데이터 마이닝의 지도학습 기법 성능향상을 위한 불일치 패턴 모델

Inconsistent Pattern Model for Improving the Performance of Supervised Learning in Data Mining

허 준*, 김종우**

* SPSS Korea (주)데이터솔루션 (hoh@spss.co.kr)
** 한양대학교 경영대학 경영학부 (kjw@hanyang.ac.kr)

Abstract

본 논문은 데이터 마이닝의 기법 중 가장 잘 알려진 지도학습 기법의 성능 향상을 위한 새로운 Hybrid 및 Combined 기법인 불일치 패턴 모델(오차 패턴 모델)에 대한 연구 논문이다. 불일치 패턴 모델이란 2개 이상의 기법 중 향후 더 레코드별로 더 잘 맞출 수 있는 기법을 메타 분류하는 불일치 패턴 모델을 개발하여, 최종적으로는 기존의 기법보다 더 좋은 분류 정확도 및 예측 향상율을 기대하기 위한 기법을 의미한다. 본 논문에서는 의사결정나무 추론 기법인 C5.0과 C&RT 그리고 신경망 분석, 그리고 로지스틱 회귀분석과 같은 대표적인 데이터 마이닝의 지도학습 기법을 이용하여 불일치 패턴 모델을 생성하여 보고, 이들이 기존 단일 기법과 기존의 Combined 모델인 Bagging, Boosting 그리고 Stacking 기법보다 성능이 우수함을 23개의 실제 데이터 및 공신력 있는 공개 데이터를 이용하여 증명하여 보였다. 또한 데이터의 특성에 따라서 불일치 패턴 모델의 성능의 변화 및 더 우수해 지는지를 알아보기 위한 연구도 같이 수행을 하여 본 모델의 활용성을 높이고자 하였다.

1. 서론

1.1 논문의 배경

1980년대 컴퓨터의 데이터 저장 능력이 발전하면서 성장하면서, 이에 대한 데이터 분석 역시 대용량 데이터를 효율적으로 분석할 수 있는 것이 필요하게 되었고, 이를 위해 생겨난 것이 데이터 마이닝(Data Mining)이다[조용준, 2003]. 데이터 마이닝은 KDD(Knowledge Discovery in Database)라는 개념에서 나온 것으로 대용량의 자료를 취급하는 데이터베이스와 매우 밀접한 관련이 있다는 것을 알 수 있다. 이는 다시 말해서, 실제 기업이나 조직의 현실에서 지식이 되는 실용적인 패턴(Pattern)을 찾아내는 것이라고 할 수 있다. 실제 상황을 중시하는 만큼 각종 분석 상황이 변할수록 데이터 마이닝은 이에 맞추기 위해 급속적인 발전을 계속해오고 있으며, 미래에 대한 잠재력을 계속 보여주는 방향으로 계속적으로 기법들이 진화 발전해 오고 있다[Han and Kamber, 2001]. 또한 근래에 들어 데이터 마이닝은 CRM(Customer Relationship Management)의 주요한 수단[강명구 외, 2001]으로 인식되면서, 더욱 실용적인 검증과 특히 각종 기법 성능의 향상을 계량적인 수단을 통해서 확인하는 것이 필수적인 것으로 되고 있다. 데이터 마이닝은 통계분석과 인공지능 분석 그리고 데이터베이스(database)의 각종 기법이 혼합된 분야이다[Quinlan, 1993]. 그런 만큼 다양한 분석 기법이 존재하는데, 그 중 가장 널리 활용되는 것이 지도학습 기법(Supervised Learning)이다. 의사결정나무 추론(Tree Induction)과 신경망(Neural Networks)분석 그리고 통계분석 방법인 로지스틱 회귀분석(Logistic Regression) 등이 대표적인 이 지도학습 기법은 데이터 마이닝의 전 분야에서 가장 많이 활용되는 분석 방법[www.kdnuggets.com]이며, 이 지도학습 기법의 성능을 향상시킬 수 있는 다양한 비교 분석과 방법의 개발은 데이터 마이닝의 성공적인 수행과 적용에 큰 영향을 미친다는 것은 예상이 쉽게 될 수 있다.

1.2 연구 범위와 목적

데이터 마이닝에서 지도학습 기법을 수행하는 가장 큰 목적은 목적변수를 잘 분류하거나 예측하는 것이다. 이 지도학습기법의 성능이라고 말할 수 있는 분류 정확도와 예측력을 높이기 위한 수많은 연구가 있었다. 그 중 한 방법이, 기존의 기법을 여러 형태로 결합 및 혼합시킨 Combined 모델과 Hybrid 모델이라고 할 수 있다. 그 대표적인 기법으로 Bagging[Breiman, 1996], Boosting[Freund and Schapire, 1996] 그리고 Stacking[Wolpert, 1992]등이 있다. 본 논문은 지도학습 기법의 성능을 향상시키는 방법으로 새로운 Hybrid 및 Combined 모델인 "불일치 패턴 모델" 이라는 것을 소개하고, 이 방법을 여러 가지 다양한 데이터 집합들에 적용하여, 단일한 기법 또는 유사한 Combined 기법보다 성능이 향상됨을 보이고, 또한 한계점과 어떤 데이터 형태 및 기법에서 더욱 효율적인지를 연구하여 실제 데이터 마이닝 적용 분야에서 적극적으로 활용하고자 하는 목적을 가지고 있다. 불일치 패턴 모델은 아래의 3장에서 자세히 설명하겠지만, 예를 들어 2개 이상의 상이한 지도학습 기법들에서, 동일한 데이터에 각 기법들을 적용하면, 공통으로 맞추는 사례(Case)가 있고 서로 다른 결과를 내는 사례가 있기 마련인데, 이 중 서로 상이한 결과를 발생시키는 사례들을 모아 새로운 데이터 집합을 만들고, 적용한 상이한 기법 중 어느 기법이 더 잘 맞추는가를 판별한 불일치 패턴 모델을 별도로 만들어, 향후 최종 적용하는 실제 데이터 또는 Test 데이터 집합에 적용하여, 더 좋은 성능 향상(분류도 또는 예측도)을 꾀하고자 하는 기법이다. 불일치 패턴 모델이라는 명칭은 2개 이상의 지도학습 기법에서 서로 불일치가 되는 데이터만을 모으고, 이렇게 불일치가 발생하는 원인이 무엇인지를 찾아내는 패턴을 찾아서 모델을 만든다는 의미에서 붙여졌고, 정확하게는 불일치 원인 패턴 모델이라고 할 수 있으며, 이를 줄여 불일치 패턴 모델이라고 하였다.

본 논문에서는 불일치 패턴 모델이 기본적인 단일 기법이나 현재 알려진 다양한 결합 방법들보다 성능이 향상되는지를 다양한 각도로 살펴보고, 어떤 데이터에서 불일치 패턴 모델의 성능 향상이 될지를 살펴보고, 그 외 다른 한계점을 살펴보기 위해서 다음과 같은 사항을 실험을 통하여 분석하고자 한다.

첫째 소개되는 불일치 패턴 모델은 내부에 결합된 단일 지도학습 기법의 모델과 비교하여 예측 및 분류 정확도에서 성능 향상이 이루어지는가?

둘째 불일치 패턴 모델은 기존의 데이터 마이닝 Combined 모델인 Bagging, Boosting, Stacking과 방법과 비교하여 성능 향상이 이루어지는가? 그리고 이들이 불일치 패턴 모델에서 통합적 활용을 하는 경우에는 성능 향상이 이루어지는가?

셋째 불일치 패턴 모델은 동시에 적용되는 기법 수가

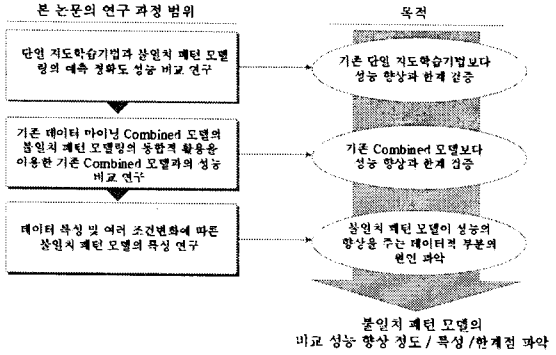
증가할수록 더 성능이 좋아지는가?

넷째 어떤 성격을 가진 데이터에서 정확도의 향상을 보이고, 데이터의 변화에 따라 어떻게 불일치 패턴 모델이 변화가 되는가?

위의 4가지를 파악하는 것이 본 연구의 세부적인 목적이고, 연구 범위가 될 것이다. 이와 같은 세부 목적을 통해서, 다른 일반적인 데이터 마이닝의 기법들과 비교하여 불일치 패턴 모델의 성능향상 정도와 성격 그리고 한계점을 파악하게 될 것이다.

본 논문에서 소개하는 불일치 패턴 모델은 기존의 데이터 마이닝 지도학습 기법의 각종 분류 정확도와 예측도를 향상시키는 것과는 다른 방법이기에 아직 관련된 많은 연구가 있지 못한 분야이다. 따라서 기존에 있던 모델이거나 발전된 모델인 경우 다양한 연구 방향과 확장된 연구가 가능하겠지만, 아직은 그렇지 못하므로 본 논문에서는 위에서 언급한 목적을 달성하고 불일치 패턴 모델을 일반화시킬 수 있는 점을 중심으로 분석 정리할 수 있도록 연구의 범위와 목적을 다음의 <그림 1>과 같이 도식화하여 정리하여 보았다.

1.3 논문의 구성



<그림 1> 본 논문의 연구 범위 도식화

이러한 연구 목적의 달성을 위해서, 본 논문은 다음과 같이 구성을 하였다. 제 1장 서론에서는 연구의 배경과 목적 및 본 논문의 핵심이 되는 지도학습 기법에 대한 설명과 본 논문의 목적 및 의의(意義)을 기술하였고, 제 2장에서는 Hybrid 모델과 Combined 모델의 개념과 이를 이용한 관련된 연구에 대한 정리를 하고, 추가하여 본 논문에서 계속 활용되어질 기존의 Hybrid 및 Combined 모델의 기본적인 이론에 대한 소개를 하고자 한다. 제 3장에서는 본 논문의 이론적 알고리즘인 불일치 패턴 모델에 대한 설명과 이 불일치 패턴 모델에 Bagging, Boosting, Stacking을 통합하여 활용하는 불일치 패턴 모델 방법 그리고 마지막으로 동시에 3개를 사용하는 불일치 패턴 모델의 알고리즘에 대한 설명을 할 것이다. 제 4장에서는 본 연구를 수행하는 실험의 기본적인 가정과 실험에 사용할 실제 데이터 집합 15개와 UCI Machine Learning Repository[www.ics.uci.edu] 데이터를 비롯한 공개된 데이터 집합 8개 등 총 23개의 데이터에 대한 설명을 하고, 5장에서 실험을 할 실험 설계안을 정리할 것이다. 다음 제 5장에서는 설계되어진 계획대로, 각종 실험을 수행하고, 이에 대한 분석을 하여 결과를 도출하고, 동시에 모델의 한계점에 대한 고찰을 하고자 한다. 마지막으로 6장에서는 전체 실험을 통해 나온 최종 결론을 내린 다음 본 논문의 모델이 응용이 잘 될 수 있는 분야에 대한 정리와 함께 향후에 필요한 연구 과제에 대한 정리를 하고자 한다.

2. 관련연구

데이터 마이닝 분야에서 Hybrid 모델이나 Combined 모델을 이용하여, 각 분야 및 일반적으로 성과를 거둔 연구는 매우 활발하게 이루어지고 있으며, 그 유형별로도 다양하다. 먼저 2개 이상의 상이한 기법을 혼합하여 만든 Hybrid 모델을 가지고 모델의 각종 성능 개선을 유도한 연구를 살펴보면, 먼저 Coenen et al.[2000] 등은 C5.0 알고리즘을 이용하여, Direct Mail 반응 예상고객에 대하여, 초기 분류자(Classifier)를 생성하고, 사례기반 추론(Case-based reasoning) 기법을 이용하여 그 순위를 매기는 Hybrid 모델을 사용하여 캠페인의 응답율이 향상하였다고 보고하였다. 다음 Carvalho and Alex[2004]의 경우 C5.0을 이용한 일반적인 의사규칙 Rule을 생성하고, 거기에 Genetic 알고리즘을 융합한 새로운 Hybrid 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 수행하였는데, 이 연구에서 HGT (Hierarchical Generalization Trees) 방법을 이용하여 기초 Rule을 생성하고, Deviation Detection을 이용하여, 단계적으로 불필요한 정보를 제거하는 방식의 Hybrid 모델을 제안하였다. 또한 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 데이터에서 범주형 변수인 경우에는 연관성 분석을 이용하여, Tree 모델을 만들어 내고, 연속형 변수인 경우에는 별도의 변환 없이 바로 Tree 모델을 만들어 이들을 이용한 Hybrid 모델을 만들어, 학생들의 학습 능력 향상 사례에 적용하는 연구를 수행하였다. 의사결정나무 추론을 이용한 또 하나의 Hybrid 방법으로 Lu, Setiono and Liu[1996]는 신경망 기법을 적용할 때 입력변수의 선택 등을 의사결정나무 추론을 이용하였다. 또한 국내에서는 이극노와 이홍철[2003]이 의사결정나무 추론 기법인 C4.5를 이용하여, 주요 설명변수를 도출해 내고, 다음 이 변수들을 사용하여, 신경망 분석을 수행하는 모델을 이동통신 고객 분류에 적용하여 성과를 나타내었으며, 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 Hybrid 모델을 제안하였다. 위에서 언급한 연구 사례들은 주로 Hybrid 방법을 순차적으로 사용한 예라고 할 수 있다. 즉, 어느 한 가지 방법을 수행하고 난 다음 다른 방법을 사용하는 것을 의미한다. 이와는 달리 Hybrid 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 Hybrid 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 Hybrid 모델을 연구하여 분류 정확도를 향상시키었다고 연구하였다. 국내에서는 김진성[2003]이 지도학습 기법인 Fuzzy 신경망에 연관성 분석을 결합한 모델을 연구하기도 하였다. 이렇게 Hybrid 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 Ensemble 학습은 단일한 하나의 Ensemble 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 Hybrid 모델의 개발과 비교에 관한 연구 외에, Hybrid 모델 또는 Combined 모델 자체에 대한 연구도 많이 이루어졌다. Kuncheva, Bezdek, and Shutton[1998]은 Hybrid 모델을 이용하여 예측력이 향상된 사례 자체들을 연구하였고, Suh[1999] 등은 RFM, 로지스틱 회귀분석, 신경망 모델을 가지고 기법 간에 상관성이 낮을수록 Hybrid 모델의 성능이 더욱 더 좋아진다는 연구를 수행하였다. Zhang and Zhang[2004]은 자신들의 저서에서 모든 데이터 셋에 Hybrid 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 Ensemble 에서 전체(all) 네트워트를 사용하여, Ensemble을 만드는 것보다 선택된 몇 가지 네트워크만을 이용하여 만드는 Ensemble 모델이 더욱 더 효율적임을 언급하고, 이 방법을 GASEN(Genetic Algorithm based Selective ENsemble)이라는 이름을 붙이기도 하였다. 또한 Webb and Zheng[2004]은 Ensemble 모델을 여러 개 사용하는 전략이 단순한 Ensemble 모델보다 더 오류(Error)를 감소시킬 수 있다는 것에 대한 전체적인 연구를 수행하였다.

3. 제안 모델링 알고리즘

3.1 불일치 패턴 모델을 이용한 Hybrid 모델

본 장에서는 제시하고자 하는 불일치 패턴 모델을 이용한 Hybrid 모델을 기술하도록 한다. 예를 들어, 지도학습 알고리즘을 이용하는 데이터 마이닝에서 어떤 분석 데이터가 10건이 있고, 어떤 A기법(예를 들어 의사결정나무 추론 기법)을 이용하여 모델을 만들었다고 가정하자. 그리고 그 모델을 시험용 데이터를 이용하여 검정한 결과 정확도가 70%(7건) 그리고 오분류 또는 예측을 잘못한 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 잘못한 것이 30%(3건)이었다고 가정한다. 이렇게 단순한 요약 정보만 있다면, 이 2개의 모델은 동일한 성능을 가졌다고 판단할 수 있다. 하지만 다음 <그림 2>와 같은 경우를 생각해 본다.

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 2> Hybrid 기법의 예측값 선택 과정

<그림 2>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 2>의 가장 우측의 결과처럼 정확도는 80%로 올라가게 될 것이다.

본 연구에서 제시하는 불일치 패턴 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 불일치 패턴 모델을 생성한 다음, 실제 적용할 데이터 집합에서는 각 사례에 대하여 <그림 2>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 불일치 패턴 모델의 기본 과정

구체적으로 불일치 패턴 모델을 이용하여 Hybrid 모델을 만드는 과정을 정리하면 다음과 같다.

먼저 전체 훈련용 데이터 집합을 $L=\{(y_n, x_n), n=1,2,....N\}$ 이라고 한다. 여기서 y_n은 목적 변수를 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 $L_t=\{(y_t, x_t), t=1,2,....T\}$ 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.

$$L_1=\{(y_m, x_m), m=1,2,....M\},$$
$$L_2=\{(y_p, x_p), p=1,2,....P\}$$
단, $M+P=N$

본 단계에서는 처음에 나눈 훈련용 데이터 L을 임의추출을 통해 다시 2개의 훈련용 데이터 집합으로 분리하는 단계로, 이는 내부적으로 또 하나의 모델인 불일치 패턴 모델을 만들기 위하여, 훈련용 데이터를 다시 L_1과 L_2 데이터 집합으로 분할을 하는 것이다.

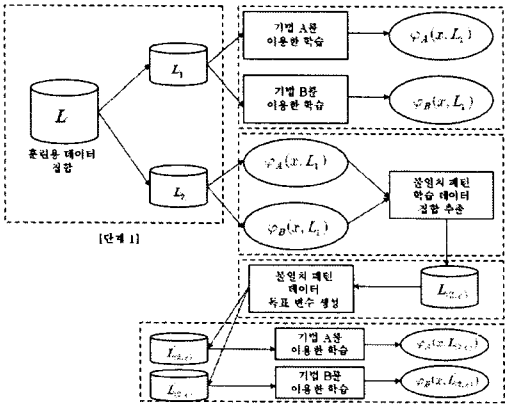
[단계 2] 먼저 L_1 데이터 집합을 이용하여, 분석 기법 A를 이용하여, 모델링을 수행한다. 이 때 모델링을 통해서 생성한 기법 A의 로직, 즉, Predictor를 $\varphi_A(x, L_1)$이라고 한다. 다음 동일한 데이터 집합에 A와는 다른 분석 기법 B를 이용하여 모델링을 수행한다. B기법을 이용하여 모델링을 수행하여 나오게 되는 로직을 $\varphi_B(x, L_1)$이라고 한다. 본 단계는 1 단계에서 나누어진 2개의 훈련용 데이터 중 하나를 불일치 패턴 모델링을 만들기 위해 훈련용 데이터로 지정을 하고, 기법을 적용시킨 것이며, 이 단계를 통해서 나온 로직을, 남아 있는 다른 데이터 집합에 적용하여, 결과를 산출하는 것이 다음 단계이다.

[단계 3] 다음으로 1 단계에서 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 형태가 될 것이다.)를 $T_{(A, L_2)}$라고 하자. 마찬가지로 기법 B를 적용시켜서 나온 결과를 $T_{(B, L_2)}$이라고 하자. 데이터 집합 L_2를 통해서 나온 2개의 결과값을 서로 비교하여 결과 값이 서로 틀린 데이터 집합만을 추출한다. 이것은 <그림 5>에서 번호 3번, 9번 데이터만 추출하는 것과 동일하다고 할 수 있다. 이렇게 추출해 낸 데이터 집합을 $L_{(2,e)}$라고 정의한다. 즉, $L_{(2,e)}=\{(y_i, x_i) \mid (y_i, x_i) \in L_2 \text{ and } \varphi_A(x_i, L_2) \neq \varphi_B(x_i, L_2)\}$ 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 목적변수 y_i값과 기법 A를 이용하여, 생성된 Predictor

$\varphi_A(x_i, L_1)$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 Predictor $\varphi_B(x_i, L_1)$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 Predictor를 $\varphi_A(x, L'_{(2,e)})$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) 또는 불일치 원인 패턴 모델(Inconsistent Cause Pattern Model)이라고 정의한다. 이 5단계까지의 과정을 도식화하면 다음의 <그림 3>과 같다.



<그림 3> 불일치 패턴 모델의 생성과정 도식화

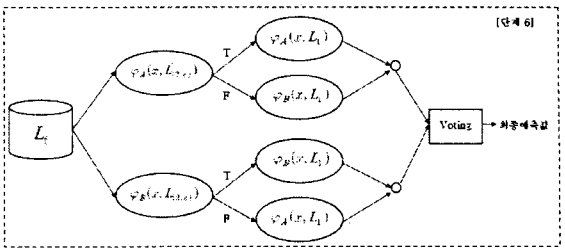
[단계 6] 다음 이렇게 불일치 패턴 모델(또는 불일치 원인 패턴 모델)을 구했으면, 이를 적용한 최종 Predictor를 생성하게 되는데 이 과정은 Voting 방법을 이용한다. 예를 들어서 시험용 데이터 집합인 $L_t = \{(y_t, x_t), t = 1,2,\ldots, T\}$에서 먼저 $\varphi_A(x, L'_{(2,e)})$의 로직을 적용하여, 예측값이 T가 되는 사례는 $\varphi_A(x, L_1)$ 로직을 이용한 결과 값을 선택하고, 아닌 것은 $\varphi_B(x, L_1)$로직을 이용한 결과 값을 선택한다. 이렇게 판별하여 생성된 Predictor를 $\varphi_{(A,B)}(x, L)$ 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 Predictor를 $\varphi_{(B,A)}(x, L)$라고 한다. 이들이 각기 다를 수 있으므로 안정적인 Predictor를 생성하기 위해 이 2개의 계산된 Predictor들에게서 확률값이 큰 쪽을 선택하여 만들어낸 최종 Predictor인 $\varphi_{<A,B>}(x, L)$를 생성하게 되면, 모든 과정이 완료된다. 본 내용을 예제 데이터를 사용하여 표현하면 다음 <표 1>, <표 2>와 같다. 또한 이를 도식화 하면 <그림 4>와 같다.

<표 1> 시험용 데이터에서 불일치 패턴 모델의 판별 결과와 두 기법의 Predictor 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 Voting한 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 4> 불일치 패턴 모델의 적용을 위한 최종 Voting 과정

3.3 기존의 데이터 마이닝 Combined 모델을 통합한 불일치 패턴 모델

위의 제 2절에서는 가장 간단한 형태의 불일치 패턴 모델을 설명하였다. 불일치 패턴 모델은 기존에 단일 지도학습 기법(예를 들어 C5.0, 신경망)뿐만 아니라 데이터 마이닝에서 잘 알려진 Combined 모델도 불일치 패턴 모델에서 하나의 내부 모델로 활용할 수 있다. 본 논문에서는 제 2절에서 언급한 불일치 패턴 모델에서 사용되는 기법 중 1개 또는 2개 모두를 단순한 일반 기법을 사용하는 것이 아니라, 많은 연구에서 그 우수성이 입증된, Bagging, Boosting, Stacking과 같은 Combined 모델을 통해 최종 Predictor를 생성하고, 이것을 가지고 불일치 패턴 모델을 만드는 방식으로 불일치 패턴 모델을 데이터 마이닝의 기존 Combined 모델과 통합적으로 활용하여 보고, 그 성능을 연구하고자 한다. 그 통합적 활용 방법은 다음과 같다.

3.3.1. Bagging과 불일치 패턴 모델의 통합

먼저 앞의 과정에 [단계 1]과 [단계 2]까지는 동일하다. 본 단계는 전체 데이터 집합 L을 2개의 훈련용 데이터와 1개의 검증용 데이터로 분할을 하는 과정이었다. 다음 [단계 3]에서 분석 기법 A를 그대로 적용하여 기법 A의 Predictor인 $\varphi_A(x, L_1)$를 생성하는 것이 아니라 Bagging을 이용한 새로운 Predictor인 $\varphi_{A_{Bagging}}(x, L_1)$과 을 만든다. 마찬가지로 다른 분석 기법 B의 경우도 해당 기법을 그대로 적용하는 것이 아니라 Bagging을 적용하여 $\varphi_{B_{Bagging}}(x, L_1)$과 을 만든다. Bagging은 다음과 같이 개발한다.

[Bagging 1단계] 먼저 데이터 집합 L_1을 5배 Over Sampling을 수행한다. 이렇게 만들어진 데이터 집합을 L'_1이라고 한다.

[Bagging 2단계] 다음 L'_1 집합에서 임의의 난수를 가지는 변수를 생성시킨 다음 이들을 난수의 순서대로 정렬을

한다.

[Bagging 3단계] 난수의 순서대로 정렬된 자료 L_1'을 5등분한 데이터 집합 $L_{1,1}' \sim L_{1,5}'$를 만든다. 즉, $L_1' = L_{1,1}' + L_{1,2}' + L_{1,3}' + L_{1,4}' + L_{1,5}'$가 된다. 이렇게 함으로써, 원래 훈련 데이터 L_1의 1개 레코드가 Over Sampling한 L_1' 데이터 집합에 나타나는 회수는 이항분포 B(5,0,2)를 따르게 되고, 이는 Bagging이 요구하는 포아송 분포, Poisson(1)과 거의 확률값이 비슷해지게 된다[허명회, 2004].

[Bagging 4단계] 이렇게 만들어진 5개의 데이터 집합 $L_{1,1}' \sim L_{1,5}'$에 각각 지도학습 기법 A와 B 를 수행시킨다.

[Bagging 5단계] 먼저 5개의 데이터 집합에 A기법을 적용하면 이를 통해서 Predictor가 $\varphi_A(x, L_{1,1}')$부터 $\varphi_A(x, L_{1,5}')$ 등 5개가 나오게 된다. B기법을 적용해도 동일하다.

[Bagging 6단계] 이렇게 만들어진 5개의 Predictor로부터 나오는 결과를 Voting하면 최종적으로 A기법(또는 B기법)을 Bagging을 수행한 Predictor가 완성된다. 즉, $\varphi_{A_{Bagging}}(x, L_1)$ = Voting($\varphi_A(x, L_{1,1}')$, $\cdots$, $\varphi_A(x, L_{1,5}')$)이 된다. $\varphi_{B_{Bagging}}(x, L_1)$도 마찬가지 방법으로 생성을 한다.

이렇게 Bagging을 적용한 Predictor가 만들어지면, 2절의 [단계 4]부터 발생되는 모든 기법 A와 B 의 Predictor인 $\varphi_A(x, L.)$과 $\varphi_B(x, L.)$는 전부 $\varphi_{A_{Bagging}}(x, L.)$과 $\varphi_{B_{Bagging}}(x, L.)$을 적용하면 된다.

3.3.2. Boosting과 불일치 패턴 모델의 통합

Boosting을 이용한 확장 또한 Bagging을 이용한 확장과 절차와 프로세스는 같다. 다만 생성하는 최종 Predictor가 Bagging이 아닌 Boosting 방법이라는 것이 다를 뿐이다. 거기에 주로 의사결정나무 추론에서 사용되는 Boosting은 기본적으로 Bagging과 같이 여러 개의 Predictor들을 결합한다는 것에서 그 개념 또한 유사하다고 할 수 있다. Bagging과 다른 점은 Predictor들을 생성하고, 초기 가중치를 준 다음 오분류율을 확인하여 그 가중치를 변화시킨 다음 생성된 Predictor들을 최종적으로 Voting을 한다는 것이 동등한 가중치를 주는 Bagging과 약간 다를 뿐이다. 가중치를 주는 방법은 Ada Boost.M1[1998]이라는 Boosting 알고리즘이 가장 정리가 잘 되어 사용되어지고 있으며 본 논문에서도, 이를 이용하였다. 이에 대하여 설명하면 다음과 같다.

[Boosting 가중치 생성 1단계] 앞으로 발생할 Predictor $\varphi_A(x, L_{r1})$ 들에게 초기 가중치 $w_i = \frac{1}{n}(i = 1, \cdots, n)$를 부여한다.

[Boosting 가중치 생성 2단계] 먼저 A기법을 이용하여 첫 번째 Predictor들을 생성한다. 즉, 10개의 Predictor들인 $\varphi_A(x, L_{r1})$ 을 생성한다.

[Boosting 가중치 생성 3단계] 다음 $\varphi_A(x, L_{r1})$의 오분류율인 err_1을 생성한다. 그리고 이것을 이용하여 다음과 같은 가중치를 생성한다. $\alpha_1 = \log(1 - err_1)/err_1$을 생성한다.

[Boosting 가중치 생성 4단계] 최초의 가중치 w_1을 다음과 같이 바꾼다.

$w_1 \leftarrow w_1 \cdot \exp(\alpha_1)$

[Boosting 가중치 생성 5단계] 이렇게 가중치가 적용된 것을 10회 반복하여 수행하고 이들을 Voting하여 최종 A기법을 이용한 Boosting 알고리즘 $\varphi_{A_{Boosting}}(x, L_1)$을 생성한다. B 기법도 방식은 동일하게 하여 $\varphi_{B_{Boosting}}(x, L_1)$을 생성한다. 단, 여기서 10회인 것은, 이는 향후 본 알고리즘을 검정 수행하는데 사용될 데이터 마이닝 솔루션인 Clementine 8.1[SPSS inc., 2003] 버전의 기본 설정값이 10회이고, 또한 R. Quinlan이 여러 Combined 모델과 자신이 제안한 C4.5를 비교한 논문에서는 Boosting의 기준을 10회로 한 다음 비교하여[Quinlan, 1996] 그에 따라서 기준을 잡았다. 이 이후의 작업은 Bagging과 동일하다. 즉, 불일치 패턴 모델의 단계 (4)부터 발생되는 모든 기법 A와 B 의 Predictor인 $\varphi_A(x, L.)$과 $\varphi_B(x, L.)$는 전부 $\varphi_{A_{Boosting}}(x, L.)$과 $\varphi_{B_{Boosting}}(x, L.)$을 적용하면 된다.

3.3.3. Stacking과 불일치 패턴 모델의 통합

Stacking을 이용한 확장 방법은 Bagging 또는 Boosting과는 약간 절차가 다르다. 일단 2절의 [단계 3]까지는 전부 동일하다. [단계 4]도 전부 동일하게 수행한 다음 1가지를 더 수행한다. 앞의 [단계 4]에서 즉, 훈련용 데이터 집합 L_2에 적용한 두 기법의 예측 로직인 Predictor $\varphi_A(x, L_1)$과 $\varphi_B(x, L_1)$를 적용시켜서. 먼저 기법 A를 적용시켜서 나온 예측 결과를 $T_{(A, L_2)}$라고 하자. 그리고 기법 B를 적용시켜서 나온 결과를 $T_{(B, L_2)}$이라고 정의하였는데, 이들은 예측값과 예측 확률로 구성된 필드가 된다. 이 필드를 다시 데이터 집합 L_2에서 설명변수로 추가로 사용을 하는 것이다. 그리고 다시 이들 기법 A를 적용시켜서 나온 $T_{(A, L_2)}$에서 나온 필드 2개와 B를 적용시켜서 나온 결과 $T_{(B, L_2)}$에서 나온 필드 2개를 기존 데이터 집합 L_2에 결합을 한다. 이것을 데이터 집합 L_2'이라고 하자. 즉, $L_2' = \{(y_i, x_i) | (y_i, x_i) = L_2 + T_{(A, L_2)} + T_{(B, L_2)}\}$이다. 단 목적변수는 L_2와 L_2'가 동일하다. 새롭게 만들어진 데이터 집합 L_2'에 다시 기법 A(또는 B)를 기존 적용시켜서 나온 Predictor가 바로 Stacking을 이용하여 생성된 Predictor인 $\varphi_{A_{Stacking}}(x', L_2')$ (또는 $\varphi_{B_{Stacking}}(x', L_2')$)이다. 이후에 2절의 [단계 5]부터 발생되는 모든 기법 A와 B 의 Predictor인 $\varphi_A(x, L.)$과 $\varphi_B(x, L.)$는 전부 $\varphi_{A_{Stacking}}(x', L.)$과 $\varphi_{B_{Stacking}}(x', L.)$을 적용하면 된다.

4. 실험의 설계

4.1 실험의 가정

3절에서 설명한 불일치 패턴 모델과 기존의 Combined 모델과 통합한 불일치 패턴 모델의 효율성에 대한 실험을 하기 위하여, 다음과 같은 가정을 둔다.

첫째 실험 데이터의 목적 변수는 전부 이분형(binary type)데이터를 선정하였다. 범주가 3개 이상의 경우도 가능하기는 하지만, 데이터 마이닝에서 가장 목적변수로 가장 많이 사용되는 것이 이분형 형태이고, 우선 본 논문의 목적은 불일치 패턴 모델을 이용한 모델이 비교 대상에 비하여 성능의 향상이라는 목적에 맞추기 위하여, 가능한 외생적인 변수를 통제하고, 해당 효과만을 살펴보기 위해서이다. 또한 목적변수가 3개 이상인 것들은 이분형으로 통합을 할 수도 있어서, 더 많은 이분형 데이터를 이용한 실험을 통해 모델의 신뢰도가 높아지면, 본 논문이 더욱 의미가 있을 것으로 판단된다. 둘째 본 논문에서 기초로 사용한 기법으로 신경망의 MLP(Multi-Layer Perceptron) 방법과 의사결정나무 추론 기법의 C5.0, C&RT 그리고 통계적인 방법으로는 로지스틱 회귀분석과 판별분석을 선정하였다. 이들 기법을 선정한 것은 신경망 분석과 의사결정나무 추론 기법, 로지스틱 회귀분석 등이 지도학습 기법 중 데이터 마이닝을 대표하는 기법[Entrue 컨설팅 CRM그룹, 2000]이기도 하고, 특히 C5.0 알고리즘은 다른 의사결정나무 추론 기법보다도 효율적이기 때문이다[Quinlan, 1996]. 그리고 신경망 분석이나 의사결정나무 추론 분석 이외에 통계적인 성격의 로지스틱 회귀분석이나 판별분석 등은 데이터 상에서 가지고 있는 결측치 또는 분석의 성격 상 일반적인 가정에 부합하는 경우에만 실험을 하여서, 전체 데이터 집합에 적용하지 못하였다. 셋째 이곳에서 적용하는 모든 기법들의 다른 옵션(option)들은 모두 데이터 마이닝 솔루션에서 제공하는 기본 옵션에서 변경을 시키지 않았다.

본 실험에 사용된 데이터 마이닝 Tool은 SPSS[1]社의 Clementine 8.1을 사용하였다. 이렇게 잘 알려진 Tool을 이용한 것은 개인이 각종 알고리즘을 프로그래밍하면서 발생할 수 있는 오류를 사전 방지하고, 다른 연구자의 실험 재현성을 위해서 이다. 다음의 <표 3>은 금번 논문에 사용된 모델링 기법을 정리한 것이다.

<표 3> 본 논문에서 비교되는 분석 기법

번호	기법	내용	비고
1	신경망	MLP만 이용하여 결과 생성	단일 기법
2	C5.0	C 5.0 기본 옵션만 이용하여 결과 생성	단일 기법

1) SPSS Inc. (http://www.spss.com)

3	C&RT	C&RT만 이용하여 결과 생성	단일 기법
4	신경망 + C5.0 불일치 패턴 모델	신경망과 C5.0을 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
5	신경망 + C&RT 불일치 패턴 모델	신경망과 C&RT를 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
6	C5.0 + C&RT 불일치 패턴 모델	C5.0과 C&RT를 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
7	로지스틱 회귀 모델	로지스틱 회귀분석만 이용하여 결과 생성	단일 기법
8	로지스틱 + 신경망 불일치 패턴 모델	로지스틱 회귀분석과 신경망을 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
9	로지스틱 + C5.0 불일치 패턴 모델	로지스틱 회귀분석과 C5.0을 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
10	로지스틱 + C&RT 불일치 패턴 모델	로지스틱 회귀분석과 C&RT을 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
11	신경망 + C5.0 – 판별분석을 이용한 불일치 패턴 모델	C5.0과 신경망을 이용한 불일치된 사항을 판별분석을 이용하여 만든 불일치 패턴 모델	불일치 패턴 모델 기본형
12	C5.0 + C&RT – 판별분석을 이용한 불일치 패턴 모델	C5.0과 C&RT을 이용한 불일치된 사항을 판별분석을 이용하여 만든 불일치 패턴 모델	불일치 패턴 모델 기본형
13	C&RT + 신경망 – 판별분석을 이용한 불일치 패턴 모델	신경망과 C&RT을 이용한 불일치된 사항을 판별분석을 이용하여 만든 불일치 패턴 모델	불일치 패턴 모델 기본형
14	C5.0 Bagging	C5.0을 Bagging을 이용한 Combined 모델	기존에 활용되는 Combined 모델
15	C5.0 Boosting	C5.0을 Boosting을 이용한 Combined 모델	기존에 활용되는 Combined 모델
16	C5.0 Stacking	C5.0을 Stacking을 이용한 Combined 모델	기존에 활용되는 Combined 모델
17	신경망 + C5.0 Voting	신경망의 결과와 C5.0의 결과에 대한 단순한 Voting	기존에 활용되는 Combined 모델
18	신경망 + C5.0 불일치 패턴 모델	신경망과 C5.0을 이용한 불일치 패턴 모델	불일치 패턴 모델 기본형
19	신경망 + C5.0 Bagging 불일치 패턴 모델	신경망과 Bagging이 적용된 C5.0 을 이용한 불일치 패턴 모델	기존 Combined모델을 통합 활용한 불일치 패턴 모델
20	신경망 + C5.0 Boosting 불일치 패턴 모델	신경망과 Boosting이 적용된 C5.0 을 이용한 불일치 패턴 모델	기존 Combined모델을 통합 활용한 불일치 패턴 모델
21	신경망 + C5.0 Stacking 불일치 패턴 모델	신경망과 Stacking이 적용된 C5.0 을 이용한 불일치 패턴 모델	기존 Combined 모델을 통합 활용한 불일치 패턴 모델

위의 <표 3>에서 먼저 단일 기법과 불일치 패턴 모델의 성능 비교를 위해서는 C5.0, 신경망, C&RT 방법 3가지의 조합을 이용하여 비교하도록 한다. 다음 기존의 데이터 마이닝 Combined 모델과의 비교 및 통합 활용에서는 3개의 기법 중 의사결정나무 추론의 대표방법인 C5.0과 의사결정나무 추론과는 성격이 이질적인 신경망 분석 2개를 선택하여 불일치 패턴 모델을 만들어 실험을 수행한다. 이렇게 수행하는 이유는 C&RT의 경우 C5.0과 유사한 의사결정나무 추론방법이기도 하고, 선행 연구[Quinlan, 1996]에서 C5.0이 C&RT보다 일반적으로 더 효율적이라는 점과 불일치 패턴 모델의 경우 2개의 기법 중 서로 좋은 부분을 선택하여, 상호 보완하여 성능을 향상시키는 성격의 특성상 성격이 좀 다른 이질적인 2개의 방법을 선택하였다. 여기서 특이한 사항은 Bagging과 Boosting 그리고 Stacking에 의사결정나무 추론기법인 C5.0에만 적용을 한 것이다. 이것은 특히 3개의 Combined 모델 중 Boosting 기법이 의사결정나무 추론 기법에만 적용이 가능하여, 3개의 Combined 모델이 동등한 조건에서 분석을 할 수 있도록 <표 3>과 같이 기법을 정리하였다. 또한 위의 방법 중 신경망과 C5.0의 Voting 방식의 경우 가장 일반적인 Voting 방법인 2개의 기법을 통해 나온 결과 중 예측 확률값이 더 높은 쪽을 선택하는 Voting 방법을 이용하였다.

본 실험에서 가장 중요한 것은 훈련용 데이터 집합을 가지고 만든 모델을 검증용 데이터 셋을 이용하여 최종 예측 및 분류 정확도를 확인하는 것이다. 본 실험에서 예측의 정확도를 계산하는 방법은 일반적인 방법으로 <표 4>와 같다

<표 4> 예측 정확도의 계산 방법

구분		예측	
		T(True)	F(False)
실제	T(True)	참 TP	거짓(FN)
	F(False)	거짓(FP)	참(TN)
* 예측정확도 = (TP+TN)/(TP+FN+FP+TN)			

4.2 실험에 사용된 데이터

본 실험에서 사용될 데이터는 총 23개이다. 앞의 15개는 실제 각종 산업군에서 데이터 마이닝을 수행하면서 사용된 데이터이고, 나머지 8개는 UCI Machine Learning Repository[www.ics.uci.edu] 및 Carnegie Mellon 대학의 AutonLab 실험 데이터[www.qutonlab.org] 등 공개되어져 있는 데이터를 이용하였다. 이렇게 다양한 데이터를 사용하는 것은 기본적으로 가능한 여러 개의 데이터를 이용하여, 불일치 패턴 모델의 성능향상과 다른 기법과의 비교 결과에서 신뢰성을 높이기 위한 것이다. 이 신뢰성은 2가지 방향에서 전부 얻고자 하였는데, 데이터 마이닝의 경우 현실성이 강한 만큼 실제 활용되었던 산업군의 다양한 데이터를 이용하였고, 동시에 학문적인 신뢰성을 확보하기 데이터 마이닝 실험에 가장 많이 활용되는 UCI 데이터 및 공개 데이터도 동시에 사용하여, 현실적인 신뢰성과 학문적인 신뢰성 2가지를 전부 얻고자 하였다. 다음 <표 5>는 실험에 사용된 데이터에 대한 설명이다.

<표 5> 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 5만 달러 이상/ 5만 달러 미만	소득 예상 계층을 분류하는 데이터
3	연속형 데이터 : 연령, 개인평균소득, 자녀수 등 3개 필드 범주형 데이터 : 차량 소유여부, 거주지역, 담	반응/비반응	미국의 어느 잡지 구독 고객을 대상으로 메일 캠페인에 대한 반응 여

	보여부, 저축여부 등 7개 필드		부 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 가입 여부 데이터
7	연속형 데이터 : 주식선물옵션 3개월 평잔, 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(유선 가입에 대한 이탈/유지)	미국의 한 유선 통신회사의 이탈/유지 여부 데이터
11	연속형 데이터 : 부가서비스 요금, 사용기간, 연령 3개 필드 범주형 데이터 : 지역, 성별, 직업유형 등 5개 필드	Churn(무선 통신 서비스에 대한 이탈/유지)	한국의 한 이동 통신사 서비스의 이탈/유지 여부 데이터
12	연속형 데이터 : 연령, 보험가입 기간 등 7개 필드 범주형 데이터 : 성별, 운전여부, 직업 3개 필드	가입 보험의 중도 해지 여부	한국의 한 생명 보험사의 고객 상품 가입 데이터
13	연속형 데이터 : 연령, 총진료비, 입원일수 등 17개 필드 범주형 데이터 : 지역, 성별, 상병유형 등 6개 필드	건강보험료 청구 심사 통과 여부	한국의 한 종합병원의 건강보험료 청구 심사 통과 자료
14	연속형 데이터 : 연령, 월수입 등 3개 필드 범주형 데이터 : 성별, 지불방법, 육류구매여부 등 13개 필드	주택의 소유여부	미국의 한 할인점 고객 정보
15	연속형 데이터 : PHT 공정과정 중 시약 처리 비율 등 17개 필드 범주형 데이터 : 생산장비종류, SLOT 종류 등 6개 필드	생산제품의 불량/양호 여부	한국의 한 LCD공장의 제품 생산 정보
16	연속형 데이터: 특정 지정 단어 e-mail에 포함 수 등 총 57개 필드 범주형 데이터 : 없음	Spam mail / non-spam mail	UCI 데이터 에서 e-mail Spam관련 자료 (Spambase)
17	연속형 데이터: 없음 범주형 데이터: 말의 위	positive / negative	UCI 데이터 에서
	치를 표시하는 9개 필드		tic-tac-toe 게임 관련 자료 (tic-tac-toe)
18	연속형 데이터: 6개 필드 범주형 데이터: 5개 필드	신용도의 양호/불량 여부 (+ / -)	UCI 데이터 에서 일본의 신용 평가 Screening 자료(CRSX)
19	연속형 데이터: 주당 근무시간, 연령 등 6개 필드 범주형 데이터: 결혼상태 등 8개 필드	소득에 대한 예측 >50K, <=50K	UCI 데이터에서 인구센서스 자료 Screening 자료(Adult)
20	연속형 데이터: 범죄율 등 13개 필드 범주형 데이터: 없음	중위수 기준의 주택 가격 $22,000 이상과 이하	UCI 데이터에서 미국 보스턴 집값 관련 데이터 (Housing)
21	연속형 데이터: 전복의 크기, 무게 등 7개 범주형 데이터: 성별 1개	전복 연령 (+1.5세)이 10세 이상과 미만	UCI 데이터에서 전복(abalone) 데이터
22	연속형 데이터: 아이의 수 등 3개 범주형 데이터: 국적, 성별 등 7개	주택 소유여부	MS SQL 2000 예제데이터에서 Food Mart 데이터
23	연속형 데이터: Compressed life science 데이터 10개 필드 범주형 데이터: 없음	정상적인 활동 및 비활동 여부	Carnegie Mellon대학 Auton Lab 데이터

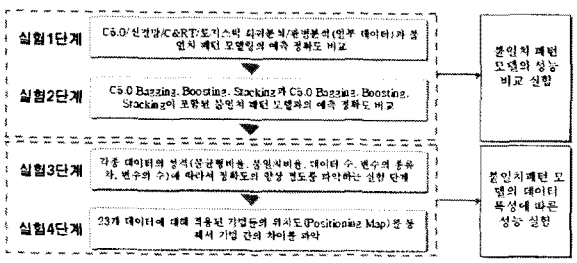
실험결과에 대한 분석을 하기에 앞서서 데이터에 대한 전체적인 상황을 파악하는 것이 필요하다. 다음의 <표 6>은 실험 대상인 23개의 데이터 집합에 모델을 적용하기 전 전체 데이터의 수와 검증용 데이터 수를 정리한 것이다.

<표 6> 실험 데이터의 검정용 데이터와 실험용 데이터의 수

데이터 번호	전체 데이터 수	검증용 데이터 수	훈련용 데이터 수(전체 데이터수-검증용 데이터수
1	14,490	4,765	9,725
2	32,561	10,750	21,811
3	12,000	3,995	8,005
4	3,284	1,079	2,205
5	4,574	1,539	3,035
6	100,000	33,013	66,987
7	228,774	76,243	152,531
8	2,181	754	1,427
9	30,000	10,000	20,000
10	16,004	5,396	10,608
11	3443	1,160	2,283
12	44,942	15,109	29,833
13	16,797	5,572	11,225
14	1,000	333	664
15	5,619	1,843	3,776
16	4,601	1,488	3,113
17	958	326	632
18	690	227	463
19	48,843	16,282	32,561
20	506	173	333
21	4,177	1,362	2,815
22	10,282	3,399	6,883
23	26,733	8,869	17,864

4.3 실험의 구성

본 논문에서의 실험은 크게 2가지의 영역으로 구성하였다. 첫 번째 영역은 기본적으로 본 논문에서 제시한 불일치 패턴 모델이 기존의 데이터 마이닝 지도학습의 단일 기법들과 기존에 잘 알려진 Combined 모델과 불일치 패턴 모델의 성능 비교를 하는 실험이며, 두 번째 영역은 어떤 데이터 특성에서 더 효율적으로 불일치 패턴 모델이 성능을 발휘하는지를 알아보는 것이다. 즉, 기본적인 성능 비교와 성능이 발휘되는 데이터의 특성을 알아보는 것으로 각각 내부적으로는 다시 2가지의 세부적인 실험 단계로 구성을 하였다. 이를 도식화 한 것이 <그림 5>와 같다.



<그림 5> 본 실험의 설계 단계

위의 <그림 5>에서와 같이 전체적으로 성능 비교 실험과 데이터 특성에 따른 성능 실험을 수행하고 세부적으로 6단계의 실험을 수행하여, 최종적으로 정확도의 향상과 데이터 특성에 따른 변화에 대한 결과를 얻고자 한다. 다음 5장에서는 위의 <그림 5>와 같이 실험을 하여, 나온 결과에 대하여, 분석 정리하도록 한다.

5. 실험결과

5.1 단일 지도학습 기법과 불일치 패턴 모델의 비교 실험 결과

위의 <그림 5>에서 실험 설계한 바와 같이 불일치 패턴 모델이 기존의 지도학습기법인 C5.0, 신경망 그리고 C&RT 그리고 로지스틱 회귀분석등과 같은 단일 지도학습 기법과 비교하여, 예측 정확도의 성능 향상이 일어나는지 확인하기 위하여 앞에서 제시한 23개 데이터를 이용하여 실험을 하였다. <표 7>은 먼저 지도학습 기법 중 신경망(MLP)분석, 의사결정 나무 추론 분석인 C5.0과 C&RT의 단일 기법과 이들을 2개씩 조합하여 만든 불일치 패턴 모델을 생성하여 동일한 23개의 데이터를 이용하여 나온 결과이다.

<표 7> 3개의 단일 지도학습 기법과 이를 이용한 불일치 패턴 모델의 실험 결과

데이터 번호	신경망 (가)	C5.0 (나)	C&RT (다)	C5.0&신경망 불일치 패턴 모델(라)	신경망&C&RT 불일치 패턴 모델(마)	C&RT&C5.0 불일치 패턴 모델(바)
1	68.96% T: 3,286 F: 1,479	73.33% T: 3,494 F: 1,271	66.44% T: 3,166 F: 1,599	77.73% T :3,704 F :1,061	73.73% T: 3,513 F: 1,252	75.80% T: 3,612 F: 1,153
2	83.23% T: 8,947 F: 1,803	85.45% T: 9,186 F: 1,564	83.52% T: 8,978 F: 1,772	85.67% T :9,209 F :1,541	84.50% T: 9,084 F: 1,666	85.42% T: 9,183 F: 1,567
3	91.34% T: 3,649 F: 346	94.67% T: 3,782 F: 213	92.07% T: 3,678 F: 317	94.67% T :3,782 F: 213	93.77% T: 3,746 F: 249	94.67% T: 3,782 F: 213
4	66.73% T: 720 F: 359	67.10% T: 724 F: 355	65.25% T: 704 F: 375	72.01% T: 777 F: 302	70.71% T: 763 F: 319	70.06% T: 756 F: 323
5	76.09% T: 1,171 F: 368	76.54% T: 1,178 F: 361	77.26% T: 1,189 F: 350	76.60% T: 1,179 F: 360	77.26% T: 1,189 F: 350	76.74% T: 1,181 F: 358
6	95.14% T: 31,409 F: 1,604	96.01% T: 31,697 F: 1,316	94.28% T: 31,125 F: 1,888	96.22% T: 31,765 F: 1,248	95.05% T: 31,379 F: 1,634	96.14% T: 31,740 F: 1,273
7	80.02% T: 61,003 F: 15,231	98.87% T: 75,371 F: 863	90.09% T: 68,679 F: 7,555	98.90% T: 75,398 F: 836	91.46% T: 69,723 F: 6,511	98.88% T: 75,380 F: 854
8	61.01% T: 460 F: 294	63.40% T: 478 F: 276	63.53% T: 479 F: 275	71.49% T: 539 F: 215	74.01% T: 558 F: 196	68.44% T: 516 F: 238
9	70.30% T: 7,030 F: 2,970	71.65% T: 7,165 F: 2,835	70.49% T: 7,049 F: 2,951	91.10% T: 9,110 F: 890	70.34% T: 7,034 F: 2,966	71.08% T: 7,108 F: 2,892
10	90.77% T: 4,898 F: 498	93.48% T: 5,044 F: 352	88.10% T: 4,754 F: 642	93.64% T: 5,053 F: 343	91.16% T: 4,919 F: 477	93.53% T: 5,047 F: 349
11	84.40% T: 979 F: 181	68.88% T: 799 F: 361	75.26% T: 873 F: 287	86.64% T: 1,005 F: 155	90.43% T: 1,049 F: 111	79.66% T: 924 F: 236
12	74.80% T: 11,302 F: 3,807	78.40% T: 11,845 F: 3,264	76.67% T: 11,584 F: 3,525	83.70% T: 12,646 F: 2.463	76.21% T: 11,515 F: 3,594	78.41% T: 11,847 F: 3.262
13	96.06%	95.98%	95.60%	96.06%	96.05%	95.98%

	T: 5,354 F: 218	T: 5,348 F: 224	T: 5,327 F: 245	T: 5,354 F: 218	T: 5,352 F: 220	T: 5,348 F: 224
14	61.26% T: 204 F: 129	56.76% T: 189 F: 144	57.66% T: 192 F: 141	62.76% T: 209 F: 124	**63.36%** T: 211 F: 122	55.86% T: 186 F: 147
15	75.53% T: 1,392 F: 451	74.17% T: 1,367 F: 476	74.93% T:1,381 F: 462	**75.64%** T: 1,394 F: 4449	75.31% T: 1,388 F: 455	75.20% T: 1,386 F: 457
16	91.87% T: 1,367 F: 121	90.03% T: 1,353 F: 135	88.91% T: 1,323 F: 165	**92.81%** T: 1,381 F: 107	91.06% T: 1,355 F: 133	90.73% T: 1,350 F: 138
17	78.83% T: 257 F: 69	80.06% T: 261 F: 65	80.67% T: 263 F: 63	81.60% T: 266 F: 60	84.05% T: 274 F: 52	**85.28%** T: 278 F: 48
18	83.70% T: 190 F: 37	87.22% T: 198 F: 29	81.06% T: 184 F: 43	87.22% T: 198 F: 29	85.46% T: 194 F: 33	**88.99%** T: 202 F: 25
19	83.91% T: 13,661 F: 2,620	85.86% T: 13,979 F: 2,302	84.70% T: 13,790 F: 2,491	86.06% T: 14,011 F: 2,270	85.79% T: 13,967 F: 2,314	**86.45%** T: 14,075 F: 2,206
20	**89.02%** T: 154 F: 19	83.82% T: 145 F: 28	85.55% T: 148 F: 25	86.71% T: 150 F: 23	88.44% T: 153 F: 20	86.13% T: 149 F: 24
21	**78.85%** T: 1,074 F: 288	76.14% T: 1,037 F: 325	74.74% T: 1,018 F: 344	78.41% T: 1,068 F: 294	78.19% T: 1,065 F: 297	76.36% T: 1,040 F: 322
22	67.78% T: 2,304 F: 1,125	66.93% T:2,275 F:1,124	68.58% T: 2,331 F: 1,068	67.84% T: 2,306 F: 1,093	68.58% T: 2,331 F:1,124	**68.64%** T: 2,333 F: 1,066
23	97.79% T: 8,673 F: 196	97.73% T: 8,668 F: 201	97.22% T: 8,622 F: 247	**97.85%** T: 8,678 F: 191	97.79% T: 8,673 F: 196	97.74% T: 8,669 F: 200

위의 <표 7>을 보면 앞의 단일한 기존의 지도학습 기법보다 불일치 패턴 모델의 성능이 좀 더 우수한 것을 알 수 있다. 이를 정확하게 통계적으로 확인하기 위하여 Wilcoxon의 대응 2표본 부호 순위 검정을 통하여 비교하여 보았다. Wilcoxon의 대응 2표본 부호 순위 검정을 사용한 것은 검정하고자 하는 집단의 개수가 23개 정도이어서, 이들이 구체적인 통계적 분포함수(Distribution Function)를 따른다고 가정하기 어려워 비모수적인 방법인 Wilcoxon의 대응 2표본 부호 순위 검정을 이용한 것이다[송문섭, 박창순, 1989]. 또한 일반적인 대응 t-test(paired t-test)를 이용하여서도 동시에 검정을 하여, 검정 통계량의 따른 차이가 있는지도 확인을 하였다. 검정의 결과는 <표 8>과 같다. Wilcoxon의 대응 2표본 부호 순위 검정 및 대응 t-test 검정 등은 모두 SPSS社의 SPSS 14.0.2 버전을 이용하여 분석 하였다.

<표 8> 3개의 단일 지도학습기법과 불일치 패턴 모델의 Wilcoxon 부호순위 및 대응 t-test 검정결과

대응기법	순위N		근사유의확률 (Z-값)	대응 t-test
신경망(가) vs. 불일치 패턴 모델-C5.0& 신경망(라)	가 > 라	2	0.000 (-3.620)	0.003 (-3.323)
	가 < 라	20		
	가 = 라	1		
신경망(가) vs. 불일치 패턴 모델- 신경망& C&RT(마)	가 > 마	6	0.002 (-3.133)	0.005 (-3.160)
	가 < 마	16		
	가 = 마	1		
C5.0(나) vs. 불일치 패턴 모델-C5.0& 신경망(라)	나 > 라	0	0.000 (-4.015)	0.005 (-3.089)
	나 < 라	21		
	나 = 라	2		
C5.0(나) vs. 불일치 패턴 모델-C5.0& C&RT(바)	나 > 바	4	0.002 (-3.147)	0.013 (-2.702)
	나 < 바	17		
	나 = 바	2		
C&RT(다) vs. 불일치 패턴 모델-신경망& C&RT(마)	다 > 마	2	0.000 (-3.841)	0.001 (-3.864)
	다 < 마	19		
	다 = 마	2		
C&RT(다)vs. 불일치 패턴 모델-C5.0& C&RT(바)	다 > 바	2	0.000 (-3.711)	0.000 (-4.402)
	다 < 바	21		
	다 = 바	0		

<표 8>을 보면 통계적으로 단일기법보다 불일치 패턴 모델의 성능이 유의수준 95%에서 매우 우수한 것으로 나타났다. 이는 일단, 불일치 패턴 모델이 단일한 지도학습 기법을 이용하여, 모델을 만드는 것보다 좀 더 우수한 결과를 나타낸다고 할 수 있다. 또한 일반적으로 한 가지 방법을 이용한 모델보다 여러 기법을 혼합한 Hybrid 모델이나 결합한 Combined 모델의 성능이 우수하다는 여러 선행 연구와도 부합하는 결과라고 할 수 있다.

다음은 지도학습 기법 중 위에서 언급한 신경망 분석이나 의사결정 나무 추론 이외에 통계적인 방법인 로지스틱 회귀분석을 이용하여 불일치 패턴 모델을 생성하여 예측 정확도에 대한 비교 분석을 수행하여 보고자 한다. 그 결과가 <표 9>와 같다.

<표 9> 로지스틱 회귀분석과 신경망, C5.0, C&RT를 이용하여 개발한 불일치 패턴 모델의 비교 실험결과

데이터 번호	로지스틱 회귀분석 (사)	로지스틱 & 신경망 불일치 패턴 모델 (아)	로지스틱& C5.0 불일치 패턴 모델 (자)	로지스틱 & C&RT 불일치 패턴 모델 (차)	비고
1	...	...	...	...	결측치로 인한null 정보가 너무 많음
2	84.66% T: 9,101 F: 1,649	84.47% T: 9,081 F: 1,669	85.73% T: 9,216 F: 1,534	85.04% T: 9,142 F: 1,608	85.04% T: 9,142 F: 1,608
3	82.80% T: 3,308 F: 687	91.56% T: 3,658 F: 337	94.74% T: 3,785 F: 346	93.32% T: 3,728 F: 267	93.32% T: 3,728 F: 267
4	59.04% T: 637 F: 442	66.54% T: 718 F: 361	69.32% T: 748 F: 331	68.12% T: 735 F: 344	68.12% T: 735 F: 344
5	74.92% T: 1,153 F: 386	76.15% T: 1,172 F: 367	76.74% T: 1,181 F: 358	77.00% T: 1,185 F: 354	77.00% T: 1,185 F: 354
6	94.33% T: 31,140 F: 1,873	94.92% T: 31,336 F: 1,677	96.19% T: 31,754 F: 1,259	95.38% T: 31,488 F: 1,525	95.38% T: 31,488 F: 1,525
7	79.21% T: 60,383 F: 15,851	80.82% T: 61,613 F: 14,621	98.89% T: 75,389 F: 845	92.20% T: 70,284 F: 5,950	92.20% T: 70,284 F: 5,950
8	59.95% T: 452 F: 302	65.38% T: 493 F: 261	63.93% T: 482 F: 272	63.66% T: 480 F: 274	63.66% T: 480 F: 274
9	67.46%	81.83%	79.98%	76.38%	76.38%

	T: 6,746 F: 3,254	T: 8,183 F: 1,817	T: 7,998 F: 2,002	T: 7,638 F: 2,362	T: 7,638 F: 2,362
10	88.90% T: 4,797 F: 599	91.14% T: 4,918 F: 478	93.24% T: 5,031 F: 365	89.07% T: 4,806 F: 590	89.07% T: 4,806 F: 590
11	...	...	...	...	결측치로 인한null 정보가 너무 많음
12	72.84% T: 11,005 F: 4,104	78.45% T: 11,853 F: 3,256	78.50% T: 11,861 F: 3,248	78.83% T: 11,911 F: 3,198	78.83% T: 11,911 F: 3,198
13	93.61% T: 5,216 F: 356	96.18% T: 5,359 F: 213	96.30% T: 5,366 F: 206	96.20% T: 5,360 F: 212	96.20% T: 5,360 F: 212
14	55.86% T: 186 F: 147	60.36% T: 201 F: 132	57.36% T: 191 F: 142	59.16% T: 197 F: 136	59.16% T: 197 F: 136
15	...	...	...	...	결측치로 인한null 정보가 너무 많음
16	92.34% T: 1,374 F: 114	92.94% T: 1,383 F: 105	92.54% T: 1,377 F: 111	92.41% T: 1,375 F: 113	92.41% T: 1,375 F: 113
17	95.71% T: 312 F: 14	95.71% T: 312 F: 14	95.71% T: 312 F: 14	94.48% T: 308 F: 18	94.48% T: 308 F: 18
18	...	...	...	...	결측치로 인한null 정보가 너무 많음
19	85.25% T: 13,879 F: 2,402	85.80% T: 13,969 F: 2,312	86.17% T: 14,029 F: 2,252	85.26% T: 13,881 F: 2,400	85.26% T: 13,881 F: 2,400
20	87.28% T: 151 F: 22	89.02% T: 154 F: 19	87.28% T: 151 F: 22	84.97% T: 147 F: 26	84.97% T: 147 F: 26
21	79.66% T: 1,085 F: 277	79.00% T: 1,076 F: 286	77.97% T: 1,062 F: 300	77.97% T: 1,062 F: 300	77.97% T: 1,062 F: 300
22	68.37% T: 2,324 F: 1,075	68.37% T: 2,324 F: 1,075	68.34% T: 2,323 F: 1,076	68.78% T: 2,338 F:1,061	68.78% T: 2,338 F:1,061
23	97.33% T: 8,632 F: 237	97.85% T: 8,678 F: 191	97.76% T: 8,670 F: 199	97.33% T: 8,632 F: 237	97.33% T: 8,632 F: 237

위의 <표 9>에서 보면, 1, 11, 15, 18번 데이터의 경우 분석을 수행하지 않았다. 이는 로지스틱 회귀분석의 특성상 설명변수에 결측치가 발생을 하면 전체 결과를 결측된 것으로 처리하여 목표 변수 중 초기 설정값으로 나타내게 된다. 그로 인하여, 올바른 분석이 수행되지 않을 수도 있다. 따라서, 다른 3개의(신경망, C5.0, C&RT) 단일 기법 예측값보다 예측 값이 현저히 낮은(차이가 평균 10% 이상 발생하는) 위의 4개 데이터는 삭제하고 분석을 수행하였다. 이렇게 만들어진 결과를 토대로 역시 Wilcoxon의 대응 2표본 부호 순위 검정과 대응 t-test 검정을 수행한 결과가 <표 10>과 같다.

<표 10> 로지스틱 회귀분석을 이용한 불일치 패턴 모델의 Wilcoxon 부호순위 검정과 대응 t-test 검정 결과

대응기법	순위N		근사유의 확률(Z-값)	대응 t-test
로지스틱 회귀(사) vs. 불일치 패턴 모델-C5.0& 로지스틱(자)	사 > 자	2	0.001 (-3.243)	0.005 (-3.155)
	사 < 자	15		
	사 = 자	2		
로지스틱 회귀(사) vs. 불일치 패턴 모델-신경망&로지스틱(아)	사 > 아	2	0.001 (-3.290)	0.003 (-3.160)
	사 < 아	15		
	사 = 아	2		
로지스틱 회귀(사) vs. 불일치 패턴 모델-C&RT& 로지스틱(차)	사 > 차	3	0.008 (-2.744)	0.010 (-3.089)
	사 < 차	15		
	사 = 차	1		
신경망(가) vs. 불일치 패턴 모델-신경망& 로지스틱(아)	가 > 아	3	0.006 (-2.744)	0.048 (-2.118)
	가 < 아	15		
	가 = 아	1		
C5.0(나) vs. 불일치 패턴 모델-C5.0& 로지스틱(자)	나 > 자	1	0.001 (-3.541)	0.038 (-2.245)
	나 < 자	18		
	나 = 자	0		
C&RT(다) vs. 불일치 패턴 모델-C&RT & 로지스틱(차)	다 > 차	2	0.001 (-3.421)	0.010 (-2.889)
	다 < 차	17		
	다 = 차	0		

위의 <표 10>에서와 같이 로지스틱 회귀분석을 19개의 데이터에 불일치 패턴 모델에 적용한 결과 역시 단일 기법들 보다 불일치 패턴 모델의 성능이 더욱 효율적인 것을 알 수가 있다. 이렇게 기존의 단일 데이터 마이닝 지도학습 기법과 이를 Hybrid 및 Combined 시킨 불일치 패턴 모델의 비교에서는 불일치 패턴 모델이 성능이 더 우수한 것으로 나타났다. 그러면 불일치 패턴 모델끼리는 어떠한 성능의 차이를 보이는지 확인하기 위하여 다음의 <표 11>과 같이 단일기법과 불일치 패턴 모델끼리의 성능을 비교하여 보았다.

<표 11> 6개의 불일치 패턴 간의 Wilcoxon 부호순위 검정과 대응 t-test 검정 결과

대응기법	순위N		근사유의확률(Z-값)	대응 t-test
불일치 패턴 모델-C5.0&신경망(라) vs. 불일치 패턴 모델-신경망&C&RT(마)	라 > 마	17	0.121 (-1.551)	0.119 (1.622)
	라 < 마	5		
	라 = 마	1		
불일치 패턴 모델-C5.0& 신경망(라) vs. 불일치 패턴 모델-C5.0& C&RT(바)	라 > 바	16	0.020 (-2.321)	0.059 (1.996)
	라 < 바	7		
	라 = 바	0		
불일치 패턴 모델-신경망&C&RT(마) vs. 불일치 패턴 모델-C5.0& 신경망(라)	마 > 라	11	0.648 (-0.456)	0.733 (0.345)
	마 < 라	12		
	마 = 라	0		
불일치 패턴 모델-C5.0& 로지스틱(자) vs. 불일치 패턴 모델-신경망& 로지스틱(아)	자 > 아	10	0.500 (-0.675)	0.307 (1.051)
	자 < 아	8		
	자 = 아	1		
불일치 패턴 모델-신경망& 로지스틱(아) vs.불일치 패턴 모델-C&RT & 로지스틱(차)	아 > 차	10	0.398 (-0.845)	0.949 (0.065)
	아 < 차	9		
	아 = 차	0		
불일치 패턴 모델-C5.0& 로지스틱(자) vs.불일치 패턴 모델-C&RT & 로지스틱(차)	자 > 차	14	0.014 (-2.461)	0.022 (2.498)
	지 < 차	4		
	자 = 차	1		

위의 <표 11>을 보면 불일치 패턴 모델끼리 중 C5.0과 신경망을 가지고 만든 것이 C5.0과 C&RT를 이용하여 만든 불일치 패턴 모델과 C5.0과 로지스틱으로 만든 불일치 패턴 모델이 C&RT와 로지스틱 회귀분석을 이용하여 만든 불일치 패턴 모델보다 성능이 좋은 것을 제외하고는, 나머지 4가지의 경우에서는 별다른 차이가 없다고 할 수 있다. 이는 불일치 패턴 모델은 내부의 2개 기법들 보다는 더 성능을 좋게 만들어 주는 매개체 역할을 하지만, 불일치 패턴 모델 간에는 반드시 성능의 차이를 나타내지 않는다고 말할 수 있다. 이를 좀 더 구체적으로 확인해 볼 수 있는 것이 다음의 <표 12>의 결과이다.

<표 12> 단일기법 중 최고 예측율 모델과 불일치 패턴 모델 중 최고 예측율 모델의 비교

데이터번호	신경망 (가)	C5.0 (나)	C&RT (다)	C5.0&신경망 불일치 패턴 모델 (라)	신경망 & C&RT 불일치 패턴 모델 (마)	C&RT & C5.0 불일치 패턴 모델 (바)
1	68.96%	73.33%	66.44%	77.73%	73.73%	75.80%
2	83.23%	85.45%	83.52%	85.67%	84.50%	85.42%
3	91.34%	94.67%	92.07%	94.67%	93.77%	94.67%
4	66.73%	67.10%	65.25%	72.01%	70.71%	70.06%
5	76.09%	76.54%	77.26%	76.60%	77.26%	76.74%
6	95.14%	96.01%	94.28%	96.22%	95.05%	96.14%
7	80.02%	98.87%	90.09%	98.90%	91.46%	98.88%
8	61.01%	63.40%	63.53%	71.49%	74.01%	68.44%
9	70.30%	71.65%	70.49%	91.10%	70.34%	71.08%
10	90.77%	93.48%	88.10%	93.64%	91.16%	93.53%
11	84.40%	68.88%	75.26%	86.64%	90.43%	79.66%
12	74.80%	78.40%	76.67%	83.70%	76.21%	78.41%
13	96.06%	95.98%	95.60%	96.06%	96.05%	95.98%
14	61.26%	56.76%	57.66%	62.76%	63.36%	55.86%
15	75.53%	74.17%	74.93%	75.64%	75.31%	75.20%
16	91.87%	90.03%	88.91%	92.81%	91.06%	90.73%
17	78.83%	80.06%	80.67%	81.60%	84.05%	85.28%
18	83.70%	87.22%	81.06%	87.22%	85.46%	88.99%
19	83.91%	85.86%	84.70%	86.06%	85.79%	86.45%
20	89.02%	83.82%	85.55%	86.71%	88.44%	86.13%
21	78.85%	76.14%	74.74%	78.41%	78.19%	76.87%
22	67.78%	66.93%	68.58%	67.84%	68.58%	68.64%
23	97.79%	97.73%	97.22%	97.85%	97.79%	97.67%

위의 <표 12>에서 비교의 편의를 위하여 23개의 데이터를 충족시키지 못하는 로지스틱 회귀 분석은 제외를 하고, 신경망 분석, C5.0, C&RT 단일 기법과 이들을 이용한 불일치 패턴 모델을 비교 분석하였다. <표 12>에서 보면 단일 기법 중 최고 예측 확률값을 포함한 불일치 패턴 모델이 다른 불일치 패턴 모델보다 더 좋은 예측력을 보여주고 있다는 것을 알 수 있다. 그렇지 않은 것은 단 한 경우도 없다. 이는 너무 당연한 것이겠지만, 불일치 패턴 모델은 그 내부를 구성하고 있는 2개의 기법 알고리즘에 매우 큰 영향을 받고 있다는 것을 다시 한 번 알 수 있다. 결론적으로 불일치 패턴 모델은 내부의 기법들의 기본적인 성능에 영향을 받고 있고, 2개 기법들의 장점을 가지고와서 기존 2개의 기법보다 일반적으로 더 좋은 모델을 생성하여 주는 매개체적인(또는 메타 성격의) 모델이라는 것을 알 수 있다.

5.2 데이터 마이닝 Combined기법과 불일치 패턴 모델의 비교 및 통합적 활용 실험 결과

본 논문의 목적의 불일치 패턴 모델과 다양한 비교 대상과의 성능 비교를 통해서 불일치 패턴 모델의 특성 및 효율성을 파악하는 것이다. 실험설계에 따른 2번째 실험은 위의 1절에서 단일 지도학습 기법과 불일치 패턴 모델의 성능 비교에 이어서 다양한 기존의 Combined 모델과 불일치 패턴 모델의 비교 그리고 Combined 모델을 불일치 패턴 모델에서 통합적으로 사용한 성능 비교 실험을 수행하였다. 그 결과가 <표 13>과 같다.

<표 13> 다양한 Combined 모델과 불일치 패턴 모델, Combined 모델을 통합한 불일치 패턴 모델의 성능 비교

데이터번호	신경망 (A)	C5.0 (B)	C5.0 - Bag	C5.0 - Boo	C5.0 - Stac	C5.0 /신경망	불일치 패턴	불일치 패턴	불일치패턴모	불일치패턴교

			ging (C)	sting (D)	king (E)	Voting (F)	모델 (G)	모델-C5.0 Bagging (H)	델-C5.0 Boosting (I)	델-C5.0 Stacking (J)
1	68.96% T: 3,286 F: 1,479	73.33% T: 3,494 F: 1,271	74.61% T: 3,555 F: 1,210	76.24% T: 3,633 F: 1,132	75.84% T: 3,614 F: 1,151	75.30% T: 3,588 F: 1,177	77.73% T: 3,704 F: 1,061	77.71% T: 3,703 F: 1,062	78.91% T: 3,760 F: 1,005	78.68% T: 3,749 F: 1,016
2	83.23% T: 8,947 F: 1,803	85.45% T: 9,186 F: 1,564	85.84% T: 9,228 F: 1,522	85.45% T: 9,186 F: 1,564	82.91% T: 8,913 F: 1,837	85.50% T: 9,191 F: 1,559	85.67% T: 9,209 F: 1,541	85.96% T: 9,241 F: 1,509	86.31% T: 9,278 F: 1,472	83.07% T: 8,930 F: 1,820
3	91.34% T: 3,649 F: 346	94.67% T: 3,782 F: 213	94.82% T: 3,788 F: 207	93.97% T: 3,754 F: 241	94.94% T: 3,793 F: 202	94.42% T: 3,772 F: 223	94.67% T: 3,782 F: 213	94.59% T: 3,779 F: 216	94.07% T: 3,758 F: 237	94.82% T: 3,788 F: 207
4	66.73% T: 720 F: 359	67.10% T: 724 F: 355	67.38% T: 727 F: 352	68.56% T: 740 F: 339	78.84% T: 851 F: 228	68.47% T: 739 F: 340	72.01% T: 777 F: 302	72.27% T: 779 F: 300	71.17% T: 768 F: 311	79.05% T: 853 F: 226
5	76.09% T: 1,171 F: 368	76.54% T: 1,178 F: 361	75.76% T: 1,166 F: 373	76.69% T: 1,180 F: 360	77.52% T: 1,193 F: 346	76.43% T: 1,176 F: 363	76.62% T: 1,179 F: 360	75.84% T: 1,167 F: 372	77.01% T: 1,185 F: 354	77.53% T: 1,193 F: 345
6	95.14% T: 31,409 F: 1,604	96.01% T: 31,697 F: 1,316	96.31% T: 31,801 F: 1,212	96.66% T: 31,890 F: 1,123	96.34% T: 31,804 F: 1,209	96.18% T: 31,751 F: 1,262	96.22% T: 31,765 F: 1,248	96.44% T: 31,838 F: 1,175	96.85% T: 31,974 F: 1,039	96.25% T: 31,775 F: 1,238
7	80.02% T: 61,003 F: 15,231	98.87% T: 75,371 F: 863	98.88% T: 75,382 F: 852	98.87% T: 75,371 F: 863	98.92% T: 75,409 F: 825	98.89% T: 75,384 F: 850	98.90% T: 75,398 F: 836	98.92% T: 75,409 F: 825	99.17% T: 75,609 F: 632	98.92% T: 75,409 F: 825
8	61.01% T: 460 F: 294	63.40% T: 478 F: 276	64.19% T: 484 F: 270	63.53% T: 479 F: 275	65.12% T: 491 F: 263	63.40% T: 478 F: 276	71.49% T: 539 F: 215	72.81% T: 549 F: 205	66.71% T: 503 F: 251	67.97% T: 512 F: 242
9	70.30% T: 7,030 F: 2,970	71.65% T: 7,165 F: 2,835	84.02% T: 8,402 F: 1,598	83.95% T: 8,395 F: 1,605	93.86% T: 9,386 F: 614	71.46% T: 7,146 F: 2,854	91.10% T: 9,110 F: 890	91.02% T: 9,102 F: 898	87.78% T: 8,778 F: 1,222	91.55% T: 9,155 F: 845
10	90.77% T: 4,898 F: 498	93.48% T: 5,044 F: 352	93.46% T: 5,042 F: 354	92.98% T: 5,017 F: 379	93.50% T: 5,045 F: 351	93.50% T: 5,045 F: 351	93.64% T: 5,053 F: 343	93.31% T: 5,035 F: 361	93.05% T: 5,021 F: 375	93.52% T: 5,046 F: 348
11	84.40% T: 979 F: 181	68.88% T: 799 F: 361	69.46% T: 805 F: 355	64.22% T: 745 F: 415	85.60% T: 993 F: 167	80.95% T: 939 F: 221	86.64% T: 1,005 F: 155	86.61% T: 1007 F: 153	86.54% T: 1,004 F: 156	86.29% T: 1,001 F: 159
12	74.80% T: 11,302 F: 3,807	78.40% T: 11,845 F: 3,264	79.16% T: 11,961 F: 3,148	79.04% T: 11,945 F: 3,164	79.01% T: 11,937 F: 3,172	78.48% T: 11,858 F: 3,251	83.70% T: 12,646 F: 2,463	83.39% T: 12,599 F: 2,510	82.08% T: 12,402 F: 2,707	79.15% T: 11,958 F: 3,151
13	96.06% T: 5,354 F: 218	95.98% T: 5,348 F: 224	95.73% T: 5,334 F: 238	95.98% T: 5,348 F: 224	96.31% T: 5,367 F: 205	95.98% T: 5,348 F: 224	96.08% T: 5,354 F: 218	96.02% T: 5,350 F: 222	96.09% T: 5,354 F: 218	96.17% T: 5,359 F: 213
14	61.26% T: 204 F: 129	56.76% T: 189 F: 144	61.26% T: 204 F: 129	58.26% T: 194 F: 139	57.96% T: 193 F: 140	56.76% T: 189 F: 144	62.76% T: 209 F: 124	63.66% T: 212 F: 121	58.86% T: 196 F: 137	62.76% T: 209 F: 124
15	75.53% T: 1,392 F: 451	74.17% T: 1,367 F: 476	75.04% T:1,383 F: 460	74.17% T: 1,367 F: 476	74.24% T: 1,368 F: 474	74.17% T: 1,367 F: 476	75.66% T: 1,394 F: 449	75.66% T: 1,394 F: 449	75.55% T: 1,392 F: 450	75.66% T: 1,394 F: 449
16	91.87% T: 1,367 F: 121	90.03% T: 1,353 F: 135	92.47% T: 1,376 F: 112	92.74% T: 1,380 F: 108	91.93% T: 1,368 F: 120	92.87% T: 1,382 F: 106	92.87% T: 1,381 F: 107	92.66% T: 1,378 F: 121	93.12% T: 1,385 F: 102	93.21% T: 1,386 F: 101
17	78.83% T: 257 F: 69	80.06% T: 261 F: 65	73.31% T: 239 F: 87	84.66% T: 276 F: 50	80.37% T: 262 F: 64	80.98% T: 264 F: 62	81.60% T: 266 F: 60	80.67% T: 263 F: 63	84.05% T: 274 F: 52	80.37% T: 262 F: 64
18	83.70% T: 190 F: 37	87.22% T: 198 F: 29	87.22% T: 198 F: 29	87.22% T: 198 F: 29	87.22% T: 198 F: 29	87.67% T: 199 F: 28	87.22% T: 198 F: 29	87.22% T: 198 F: 29	87.22% T: 198 F: 29	87.22% T: 198 F: 29
19	83.91% T: 13,661 F: 2,620	85.86% T: 13,979 F: 2,302	86.19% T: 14,032 F: 2,249	86.76% T: 14,125 F: 2,156	86.19% T: 14,032 F: 2,249	86.13% T: 14,023 F: 2,258	86.06% T: 14,011 F: 2,270	86.19% T: 14,032 F: 2,249	86.76% T: 14,125 F: 2,156	86.24% T: 14,040 F: 2,241
20	89.02% T: 154 F: 19	83.82% T: 145 F: 28	85.55% T: 148 F: 25	84.97% T: 147 F: 26	81.50% T: 141 F: 32	84.97% T: 147 F: 26	86.71% T: 150 F: 23	87.28% T: 151 F: 22	86.71% T: 150 F: 23	84.97% T: 147 F: 26
21	78.8	76.1	76.7	77.1	77.8	76.4	78.4	78.0	78.8	79.5

	5% T: 1,074 F: 288	4% T: 1,037 F: 325	3% T: 1,045 F: 317	7% T: 1,051 F: 311	3% T: 1,060 F:302	3% T:1,041 F:321	1% T: 1,068 F: 294	5% T: 1,063 F: 299	5% T:1,074 F:288	9% T: 1,084 F: 278
22	67.78% T: 2,304 F: 1,095	66.93% T:2,275 F:1,124	67.84% T: 2,306 F: 1,093	66.93% T:2,275 F:1,123	63.40% T:2,155 F:1,244	66.90% T:2,274 F:1,125	67.84% T:2,306 F:1,093	68.43% T:2,326 F:1,073	67.81% T: 2,305 F: 1,094	67.34% T:2,289 F:1,110
23	97.79% T:8,673 F:196	97.73% T:8,668 F:201	97.80% T: 8,674 F: 195	97.73% T: 8,668 F: 201	96.41% T: 8,551 F: 318	97.84% T: 8,677 F: 192	97.85% T:8,678 F:191	97.81% T: 8,675 F: 196	97.85% T: 8,678 F: 191	97.79% T: 8,673 F: 196

위의 <표 13>을 보면, 음영으로 처리한 부분은 불일치 패턴 모델 또는 Combined 모델이 통합된 불일치 패턴 모델의 결과이다. <표 13>을 간단하게 탐색하면 단일 기법이나 기존의 Combined 모델보다 전반적으로 성능이 우수한 것을 알 수 있다. 이에 대한 분명한 검정을 위해서, 통계적 검증을 이용하여 분석하도록 한다. 먼저 불일치 패턴 모델(G)이 기존의 단일 기법이나 Combined 모델과 차이가 있는지 확인하기 위하여, Wilcoxon의 대응 2표본 부호 순위 검정과 대응 t-test를 실시하였다. 검정을 한 결과가 <표 14>와 같다.

<표 14> 기존의 Combined 모델과 불일치 패턴 모델의 성능 비교

대응기법	순위N		근사유의확률(Z-값)	대응 t-test
C5.0(Bagging)(C) vs. 불일치 패턴 모델(G)	C > G	4	0.001 (-3.389)	0.008 (-2.925)
	C < G	17		
	C = G	2		
C5.0(Boosting)(D) vs. 불일치 패턴 모델(G)	D > G	3	0.002 (-3.041)	0.037 (-2.215)
	D < G	18		
	D = G	2		
C5.0(Stacking)(E) vs. 불일치 패턴 모델(G)	E > G	8	0.041 (-2.045)	0.076 (-1.864)
	E < G	14		
	E = G	1		
C5.0과 신경망 Voting(F) vs. 불일치 패턴 모델(G)	F > G	3	0.000 (-3.589)	0.013 (-2.714)
	F < G	20		
	F = G	0		

위의 <표 14>에서 보면 23개의 데이터 집합에서 C5.0의 Bagging(C)과 C5.0의 Boosting(D)그리고 Voting(F)의 경우에도 역시 유의수준 95%에서 불일치 패턴 모델이 우수한 것을 알 수 있다. 이는 전체적으로 일반적인 Combined 모델보다 불일치 패턴 모델이 더 효율적이라는 결과를 말해 주고 있다. 그러나 C5.0 Stacking(E)과 비교했을 때는 대응 t-test 결과 불일치 패턴 모델과 성능의 차이가 없는 것으로 나타났다. Wilcoxon의 대응 부호 순위 검정의 경우 유의한 것으로 나타났으나 유의수준 차가 다른 검정과 비교해보면, 거의 기준치를 벗어난 수준이어서, C5.0 Stacking과의 비교에서 보면 불일치 패턴 모델이 좀 더 우수한 정도가 다른 Combined 모델과 비교하여, 약한 것으로 검정 결과가 나타났다.

다음 <표 15>는 기존의 Combined 모델을 활용하여 개발한 불일치 패턴 모델과 기존 Combined 모델과의 성능 차이를 역시 Wilcoxon의 대응 2표본 부호 순위 검정과 대응 t-test를 이용하여 비교한 결과이다.

<표 15> 기존의 Combined 모델을 활용하여 개발한 불일치 패턴 모델과의 비교

대응기법	순위N		근사유의확률(Z-값)	대응 t-test
C5.0(Bagging)(C) vs. 불일치 패턴 모델-C5.0 Bagging(H)	C > H	2	0.000 (-3.546)	0.007 (-2.971)
	C < H	19		
	C = H	2		
C5.0(Boosting)(D) vs. 불일치 패턴 모델-C5.0 Boosting(I)	D > I	1	0.000 (-3.750)	0.060 (-1.986)
	D < I	21		
	D = I	1		
C5.0(Stacking)(E) vs. 불일치 패턴 모델-C5.0 Stacking(J)	E > J	4	0.006 (-2.725)	0.010 (-2.829)
	E < J	16		
	E = J	3		

<표 15>를 보면 불일치 패턴 모델에서 C5.0을 이용할 때 C5.0의 일반적인 기법을 사용하는 것 보다, 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델을 사용하는 경우 기존의 Combined 모델보다 3가지 경우(Bagging, Boosting, Stacking)에서 모두 유의 수준 95% 수준에서 성능이 더 좋아지는 것을 알 수 있다. 연역적으로 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델이 당연히 단일 기법보다 성능이 좋은 것은 말할 것도 없다. 다음의 <표 16>은 일반적인 불일치 패턴 모델과 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델의 비교를 한 결과이다.

<표 16> 일반 불일치 패턴 모델과 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델 간의 비교

대응기법	순위N		근사유의확률(Z-값)	대응 t-test
불일치 패턴 모델(G) vs. 불일치 패턴 모델-C5.0 Bagging(H)	G > H	9	1.000 (-0.000)	0.665 (-0.439)
	G < H	11		
	G = H	3		
불일치 패턴 모델(G) vs. 불일치 패턴 모델-C5.0 Boosting(I)	G > I	10	0.651 (-0.452)	0.265 (1.144)
	G < I	11		
	G = I	2		
불일치 패턴 모델(G) vs. 불일치 패턴 모델-C5.0 Stacking(J)	G > J	8	0.841 (-0.201)	0.756 (0.315)
	G < J	11		
	G = J	4		
불일치 패턴 모델-C5.0 Boosting(I) vs. 불일치 패턴 모델-C5.0 Bagging(H)	I > H	10	0.638 (-0.471)	0.303 (1.054)
	I < H	12		
	I = H	1		
불일치 패턴 모델-C5.0 Stacking(J) vs. 불일치 패턴 모델-C5.0 Bagging(H)	J > H	9	0.748 (-0.322)	0.699 (0.391)
	J < H	10		
	J = H	4		
불일치 패턴 모델-C5.0 Stacking(J) vs. 불일치 패턴 모델-C5.0 Boosting(I)	J > I	10	0.733 (-0.341)	0.625 (-0.496)
	J < I	12		
	J = I	1		

<표 16>을 보면 기존의 일반적인 불일치 패턴 모델과 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델의 검정 비교 결과를 보면 3개 모두 유의수준 95%에서 별다른 차이가 나지 않는 것을 볼 수 있다. 기존 Combined 모델을 통합하여 개발한 불일치 패턴 모델 사이에서도 별다른 차이가 나타나지 않는 것을 알 수 있다. 이는 앞의 기본적인 단일 지도학습 기법인 C5.0, 신경망, C&RT, 로지스틱 회귀분석을 이용한 불일치 패턴 모델간의 성능 비교를 수행한 <표 10>의 결과와 동일하다고 할 수 있다. 전체적으로 <표 14>, <표 15>, <표 16>의 결과를 통해서 볼 때, 불일치 패턴 모델은 내부에 적용되는 2개의 기법 자체보다는 더욱 좋은 효과를 나타내게 하는 모델임은 확실하다고 할 수 있으나, 불일치 패턴 모델을

위해서 사용되는 내부 기법이 어떤 기법을 이용하는지에 따라서는 차이가 없다는 것을 알 수 있다. 이는 다시 말해서 불일치 패턴 모델 자체의 성격은 서로 다른 2개의 기법을 혼합하여, 시너지(Synergy)를 도출시키는 성격을 가지고 있다는 것을 알 수 있으며, 이는 특정한 알고리즘에만 적용되는 것이 아니라 여러 다양한 알고리즘에 적용시킬 수 있다는 것을 알 수 있다. 즉, 불일치 패턴 모델은 어떤 알고리즘을 적용해도, 분류 정확도와 예측도와 같은 지도학습 기법의 성능을 올릴 수 있는 일반화의 가능성을 보여준다고 할 수 있을 것이다.

5.3 데이터 특성에 따른 불일치 패턴 모델 성능 향상 실험 결과

앞의 5.1과 5.2까지는 불일치 패턴 모델이 기존에 있던 데이터 마이닝 지도학습 기법과 성능 비교를 하여, 좀 더 우수하다는 것을 실험을 통해서 증명하는 것이 목적이었다면, 다음의 실험은 데이터 특성과 기법의 특성에 따라 불일치 패턴 모델의 정확도 및 각종 상황의 변화가 어떻게 되는지 확인하고자 하는 실험들이다. 먼저 본 실험을 수행하기 전에 데이터의 특성을 파악하기 위해 지정한 지표들의 정의를 살펴보면 다음의 <표 17>과 같다.

<표 17> 데이터 특성을 위한 지표의 정의

데이터 특성 지표명	정의
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 가짓수 X 100
데이터의 수	데이터의 전체 사례 수
변수의 수	전체 설명변수의 전체 가짓 수

<표 17>에서 정의한 지표들은 본 논문에서 사용한 23개의 데이터 집합들이 각기 다양한 성격을 가지기에 이들에게서 공통적으로 특성을 정의하여 줄 수 있는 것만 추출을 한 것이다. 다음의 <표 18>은 23개의 데이터 집합별로 데이터 특성 지표들의 값이 어떻게 되는지 정리한 것이다.

<표 18> 데이터 집합별 데이터 특성 지표값

데이터번호	불일치 패턴 모델목표변수 불일치 비율 (%)	목표변수 불균형 비율 (%)	설명변수의 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

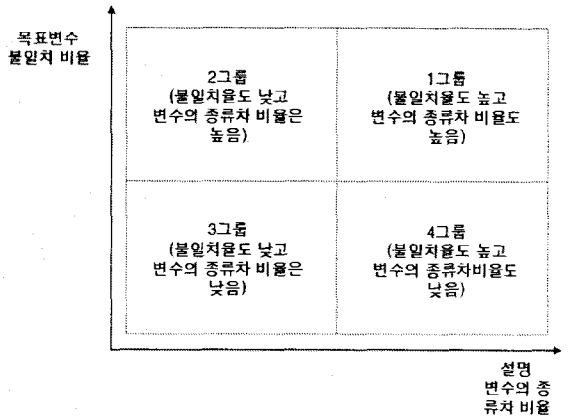
위의 <표 18>에서 최소 정확도 향상율이라는 것은 불일치 패턴 모델을 수행한 결과 비교하고자 하는 기존의 방법보다 더 예측율이 향상된 정도를 의미 한다. 예를 들어 C5.0과 신경망을 이용하여 개발한 불일치 패턴 모델의 경우 신경망 분석 단일 기법보다 향상된 정도 그리고 단일기법 C5.0보다 정확도가 향상된 비율 2가지의 값 중 최소값을 의미한다. 그 옆의 최대 정확도 향상율은 향상된 비율 2가지 값 중 최대 값을 의미한다. 여기의 불일치 비율과 정확도 향상율은 앞서 6개의 불일치 패턴 모델 신경망/C5.0 불일치 패턴 모델, 신경망/C&RT 불일치 패턴 모델, C5.0/C&RT 불일치 패턴 모델 등 단일 기법 3가지의 불일치 패턴 모델과 Bagging, Boosting, Stacking 등 Combined 모델을 통합 3개의 불일치 패턴 모델 등 총 6개의 불일치 패턴 모델의 평균값을 낸 것이다. 한 가지 언급할 사항은 최대 정확도 향상율은 만약 동일한 데이터에서 기법을 잘 선택하는 경우 최대 얻을 수 있는 향상율이라는 의미를 가지고 있다. 다음의 <표 19>는 이들 간의 지표와 정확도 향상율과의 Pearson의 상관관계 분석을 수행한 결과이다.

<표 19> 데이터 특성 지표와 정확도 향상율과의 상관관계

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

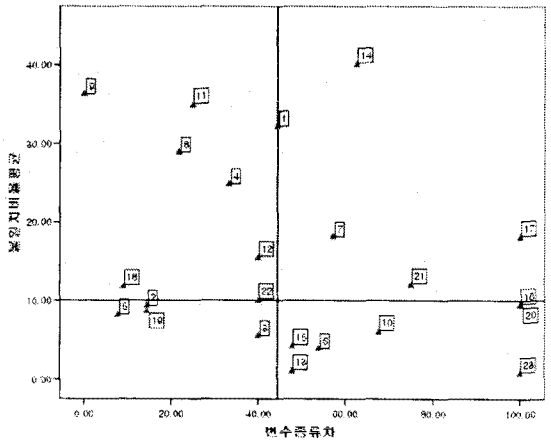
** 0.01 수준에서 유의함 / * 0.05 수준에서 유의함

위의 <표 19>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 비과 유의한 상관성을 보이는 것으로 나타났다. 그 외에 다른 데이터 특성 지표들은 별다른 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) 그리고 목표변수의 균형이 되도록 균등하고 또한 설명변수에 연속형과 범주형 변수가 골고루 분포되어져 있는 경우에 불일치 패턴 모델의 성능이 더 우수한 것을 알 수 있다. 이들을 좀 더 정확하게 파악하기 위하여, 다음의 <그림 6>과 같이 평균 정확도 향상율을 위한 목표 변수 불일치 비율과 설명변수의 종류차 비율로 분할표를 만들어 산점도를 구성하여 본다.



<그림 6> 목표변수 불일치 비율과 설명변수의 종류차 비율의 분할표 설명

위의 <그림 6>과 같은 분할표에 따라 산점도를 그린 결과는 다음의 <그림 7>과 같다.



<그림 7> 불일치 비율과 설면변수의 종류차의 분할표에 따른 23개 데이터의 산포

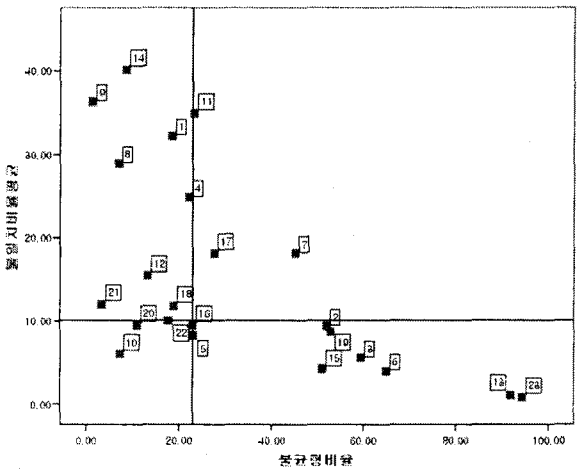
위의 <그림 7>에서 분할면을 자르는 기준으로는 실험에 활용된 23개의 데이터 집합에서 불일치 비율과 설명변수의 종류차의 중위수를 가지고 분할의 기준으로 삼았다. 불일치 비율의 경우 중위수가 10.04이고, 설명 변수의 종류차는 44.44 였다.

다음 <그림 7>에서 만들어진 4개의 분할면을 그룹으로 하여 이들 4개 그룹이 평균 정확도 향상율의 차이를 보이는지 One-Way ANOVA 분석과 데이터의 수가 작은 관계로 분포를 가정하지 않을 수 있기 때문에 비모수의 Kruskal-Wallis 검정을 동시에 확인한 결과가 <표 20>과 같다.

<표 20> 4개의 사분면 그룹별 최소 정확도 향상율의 검정 결과 1

그룹	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		

위의 <표 20>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 설명변수가 연속형과 범주형이 골고루 분포되어져 있는 것을 의미하는 변수의 종류차는 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 점검이 필요할 것이다. 다음은 최대 정확도 향상율과 상관성이 높은 불일치 비율과 목표변수의 불균형 비율을 분할표로 만들어 23개 실험 데이터의 산포를 살펴본 결과가 <그림 8>과 같다.



<그림 8> 불일치 비율과 목표 변수의 불균형 비율의 분할표에 따른 23개 데이터의 산포

위의 <그림 8>에서 불일치 비율의 분할 기준이 되는 중위수는 앞서 <그림 7>과 같이 10.04이며 목표변수의 불균형 비율의 분할 기준이 되는 중위수의 값은 22.98이다. 역시 마찬가지로 One-Way ANOVA 분석과 비모수의 Kruskal-Wallis 검정을 통해서 각 사분면의 그룹들이 차이가 있는지 확인한 결과가 <표 21>과 같다. 단, 여기서 검정 대상이 되는 값은 <표 19>의 결과에 의거하여 최대 정확도 향상율을 이용하였다.

<표 21> 4개의 사분면 그룹별 평균 정확도 향상율의 검정 결과 2

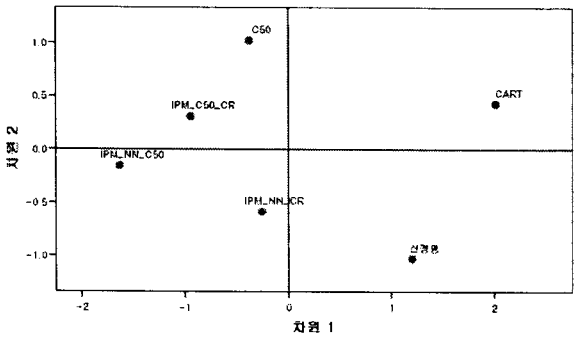
그룹	최대 정확도 향상율의 그룹별 평균값	One-Way ANOVA F통계량(p-value)	Kruskal-Wallis Chi-제곱 통계량(p-value)
1사분면	4.0033	3.518 (0.035)	14.208 (0.003)
2사분면	6.2275		
3사분면	0.6460		
4사분면	0.6200		
Duncan의 사후검정	2사분면=1사분면 > 3사분면=4사분면		

위의 <표 21>을 보면, 역시 불일치 비율이 높고, 목표 변수의 불균형 비율이 낮은 2사분면이 가장 높은 정확도 향상율을 보인 것으로 나타났다. 일반적으로 목표 변수의 불균형(imbalance)는 지도학습 기법의 상당한 걸림돌인데, 이는 마찬가지로 불일치 패턴 모델에서도 그대로 반영이 되어 정확도의 향상을 위한 향후 연구로 불균형한 데이터를 각종 Sampling 기법을 이용하여, 균형을 맞춘 후 불일치 패턴을 이용하여 분석을 수행하는 방법에 대한 것이 필요하다고 하겠다.

위의 <표 20>과 <표 21>을 통해서 보면 불일치 패턴 모델의 경우 기법 간 불일치율이 높은 데이터 집합들과 설명변수가 연속형과 범주형이 골고루 섞여져 있는 데이터 그리고 목표 변수의 불균형까지 잘 맞으면, 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 유용할 수 있다는 결론을 얻을 수 있다.

5.4 불일치 패턴 모델 내부사용 기법에 따른 위치도 분석 결과

앞에서는 데이터의 특성에 따라서 불일치 패턴 모델의 정확도 향상이 어떻게 되는지 한 번 확인을 해 보았다. 다음은 데이터의 특성이 아닌 기법 간의 상관성에 대하여 위치도를 통하여 불일치 패턴 모델의 특성을 파악 해보기로 한다. 먼저 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 그려본 것이 다음의 <그림 9>와 같다.



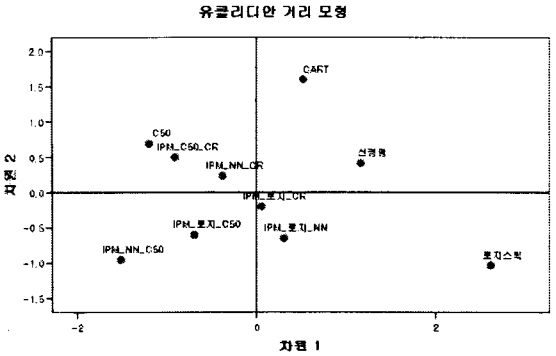
<그림 9> 6개의 모델에 대한 MDS를 이용한 위치도

다음의 <표 22>는 6개의 모델에 대한 23개 데이터 집합들의 예측 정확도의 평균값을 나타낸 표이다. 위의 <그림 9>를 이해하는데 도움을 줄 값들이다.

<표 22> 6개의 모델에 대한 23개 데이터 집합들의 예측 정확도 평균값

기법명	23개 데이터 집합의 예측 정확도 평균값	기법명	23개 데이터 집합의 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%

먼저 위의 <그림 9>를 보면 6개의 기법들의 23개 데이터에 대한 예측 정확도 값의 기법 별로 위치도를 그린 것이다. 위의 6개의 점들의 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 22>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.[2] 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 22>를 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. 이를 좀 더 확연하게 살펴보기 위하여, 로지스틱 회귀 분석을 포함한 19개 데이터 집합들을 가지고, 동일한 분석을 수행하여 본 결과가 <그림 10>과 <표 23>이다.



<그림 10> 로지스틱 회귀분석을 추가한 10개 모델에 대한 MDS를 이용한 위치도

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

<표 23> 10개의 모델에 대한 19개 데이터 집합들의 예측 정확도 평균값

기법명	23개 데이터 집합의 예측 정확도 평균값	기법명	23개 데이터 집합의 예측 정확도 평균값
신경망	80.32%	신경망-C&RT 불일치 패턴 모델	82.73%
C5.0	80.98%	C&RT-C5.0 불일치 패턴 모델	82.44%
C&RT	79.67%	로지스틱-신경망 불일치 패턴 모델	82.97%
로지스틱 회귀분석	79.97%	로지스틱-C5.0 불일치 패턴 모델	84.03%
신경망-C5.0 불일치 패턴 모델	84.41%	로지스틱-C&RT 불일치 패턴 모델	82.92%

위의 <그림 10>과 <표 23>을 보면 단일 기법인 C&RT, 신경망 분석, 로지스틱 회귀분석이 유사한 그룹을 보여주고 있고, C5.0 같은 경우 좀 다른 그룹에 있는 것을 알 수 있다. 또한 <표 23>에서 보면, 가장 평균값이 높은 것이 신경망-C5.0의 불일치 패턴 모델 그리고 C5.0-로지스틱 회귀분석 불일치 패턴 모델인데, 이 경우 C5.0과 거리가 먼 2개의 단일 기법이 로지스틱 회귀분석과 다음이 신경망인 것을 볼 때, 같은 의사결정나무 추론 기법인 C5.0과 C&RT 보다는 좀 더 기법 간 이질성이 강하고, 그로 인하여, 서로 다른 부분을 보완하여 성능을 높이는 불일치 패턴 모델의 특성 상 좀 더 효율적인 성과를 보여주는 것으로 추정하여 볼 수 있다.

<그림 9>와 <그림 10>을 통해서, 단일 기법 중 MDS의 차원 축소에 따른 오류를 감안하더라도 거리감이 있는 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.

6. 최종결론 및 향후 연구과제

6.1 실험의 최종결론

본 논문에서는 실제 기업과 조직에서 데이터 마이닝의 지도학습 기법을 이용함에 있어 가장 중요한 정확도 또는 분류도의 향상을 위해서, 제안한 불일치 패턴 모델의 성능을 다양한 상황 하에서의 비교를 통해서, 불일치 패턴 모델이 효율적인 Hybrid 모델 방법임을 증명하고, 또한 이를 어떤 데이터 및 기법 환경 하에서 활용을 하면 효율적일지에 대하여 연구하였다.

최종적으로 불일치 패턴 모델은 내부적으로 활용되는 2개의 지도학습 기법을 효율적으로 Hybrid 시켜, 지도학습 기법의 가장 큰 지향점인 정확도와 분류 예측력을 높여준다는 것을 파악할 수 있었다. 또한 불일치 패턴 모델에는 단일 기법 이외에도 Bagging, Boosting, Stacking 등 기존의 정확도를 향상시키는 Combined 모델 및 Hybrid 모델도 동시에 통합 활용이 가능하여, 최종적으로는 내부에 적용된 모델보다도 더 좋은 모델을 생성할 수 있다는 것을 알 수 있었으며, 불일치 패턴 모델 간의 성능 비교에서는 큰 차이를 보이지 않아서, 불일치 패턴 모델 내부에 활용된 기법들보다는 우수하지만, 불일치 패턴 모델 간에는 아직까지 큰 상관성이 없다는 것을 알 수 있었다. 그리고 불일치 패턴 모델의 단점이 없는 것은 아니다. 직관적으로 과정이 복잡하고 여러 단계를 거치는 것에 비교 하여, 성능의 향상이 월등하게 이루어지지 않는 경우들도 있으며, 또한 여러 개의 기법들을 혼합한다고 반드시 성능이 계속 개선되지 않는 부분도 있었다.

이러한 성능의 효율성 외에 데이터적인 특성을 살펴보면, 가장 정확도의 향상율에 기여를 하는 것은 통계적으로 검정된 것은 두 기법 간의 데이터에 대한 불일치 비율이 높은 것을 알 수 있었고, 그 외 데이터의 목표 변수의 불균형이 없고, 범주형과 연속형 데이터가 적절하게 배열되어져 있는 데이터에서 향상율이 높은 것을 알 수 있었다. 그 외에도 기법 들간의 상이성이 전체적인 기법의 차이를 많이 나타내는 것을 알 수 있었는데, 탐색적인 분포를 보았을 때 다차원 척도법을 통한 기법 간의 위치도에서 단일 기법간에 거리가 가장 많이 떨어진 기법 간에 불일치 패턴 모델을 개발하는 경우 효율성이 있는 것으로 보여, 이는 두 기법 간의 데이터 불일치 비율이 불일치 패턴 모델의 정확도 향상에 영향을 준다는 앞의 실험을 반증하여 주기도 하였다. 또한 동일한 실험에서 같은 의사결정나무 추론으로 구성된 불일치 패턴 모델보다는 서로 알고리즘이 판이하게 다른 데이터 마이닝 기법을 이용하는 것이 더 효율적일 것이라는 개연성을 파악할 수 있게 되었다. 그리고 데이터 수의 감소에 따라서는 일반적으로는 큰 정확도 향상의 변동은 없지만, 수가 원래 데이터의 1~3%까지 감소를 하면 불일치 패턴 모델의 성능이 한계를 보인다는 것도 알 수 있었다. 물론 이에 해당되지 않고, 오히려 급격하게 정확도가 상승하는 경우도 있어서 절대적 결론은 아니다.

6.2 향후 연구 과제

본 논문에서 소개한 불일치 패턴 모델은 아직 초기 연구 단계이기에 더욱 더 많은 향후 연구과제들이 주어져 있다고 할 수 있다. 먼저 이론적인 관점에서 불일치 패턴 모델과 같은 Hybrid 방법이나 Combined 기법의 경우 일반적으로 단일 기법보다 성능이 좋다고 알려져는 있으나 구체적인 이론적 근거는 미비한 실정이다[이군희, 1998]. 따라서 향후 더 많은 데이터 집합에서의 불일치 패턴 모델에 대한 적용과 더 많은 지도학습 기법의 Hybrid를 통한 비교를 통해서, 불일치 패턴 모델의 좀 더 다양한 특성과 효용성을 증명하고, 검증하는 연구가 향후에 필요할 것으로 본다. 무엇보다, 좀 더 많은 사례 및 데이터 상황에서 불일치 패턴 모델의 성능이 연구가 되어야 할 것으로 보이고, 특히 목표변수가 이분형이 아닌 다(多)범주형에서의 변화 및 성능의 평가도 하나의 향후 연구 과제이다. 그리고 무엇보다 불일치 패턴 모델이 더욱 성능 향상을 할 수 있는 데이터 아키텍처 및 데이터 모델링적 측면에 대한 다양한 연구 역시 필요하다. 그리고 또 다른 측면에서 향후 연구과제는 크게는 경영학적 관점 그리고 무엇보다 경영정보 시스템(MIS) 관점에서 본 논문에서 소개하는 모델이 다양한 산업에서 실제로 활용한 사례에 대한 추가적인 연구가 필요하다고 하겠다. 실험적 데이터를 통한 이론적 보완과 그로 인한 모델의 일반화 이론 문제도 중요하지만, 무엇보다도 산업에서 실제 적용시 나타나는 문제를 보완한 불일치 패턴 모델이 되는 것이야 말로, 데이터 마이닝 시스템과 같은 향후 고도화된 CRM 시스템이 기업과 경영의 의사결정에 큰 도움을 주기 때문이다.

참고문헌

[1] 강명구, 차진호, 김명원, "데이터 마이닝에서 교사학습에 의한 속성 가중치 최적화", 한국정보과학회 봄 학술발표논문집, Vol. 28, No. 1, 2001.

[2] 강문식, 이상용, "데이터 마이닝을 위한 경쟁학습 모델과 BP알고리즘을 결합한 하이브리드 신경망", 정보기술과 데이터베이스 저널, 제9권 2호, 2002, pp.1-16.

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

[4] 송문섭, 박창순, 「非母數統計學槪論」, 자유아카데미, 1989.

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

[6] 이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘 연구", 한국지능정보시스템학회지, 제9권 1호 ,2003, pp.139-155.

[7] 이재식, 이진천, "입력자료 판별에 의한 데이터 마이닝 성능개선",한국지능정보학회학술대회, 2000,

pp.293-303.
[8] 조용준, "신경망 모형의 초기 가중치 최적화 방법에 관한 연구", 중앙대학교 통계학과 박사학위논문, 2003.
[9] 허명회, "Clementine Stream Prototypes : Part 2", SPSS Korea White paper, 2004, pp.1-7.
[10] Entrue Counslting CRM 그룹 역, "CRM을 위한 데이터 마이닝", 대청, 2000.
[11] _,「Clementine Ver. 8 User's Guide」, SPSS Inc, 2003.
[12] Anand, S. S., Patrick, A. R., Hughes, J. G., & Bell, D. A.. "A data mining methodology for cross-sales". *Knowledge-Based Systems*, 10, 1998, pp.449-.461.
[13] Brieman, L. "Bagging Predictors", *Machine Learning*, Vol.24, No.2, 1996, pp.123-140.
[14] Breiman, L., J. H. Freidman, R. A. Olshen and C. J. Stone, "Classification and regression trees", Wadsworth, Belmont, 1984.
[15] Carvalho, Deborah R. and Alex A. Freitas "Hybrid Decision Tree/Genetic Algorithm Method for Data Mining" *Information Sciences*, Vol.163, No.1/3, 2004, pp.13-35.
[16] 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.
[17] Coenen, F. G. Swinnen, 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.
[18] Conversano, Claudio, Roberta Siciliano and Francesco Mola, "Generalized Additive Multi-mixture Model for Data Mining", *Computational Statistics & Data Analysis*, Vol.38, No.4, 2002, pp.487-500.
[19] Freund, Y. and Rober E. Schapire, "Experiments with a New Boosting Algorithm", *Proceedings of 13th International Conference on Machine Learning*, Morgan Kaufmann, 1996, pp.148-156.
[20] Han, J. and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001.
[21] Hansen, L.K. and P.Salaman, "Neural Networks Ensembles", *Transactions on Pattern Analysis and Machine Intelligence*, Vol.12, No.10, 1990, pp.993-1001.
[22] Hsu, P. L., R.Lai, CC.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.
[23] Indurkhya, Nitin and Sholom M. Weiss "Estimating Performance Gains for Voted Decision Trees" *Intelligent Data Analysis*, Vol.2, No.1/4, 1998, pp.303-310.
[24] Kuncheva, L.I.C. Bezdek and M.A.Shutton, "On Combining Multiple Classifiers by Fuzzy Templates", *International Conference on Artificial Neural Networks IEEE*, 1998, pp193-197.
[25] Li, Renpu and Zheng-ou Wang "Mining Classification Rules Using Rough Sets and Neural Networks", *European Journal of Operational Research*, Vol.157, No.2, 2004, pp.439-448.
[26] Lin, Feng Yu and Salley McClean "A Data Mining Approach to the Prediction of Corporate Failure", *Knowledge-Based Systems*, Vol.14, No.3/4, 2001, pp.189-195.
[27] Lu, h, R. Setiono and H. Liu "Effective data mining using neural networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol.8(6), 1996, pp.957-961.
[28] Pawlak, Z., J. Grzymala-Busse, R. and Slowinski, W. Ziarko, "Rough sets", *Communications of the ACM 38 (11)*, 1995, pp.88-.95.
[29] Quinlan, J. R., "Bagging, Boosting and C4.5", *Procs. 13th American Association for Artificial Intelligence*, AAAI Press, 1996.
[30] Quinlan, J. R., "C4.5 Programs for machine Learning", San Mateo: Morgan Kaufmann, 1993.
[31] Schapire, Robert, Yoav Freund, Peter Bartlett, and Wee Sun 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.
[32] 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.
[33] Versace, Massimiliano, Rushi Bhatt, Oliver Hinds and Mark 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.
[34]Webb, G. I., and Zheng, Z., "Multistrategy ensemble learning: reducing error by combining ensemble learning techniques." *IEEE Transactions on Knowledge and Data Engineering*, 16(8), 2004, pp.980-.991.
[35] Wolpert, L., "Stacked Generalization", *Neural Networks*, Vol.5, No.2, 1992, pp.241-259.
[36] Wong, M. L., Lee, S. Y., and Leung, K. S., "Data mining of Bayesian networks using cooperative coevolution", *Decision Support systems*, 38, 2004, pp.451-.472.
[37] Zhang, Z., C. Zhang, "Agent-Based Hybrid Intelligent Systems." *LNAI 2938*, 2004, pp.127-.142.
[38] Zhou, Zhi-Hua, Jianxin Wu and Wei Tang, "Ensembling Neural Networks: Many Could Be Better Than All", *Artificial Intelligence*, Vol.137, No.1/2, 2002,pp.239-263.
[39] http://www.autonlab.org/autonweb/downloads/datasets.html
[40] http://www.kdnuggets.com/polls/2006/data_mining_methods .htm
[41] http://www.ics.uci.edu/~mlearn/MLRepository.html