	0.98
3	5.68	59.44	40.00	12,000	10	0.24	1.7
4	23.90	22.16	33.33	3,284	9	3.24	4.91
5	6.82	23.06	7.69	4,574	13	-0.03	0.4
6	3.98	64.98	53.85	100,000	13	0.05	0.21
7	19.93	45.16	57.14	228,774	14	0.29	1.37
8	31.29	7.06	21.74	2,181	23	6.01	10.48
9	48.53	1.32	0.00	30,000	20	4.54	19.45
10	5.48	7.26	67.57	16,004	37	0.1	0.39
11	33.27	23.28	25.00	3443	8	2.95	6.03
12	17.84	13.20	40.00	44,942	10	2.03	5.3
13	0.32	91.96	47.83	16,797	23	-0.03	0.03
14	41.14	8.70	62.50	1,000	16	0.55	2.4
15	6.00	51.06	47.83	5,619	23	0.07	0.27
16	9.41	22.98	100.00	4,601	57	0.44	1.27
17	22.09	27.60	100.00	958	9	1.79	4.61
18	9.69	18.94	9.09	690	11	0.59	1.77
19	9.40	52.76	14.29	48,843	14	0.33	1.09
20	9.83	11.00	100.00	506	13	-1.74	0.58
21	13.58	3.37	75.00	4,177	8	-0.16	0.74
22	9.09	17.74	40.00	10,282	10	0.05	0.59
23	0.30	94.43	100.00	26,733	10	0.03	0.06

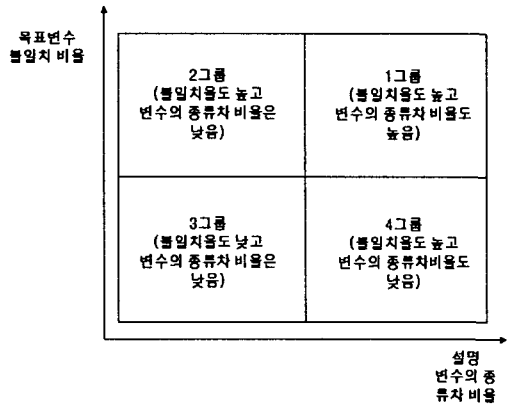
모델, C5.0/C&RT 불일치 패턴 모델 등, 단일 기법 3가지를 이용하여 만든 불일치 패턴 모델과 Bagging, Boosting, 그리고 Stacking 등 결합 모델을 통합 3개의 불일치 패턴 모델 등 총 6개의 불일치 패턴 모델의 평균값을 낸 것이다. 한 가지 정의할 사항은 최대 정확도 향상율은 만약 동일한 데이터에서 기법을 잘 선택하는 경우 최대 얻을 수 있는 향상율의 의미를 가지고 있다는 것이다. 다음의 <표 5>는 정의된 지표들과 정확도 향상율의 Pearson 상관관계 분석을 수행한 결과이다.

〈표 5〉 데이터 특성 지표와 정확도 향상율과의 상관 관계

데이터 특성 지표	최소 정확도 향상율과의 상관관계	최대 정확도 향상율 과의 상관관계
목표변수 불일치 비율	0.726**	0.705**
목표변수 불균형 비율	-0.374	-0.447*
설명변수의 종류 차 비율	-0.423*	-0.403
데이터의 수	-0.127	-0.087
변수의 수	-0.001	-0.008

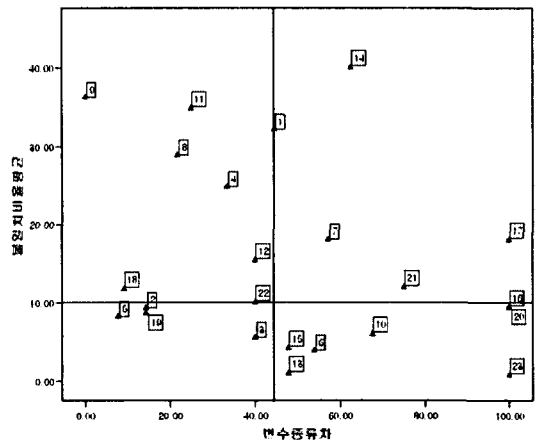
주) ** 0.01수준에서 유의함, * 0.05수준에서 유의함.

<표 5>에서 보면, 최소 정확도 향상율은 목표변수의 불일치 비율과 설명변수의 변수 종류차 비율에서 유의한 상관성을 보이고 있으며, 최대 정확도 향상율은 역시 목표변수 불일치 비율과 목표변수 불균형 비(比)와 유의한 상관성을 보이는 것으로 나타났다. 그 외에 다른 데이터 특성 지표들은 유의한 상관성이 없는 것으로 나타났다. 위의 상관관계를 보면 불일치 패턴 모델의 성능 향상율에 도움을 주는 것은 2개 기법 간의 불일치 비율(서로 맞추는 사례들의 다른 정도가 심한 것)과 목표변수의 범주 간 균형이 이루어진 정도 그리고 마지막으로 설명변수에 연속형과 범주형 변수가 골고루 분포되어져 있는 경우에 불일치 패턴 모델의 성능이 더 우수한 것을 알 수 있다. 이 사항을 좀 더 정확하게 파악하기 위하여, 다음의 <그림 5>와 같이 평균 정확도 향상율을 위한 목표 변수 불일치 비율과 설명변수의 종류차 비율로 분할표를 만들어 산점도를 구성하여 본다.



〈그림 5〉 목표변수 불일치 비율과 설명변수의 종류차 비율의 분할표 설명

<그림 5>와 같은 분할표에 따라 산점도를 그린 결과는 다음의 <그림 6>과 같다.



〈그림 6〉 불일치 비율과 설면변수의 종류차의 분할표에 따른 23개 데이터의 산포

<그림 6>에서 분할면을 자르는 기준으로는 실험에 활용된 23개의 데이터 집합에서 불일치 비율과 설명변수의 종류차의 중위수[2]를 가지고 분할의 기준으로 삼았다. 불일치 비율의 경우 중위수가 10.04이고, 설명 변수의 종류차는 44.44였다.

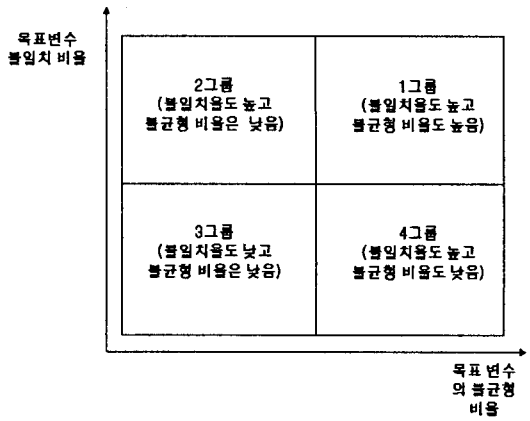
다음 <그림 6>에서 만들어진 4개의 분할면을 그룹으로 하여 이들 4개 그룹이 평균 정확도 향상율의 차이를 보이는지 확인하기 위하여, One-Way ANOVA 분석을 하였으며, 사후 검정 방법으로는 Duncan 방법을 이용하였다. 또한 데이터의 수가 작은 관계로 정규 분포를 가정하지 않을 수도 있기 때문에 비모수의 Kruskal-Wallis 검정을 동시에 수행한 결과가 <표 6>과 같다.

〈표 6〉 불일치 비율과 설명변수의 종류차 분할표에 따른 사분면 그룹별 최소 정확도 향상율의 검정 결과

그 룹	최소 정확도 향상율의 그룹별 평균값	One-Way ANOVA F통계량 (p-value)	Kruskal-Wallis Chi-제곱 통계량 (p-value)
1사분면	1.1707	5.632 (0.006)	10.177 (0.017)
2사분면	2.7731		
3사분면	0.2171		
4사분면	-0.1529		
Duncan의 사후검정	2사분면 = 1사분면 > 3사분면 = 4사분면		

<표 6>을 보면 2사분면인 데이터의 불균형 비율이 높고, 설명변수의 종류차가 낮은 그룹들의 최소 정확도 향상율의 평균이 가장 높은 것으로 나타났다. 그리고 불일치 비율이 높고 설명 변수의 종류차가 높은 그룹이 그 뒤를 이었다. 2개의 검정 모두에서 95% 유의수준을 가지고 나타났다. 이는 정확도의 향상율에 있어서, 불일치 비율이 가장 큰 영향을 미치는 것을 알 수 있다. 또한 설명변수에서 연속형과 범주형이 골고루 분포되어져 있는 것을 의미하는 변수의 종류차라는 지표는 통계적 검정은 유의하지 않았지만, 기본적으로 연속형과 범주형이 다양하게 분포되어져 있는 설명변수를 가진 데이터 집합에서 복잡한 형태의 데이터서 불일치 패턴 모델이 더 효율성이 있을 가능성이 높은 것을 또한 알 수 있었다. 이는 향후 데이터 집합을 추가로 더 분석하는 경우 다시 검증이 필요할 것이다.

다음은 최대 정확도 향상율과 상관성이 높은 불일치 비율과 목표변수의 불균형 비율을 분할표로 만들어 본 것이 <그림 7>과 같다.



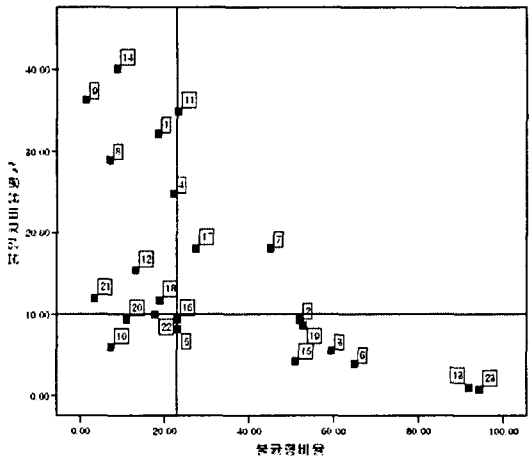
〈그림 7〉 목표변수 불일치 비율과 불균형 비율의 분할표 설명

<그림 7>의 분할표에 맞추어서, 23개 실험 데이터의 산포를 살펴본 결과가 <그림 8>과 같다.

아래의 <그림 8>에서 불일치 비율의 분할 기준이 되는 중위수는 앞의 <그림 6>과 같이 10.04이며 목표변수의 불균형 비율의 분할 기준이 되는 중위수의 값은 22.98이다. 역시 마찬가

2) 이상값 등으로 인한 평균의 왜곡을 방지하기 위하여 중위수를 이용하였다.

지로 One-Way ANOVA 분석 및 Duncan의 사후검정과 비모수 통계검정 방법인 Kruskal-Wallis 검정을 통해서 각 사분면의 그룹들이 차이가 있는지 확인한 결과가 <표 7>과 같다. 단, 여기서 검정 대상이 되는 값은 <표 5>의 결과에 의거하여 최대 정확도 향상율을 이용하였다.



〈그림 8〉 불일치 비율과 목표 변수의 불균형 비율의 분할표에 따른 23개 데이터의 산포

〈표 7〉 불일치 비율과 목표변수 불균형 비율의 분할표에 따른 사분면 그룹별 평균 정확도 향상율의 검정 결과

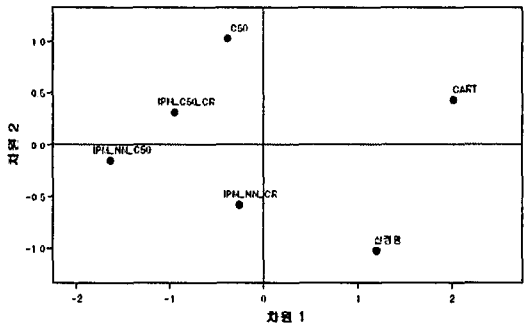
그 룹	최대 정확도 향상율의 그룹별 평균값	One-Way ANOVA F통계량 (p-value)	Kruskal-Wallis Chi-제곱 통계량 (p-value)
1사분면	4.0033	3.518 (0.035)	14.208 (0.003)
2사분면	6.2275		
3사분면	0.6460		
4사분면	0.6200		
Duncan의 사후검정	2사분면 = 1사분면 > 3사분면 = 4사분면		

위의 <표 7>을 보면, 역시 불일치 비율이 높고, 목표 변수의 불균형 비율이 낮은 2사분면이 가장 높은 정확도 향상율을 보인 것으로 나타났다. 일반적으로 목표 변수의 불균형(imbalance)은 지도학습 기법의 상당한 걸림돌인데, 이는 마찬가지로 불일치 패턴 모델에서도 그대로 나타나서, 향후 불일치 패턴 모델 정확도의 향상을 위한 추가 연구로 목표 변수가 불균형한 데이터를 각종 Sampling 기법을 이용하여, 균형을 맞춘 후 불일치 패턴을 사용하여 분석을 수행하는 방법에 대한 연구가 필요할 것이다.

위의 <표 6>과 <표 7>을 통해서 보면 불일치 패턴 모델의 경우 기법 간 불일치율이 높은 데이터 집합들과 설명변수가 연속형과 범주형이 골고루 섞여져 있는 데이터 집합 그리고 목표 변수의 불균형까지 해소가 되어져 있는 경우에, 예측력의 향상이 매우 기대되는 기법이라고 할 수 있을 것이다. 이는 나아가, 기법 간에 불일치가 높을 가능성이 많고, 다양한 종류의 데이터가 존재하는 복잡한 실무 데이터 분석에 본 모델이 유용할 수 있다는 결론을 추론할 수 있다.

5.2 불일치 패턴 모델 내부사용 기법에 따른 위치도 분석 결과

앞에서 데이터의 특성에 따라서 불일치 패턴 모델의 정확도 향상이 어떻게 되는지 확인을 해보았다. 다음은 데이터의 특성이 아닌 기법 간의 상관성에 대하여 위치도를 이용하여 불일치 패턴 모델의 특성 파악을 해보기로 한다. 단일 기법인 신경망 분석, C5.0, C&RT 그리고 이들 3개 기법을 가지고 만든 3개의 불일치 패턴 모델 등 6개 모델에 대한 23개 데이터 집합들을 가지고 다차원 척도법(Multi-Dimensional Scaling : MDS)을 이용하여 위치도를 작성한 것이 다음의 <그림 9>와 같다.



〈그림 9〉 6개의 모델에 대한 MDS를 이용한 위치도

다음의 <표 8>은 6개의 모델에 대한 23개 데이터 집합들의 예측 정확도의 평균값을 나타낸 표이다. 위의 <그림 9>를 이해하는데 도움을 줄 값들이다.

〈표 8〉 6개의 모델에 대한 23개 데이터 집합들의 예측 정확도 평균값

기법명	예측 정확도 평균값	기법명	예측 정확도 평균값
신경망	80.32%	신경망-C5.0 불일치 패턴 모델	84.41%
C5.0	80.98%	신경망-C&RT 불일치 패턴 모델	82.73%
C&RT	79.67%	C&RT-C5.0 불일치 패턴 모델	82.44%

<그림 9>를 보면 6개의 기법들의 23개 데이터에 대한 예측 정확도 값의 기법 별로 위치도를 그린 것이다. 위의 6개점들의 위치를 보면 점들이 서로 가까울수록 비슷한 결과들을 낸 기법끼리 모여 있다고 할 수 있다. <그림 9>의 그림을 보면 신경망과 C5.0의 불일치 패턴 모델과 C&RT 모델이 가장 멀리 떨어져 있는 것을 알 수 있다. 이들의 23개 데이터의 예측 정확도를 맞춘 평균값을 <표 8>에서 보면, 신경망과 C5.0의 불일치 패턴 모델이 84.41% 그리고 C&RT가 79.67%로써 가장 차이가 나서, 이들 간의 거리가 있는 것을 알 수가 있다.[3] 위의 <그림 9>를 보면 단일 기법 3개(신경망, C5.0, C&RT) 중에서 C5.0과 신경망이 가장 거리가 멀며, 이들 2개를 가지고 만든 불일치 패턴 모델이 다른 2개의 불일치 패턴 모델보다 전체적인 평균이 높은 것 <표 8>을 통해서 알 수 있는 것처럼, 기법 간의 거리가 있는 즉, 전반적으로 데이터 마이닝 지도학습 기법의 상이성이 높은 것일수록 불일치 패턴을 만들 때 좀 더 효율적이라는 것을 살펴볼 수가 있다.

<그림 9>를 통해서, 단일 기법 중 MDS의 차원 축소에 따른 오류를 감안하더라도 거리감이 있는 기법들 즉, 기법의 성격이 상이한 기법들을 가지고 불일치 패턴을 만드는 경우 좀 더 효율적인 성능 향상을 기대해 볼 수 있고, 이는 데이터의 성격과 함께 기법의 특성을 이용한다면 좀 더 효율적인 불일치 패턴 모델을 만들 수 있을 것으로 판단된다.

5.3 목표 변수 불균형 해소 후 불일치 패턴 모델의 변화 실험 결과

지도학습 기법에서 문제가 되는 큰 문제 중 하나가 바로 목표 변수의 불균형이다. 실제 대다수의 데이터는 목표 변수의 불균형이 크던 작던 있기 마련이며, 5 : 5의 비율로 있는 경우는 극히 드물다. 그렇다면, 목표변수의 불균형을 해소시킨다면, 좀 더 불일치 패턴 모델의 성능은 좋아질 것인가? 이에 대한 의문을 해결하는 실험을 통하여, 불일치 패턴 모델의 특성을 파악하고자, 목적변수의 불균형이 된 것을 샘플링(sampling)을 이용하여, 균형을 맞춘 다음 모델

3) MDS의 다차원을 2차원으로 축소하여 투영시키기 때문에 일부 오류도 있을 수 있다.

〈표 9〉 일부 데이터의 불균형 해소 후 불일치 패턴 모델링 결과

데이터번호	초기의 원 데이터 이용 불일치 패턴 모델링 결과			샘플링을 통한 균형 후 불일치 패턴 모델링 결과			목표변수 불균형비율	샘플링방법
	C5.0-신경망	신경망-C&RT	C&RT-C5.0	C5.0-신경망	신경망-C&RT	C&RT-C5.0		
3번	94.67%	93.77%	94.67%	**95.29%**	**94.29%**	**95.09%**	79 : 21	Over
6번	96.22%	95.05%	96.14%	95.41%	94.49%	95.64%	82 : 18	Over
15번	75.64%	75.31%	75.20%	**76.23%**	71.46%	**75.96%**	78 : 22	Over
16번	92.81%	91.06%	90.73%	92.20%	**91.87%**	**91.60%**	60 : 40	Over
				91.13%	**92.67%**	89.99%		Under

을 생성하여 불일치 패턴 모델을 다시 만드는 실험을 수행한 결과가 <표 9>와 같다.

<표 9>를 보면, 3번 데이터의 경우 과대 샘플링(Over Sampling; 이분형 목표 변수 중 더 많은 쪽에 기준을 맞추어서 임의 샘플링[4]을 수행하는 방법)을 이용하여, 목표변수의 불균형비를 5 : 5로 동일하게 맞추고 수행한 결과 기존의 3가지 불일치 패턴 보다 성능이 더 향상이 된 반면, 6번 데이터 경우 동일한 방법을 수행하였음에도 불구하고, 3가지의 불일치 패턴 모델 모두 성능이 기존의 불일치 패턴 모델 보다 더 안 좋은 것을 알 수 있다. 그리고 가장 아래쪽의 16번 데이터의 경우 과대 샘플링을 수행하여, 5 : 5로 목표 변수의 균형을 맞춘 결과 신경망-C&RT 그리고 C&RT-C5.0 불일치 패턴 모델의 경우 기존보다 더 성능이 향상된 반면, 이를 과소 샘플링(Under Sampling)을 수행하여 5 : 5로 수행하면, 신경망-C&RT 불일치 패턴 모델만 성능이 향상되고, 다른 모델은 더 성능이 저하되는 것을 알 수 있다. 이는 다시 말해서, 단순히 목표 변수의 불균형이 있는 경우 이를 임의 샘플링을 통해서 균형을 잡아주는 것으로 불일치 패턴 모델의 성능이 향상되는 것이 아니라 각 데이터 집합마다 사전에 가장 효율적인 성능을 발휘하는 샘플링의 비율이나 방법을 선택하고, 그 다음에 불일치 패턴 모델을 적용하면 기존보다 더욱 성능이 좋아진다고 판단할 수 있다.

4) Random Sampling.

6. 결 론

본 논문에서는 지도학습 기법의 새로운 혼합 모델인 불일치 패턴 모델의 특성이 어떤지를 다양한 실험을 통하여 파악하는 것을 목적으로 하고 있다. 서론에서도 언급을 하였지만 기존 연구에서 불일치 패턴 모델이 단일 지도학습 기법이나 기존의 결합 모델보다 더 좋은 분류 정확도 및 예측 정확도를 나타내는 것을 확인하였기에, 불일치 패턴 모델의 정확한 특성을 파악하는 것은 매우 의미 있는 작업이라 할 수 있다.

결론적으로 불일치 패턴 모델의 특성을 요약하면 다음과 같다.

먼저, 데이터적인 특성을 살펴보면, 가장 정확도의 향상율에 기여를 하는 것으로, 통계적으로 검정된 것은 두 기법 간의 데이터에 대한 불일치 비율이 높은 것을 알 수 있었고, 그 외에 데이터의 목표 변수의 불균형이 작고, 범주형과 연속형 데이터가 적절하게 구성되어져 있는 데이터에서 정확도 및 예측력의 향상율이 높은 것을 알 수 있었다. 또한 목표 변수의 불균형 비

율이 낮을수록 불일치 패턴 모델의 예측 정확도의 향상이 높을 수 있다는 판단을 할 수 있었다.

다음으로 기법의 활용 측면에서 보면, 기법들 간의 상이성(相異性)이 전체적인 기법의 차이를 많이 나타내는 것을 알 수 있었는데, 이를 확인하기 위하여, 다차원 척도법을 통한 위치도에서 기법 간의 거리가 가장 많이 떨어진 기법들을 이용하여 불일치 패턴 모델을 개발하는 경우 효율성이 있는 것으로 나타났다. 예를 들면, 동일한 실험에서 같은 의사결정나무 추론 기법들만으로 구성된 불일치 패턴 모델보다는 서로 알고리즘이 판이하게 다른 신경망 기법 1개와 의사결정나무 추론 기법 1개를 이용하여 개발하는 것이 더 효율적이라는 것이다. 일반적으로 유전학 등에서 이종교배를 하는 경우가 자손의 형질 발현에 있어서, 상호간의 우성 형질을 더 강하게 보여주는 것과 마찬가지로 기법 사용에서도 서로 이질적인 방법을 사용하여 혼합하는 경우 더욱 성능이 좋아진다는 것을 보여주는 것으로 매우 흥미로운 특성이라고 하겠다.

세 번째 특성 파악 실험에서 목표 변수의 불균형을 적절한 샘플링을 통해서 사전에 불균형을 해소시킨 다음 불일치 패턴 모델을 적용하면 더욱 효율적일 수 있다는 것을 알 수 있었다.

불일치 패턴 모델의 특성을 파악하는 연구로는 본 연구가 전부는 아닐 것이며, 더 다양한 관점에서 연구가 필요할 것으로 보인다. 특히, 단일 기법 및 기존의 결합 기법들 보다 성능의 향상이 발생하는 근원적인 이유에 대하여, 보편적인 증명을 할 수 있는 연구들은 향후에 더욱 많이 필요로 할 것이며, 더 많은 데이터 사례에서의 검증과 데이터 모델링에 따른 특성을 파악하는 상세한 연구가 필요할 것으로 본다.

참 고 문 헌

[1] 강문식, 이상용, "데이터 마이닝을 위한 경쟁학습모델과 BP알고리즘을 결합한 하이브리드 신경망", *정보기술과 데이터베이스 저널*, 제9권 제2호, 2002, pp. 1-16.

[2] 김진성, "연관규칙과 퍼지 인공신경망에 기반한 하이브리드 데이터 마이닝 메커니즘에 대한 연구", *한국경영과학회/대한산업공학회 2003 춘계 공동학술대회논문집*, 2003, pp. 884-888.

[3] 이군희, "모형평가와 앙상블을 이용한 데이터 마이닝에 관한 연구", *서강경영논총*, 제9권, 1998, pp. 293-306.

[4] 이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘 연구", *한국지능정보시스템학회지*, 제9권 제1호, 2003, pp. 139-155.

[5] 이재식, 이진천, "입력자료 판별에 의한 데이터 마이닝 성능개선", *한국지능정보학회 학술대회*, 2000, pp. 293-303.

[6] 허준, 김종우, "데이터 마이닝 결합 모델과 오차패턴 모델의 통합적 활용", *한국경영과학회/대한산업공학회 2007 춘계 공동학술대회논문집*, 2007.

[7] 허준, 김종우, "오차패턴 모델링을 이용한 Hybrid 데이터 마이닝 기법", *한국경영과학회지*, 제30권 제4호, 2005, pp. 27-43.

[8] _____, "Clementine Ver. 8 User's Guide", *SPSS Inc*, 2003.

[9] Anand, S. S., A. R. Patrick, J. G. Hughes, and D. A. Bell, "A data mining methodology for cross-sales", *Knowledge-Based Systems*, Vol. 10, 1998, pp. 449-461.

[10] Brieman, L. "Bagging Predictors", *Machine Learning*, Vol. 24, No. 2, 1996, pp. 123-140.

[11] Brieman, L. "Random Forests", *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5-32.

[12] Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone, "Classification and regression trees", Wadsworth, Belmont, 1984.

[13] 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.

[14] Chen, Y. P., "A hybrid framework using SOM and fuzzy theory for textual classification in data mining", *Modeling with Words, LNAI2873*, 2003, pp. 153-167.

[15] Coenen, F. G., S. K. Vanhoof, and G. Wets, "The Improvement of Response Modeling : Combining Rule-induction and Case-based Reasoning", Expert Systems with Application, Vol. 18, No. 4, 2000, pp. 307-313.

[16] Conversano, C., R. Siciliano, and F. Mola, "Generalized Additive Multi-mixture Model for Data Mining", *Computational Statistics and Data Analysis*, Vol. 38, No. 4, 2002, pp. 487-500.

[17] Fayyad, U. M., "Data Mining and Knowledge Discovery Making Sense Out of Data", *IEEE Expert*, Vol. 11, No. 5, 1996, pp. 20-25.

[18] Freund, Y., and R. E. Schapire, "Experiments with a New Boosting Algorithm", *Proceedings of 13th International Conference on Machine Learning*, Morgan Kaufmann, 1996, pp. 148-156.

[19] Han, J. and M. Kamber, "Data Mining : Concepts and Techniques", Morgan Kaufmann, 2001.

[20] Hansen, L. K. and P. Salaman, "Neural Networks Ensembles", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, 1990, pp. 993-1001.

[21] Hsu, P. L., R. Lai, C. C. Chui, and C. I. Hsu, "The Hybrid of Association Rule Algorithms and Genetic Algorithm for Tree Induction : An Example of Predicting the Student Course Performance", *Expert Systems with Application*, Vol. 25, No. 1, 2003, pp. 51-62.

[22] Indurkhya, N., and S. M. Weiss, "Estimating Performance Gains for Voted Decision Trees", *Intelligent Data Analysis*, Vol. 2, No. 1/4, 1998, pp. 303-310.

[23] 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.

[24] Li, R. and Z. Wang, "Mining Classification Rules Using Rough Sets and Neural Networks", *European Journal of Operational Research*, Vol. 157, No. 2, 2004, pp. 439-448.

[25] Lin, F. Yu. 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.

[26] Lu, H., R. Setiono, and H. Liu, "Effective data mining using neural networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, 1996, pp. 957-961.

[27] Pawlak, Z., J. Grzymala-Busse, R., and

Slowinski, W. Ziarko, "Rough sets", Communications of the ACM, Vol. 38, No. 11, 1995, pp. 88-95.

[28] Quinlan, J. R., "Bagging, Boosting and C4.5", *Procs. 13th American Association for Artificial Intelligence*, AAAI Press, 1996.

[29] Quinlan, J. R., "C4.5 Programs for machine Learning", San Mateo : Morgan Kaufmann, 1993.

[30] 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 International Conference on Machine Learning*, Morgan Kaufmann, 1998, pp. 322-330.

[31] Suh, E. H., K. C. Noh, and C. K. Suh "Customer List Segmentation Using the Combined Response Model", *Expert Systems with Application*, Vol. 17, No. 2, 1999, pp. 89-97.

[32] 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.

[33] Webb, G. I. and Z. Zheng, "Multistrategy ensemble learning : reducing error by combining ensemble learning techniques", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 8, 2004, pp. 980-991.

[34] Wolpert, L., "Stacked Generalization", Neural Networks, Vol. 5, No. 2, 1992, pp. 241-259.

[35] Wong, M. L., S. Y. Lee, and K. S. Leung, "Data mining of Bayesian networks using cooperative coevolution", *Decision Support systems*, Vol. 38, 2004, pp. 451-472.

[36] Zhang, Z. and C. Zhang, "Agent-Based Hybrid Intelligent Systems", *LNAI 2938*, 2004, pp. 127-142.

[37] Zhou, Z., J. Wu, and Wei Tang, "Ensembling Neural Networks : Many Could Be Better Than All", *Artificial Intelligence*, Vol. 137, No. 1/2, 2002, pp. 239-263.

[38] http://www.autonlab.org/autonweb/downloads/datasets.html.

[39] http://www.ics.uci.edu/~mlearn/MLRepository.html.

저자소개

허 준

중앙대학교 정경대학 응용통계학과에서 경제학사(1998), 중앙대학교 대학원 통계학과에서 경제학 석사(2000), 그리고 한양대학교 경영학과에서 경영학 박사(2008) 학위를 취득하였다. 현대정보기술(HIT)과 현대/기아 자동차 그룹 계열사인 오토에버닷컴에서 근무를 하였고, 현재 SPSS Korea 컨설팅팀 수석연구원으로 재직 중 이다. 주요 관심분야로는 데이터 마이닝과, 데이터베이스 마케팅/CRM 관련분야 및 수요예측 모델링 분야이다.

김 종 우

서울대학교 수학과에서 이학사(1989), 한국과학기술원 경영과학과에서 공학석사(1991), 한국과학기술원 산업경영학과에서 공학박사(1995)를 취득하였다. 현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 충남대학교 통계학과 부교수, University of Illinois at Urbana-Champaign 방문연구원 등을 역임하였다. 주요 관심분야는 상품/컨텐츠 추천 시스템, 데이터마이닝, 지능정보시스템, 의사결정지원시스템, 비즈니스 프로세스 모델링 및 통합, 데이터 품질 등이다.

◈ 이 논문은 2008년 03월 06일 접수하여 1차 수정을 거쳐 2008년 03월 14일 게재확정되었습니다.