다중모델을 이용한 자동차 보험 고객의 이탈예측

Customer Churn Prediction of Automobile Insurance by Multiple Models

저자
(Authors)

이재식, 이진천
Jae Sik Lee, Jin Chun Lee

출처
(Source)

지능정보연구 12(2), 2006.6, 167-183(17 pages)
Journal of Intelligence and Information Systems 12(2), 2006.6, 167-183(17 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE00742929

APA Style

이재식, 이진천 (2006). 다중모델을 이용한 자동차 보험 고객의 이탈예측. 지능정보연구, 12(2), 167-183

다중모델을 이용한 자동차 보험 고객의 이탈예측*

이재식
아주대학교 경영대학 e-비즈니스 학부
(leejsk@ajou.ac.kr)

이진천
아주대학교 경영대학 e-비즈니스 학부
(giny777@empal.com)

데이터마이닝은 우리가 완벽하게 알고 있지 못하는 데이터 집합으로부터 알려지지 않은 사실이나 규칙을 찾아내는 작업이기 때문에 항상 높은 오류율의 위험에 처해 있다. 다중모델은 하나의 문제에 다수의 모델을 사용함으로써 오류율을 줄이고자 하는 접근 방법이다. 본 연구에서는 데이터마이닝의 예측 성능을 개선시킬 수 있는 새로운 방식의 다중모델을 제시한다. 이 다중모델은 입력사례의 특성에 따라 그에 적합하게 개발된 모델이 선정되어 적용되는 특징을 가지고 있다. 제시된 다중모델의 현실적인 성능 검증을 위해 국내 자동차 보험 가입 고객의 이탈 예측 문제에 적용하여, 그 결과를 단일모델의 결과와 비교 평가하였다. 비교 대상 단일모델로는, 사례기반추론, 인공신경망, 의사결정나무 등이 사용되었는데, 다중모델의 예측 성능이 어떤 단일모델의 예측 성능보다 우수한 것으로 나타났다.

논문접수일 : 2006년 2월 게재확정일 : 2006년 6월 교신저자 : 이재식

1. 서론

데이터마이닝은 아직 알려지지 않았지만 타당하고 활용 가능한 규칙(Rules)이나 패턴(Patterns) 등을 데이터베이스로부터 찾아내는 과정이다(Berry and Linoff, 2004). 기업은 정보기술의 발전으로 방대한 양의 데이터를 축적하게 되었고, 이렇게 축적한 데이터로부터 비즈니스 의사결정에 필요한 정보나 지식을 제공하는 데이터마이닝의 필요성을 인식하게 되었다. 데이터마이닝은 금융, 통신, 보험, 유통, 제조, 의료, 그리고 공공분야와 같은 다양한 산업영역에서 이미 활용되고 있다. 특히, 장기간 고객데이터를 축적해온 B2C 기업들은 고객관리와 마케팅활동의 효과성을 높이기 위해 데이터마이닝을 핵심도구로 활용하고 있다. 데이터마이닝의 응용분야가 다양해지고 폭넓어짐에 따라 데이터마이닝 문제영역의 복잡성 또한 한층 심화되고 있다. 문제영역이 복잡해질수록 문제해결을 위한 최적 모델의 도출이 어려워진다. 즉, 하나의 모델링 기법만을 사용하여 최적 모델을 찾는 것이 어려워지는 것이다. 그러므로 하나의 문제에 대하여 다양한 모델의 결합을 적용하고자 하는 시도들이 이루어지고 있다.

서로 다른 특성을 가진 다수의 모델들을 결합하여 사용하는 것은 단일모델을 사용했을 때 발생하는 예측오류를 축소함으로써 예측성능을 제고하기 위한 것이다. 여기서 다수의 모델이란 하나의 모델링 기법에서 파라미터를 달리 함으로써 얻어지는

* 본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스 컴퓨팅 네트워크 원천 기반기술개발 사업의 지원에 의한 것임.

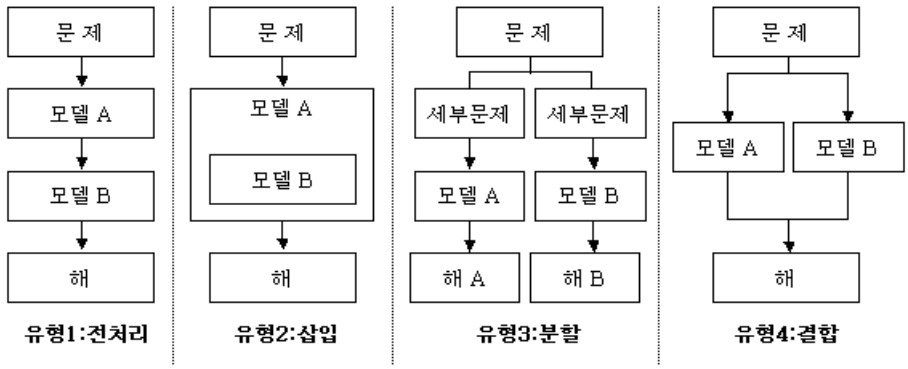
모델들일수도 있고, 서로 다른 모델링 기법을 사용하여 얻어지는 모델들일수도 있다. 특히, 후자의 방법에 의해 생성된 각각의 모델들을 결합한 최종 모델을 하이브리드(Hybrid) 모델이라고 부르는데, 본 연구에서는 전자와 후자를 통칭하여 다중모델(Multiple Model)로 부르기로 한다. 데이터마이닝에 사용되는 모델링 기법들은 저마다의 장·단점을 가지고 있다. 다중모델의 사용은 각각의 기법들이 가지는 이러한 장점의 결합과 단점의 보완을 통해 예측모델의 성능을 제고하기 위한 것이다.

본 연구에서는 분류문제의 예측성능을 향상시킬 수 있는 새로운 형태의 다중모델 구축방법을 제시하였으며, 제시한 다중모델을 국내 자동차보험 가입고객의 이탈예측문제에 적용하여 그 유용성을 검증하였다. 본 논문의 구성은 다음과 같다. 제 2절에서는 선행 연구의 검토를 토대로 다중모델의 유형을 분류하고 본 연구에서 제시하는 DyMoS 다중모델을 소개한다. 제 3절에서는 자동차보험 가입고객의 이탈예측을 위한 문제정의와 실험데이터의 구성 그리고 DyMoS 다중모델의 구축방법을 설명한다. 제 4절에서는 고객 이탈예측을 위한 단일모델 생성과정과 예측결과를 제시하고, 제 5절에서는 DyMoS 다중모델의 구축과정을 기술하고, 예측결과에 대한 성능을 평가한다. 마지막으로 제 6절에서는 본 논문의 결론과 함께 연구의 한계점과 향후 연구 과제를 제시한다.

2. 다중모델의 유형

2.1. 다중모델에 대한 선행연구

예측모델의 예측오류를 줄이기 위한 모델 설계 방법으로는 크게 두 가지 접근법이 있다. 하나는 단일모델 하에서 파라미터들(Parameters)의 최적화를 통해 최적의 모델을 찾는 단일모델 접근법이고, 다른 하나는 두개 이상의 모델을 결합하는 다중모델 접근법이다. 단일모델 접근법의 경우, 예측오류를 최소화하는 최적의 학습 파라미터를 찾기가 어렵고, 사용되는 모델링 기법에 따라 예측모델의 성능이 민감하게 달라질 수 있는 문제점이 있다. 따라서 최적의 단일모델을 찾기 위해서는 많은 시행착오를 거쳐야 한다. 더욱이 문제영역이 매우 복잡한 경우에는 학습 파라미터의 최적화만으로 예측오류를 줄이는데 한계가 있을 수 있다. 이러한 문제점을 해결하기 위한 접근법이 다중모델 접근법이다(이재식과 차봉근, 1999; Giacinto and Roli, 2001; Kuncheva, 2001; Kim et al., 2002; Daskalaki et al., 2003; Hsieh, 2005; Zhang et al., 2005; Wang and Wang, 2006; Kim et al., 2006). 다중모델은 “combination of multiple classifiers”, “classifier fusion”, “mixture of experts”, “composite classifier system”, “classifier ensembles”등 다양한 용어로 사용되고 있다(Kuncheva, 2001). 다중모델 사용의 기본 사상은 서로 다른 특성의 모델들을 결합함으로써 해당 문제영역에 포함되어 있는 다양한 패턴들을 더 많이 학습할 수 있다는 것이다. Kuncheva(2001)는 다중모델의 예측성능이 단일모델보다 더 우수함을 보였고, Daskalaki등(2003)은 통신회사에서 고객의 채무불이행을 예측하기 위해 다중모델을 사용하였다. 그들은 먼저 판별분석(Discriminant Analysis) 모델, 의사결정나무(Decision Tree) 모델, 그리고 인공신경망(Artificial Neural Networks) 모델을 독립적으로 생성한 후, 세 모델이 동일한 예측결과를 제시하는 경우에만 그 결과를 최종 예측값으로 사용하는 다중모델을 구축하였다. 이 연구에서, 다중모델의 예측성능이 단일모델보다 더 우수함을 보였다.



[그림 1] 다중모델의 네 가지 유형

다중모델은 모델의 구축과정에서 각각의 모델들이 어떠한 방식으로 결합되는지에 따라 다양한 유형으로 분류된다. 선행 연구된 다중모델의 유형은 [그림 1]과 같다(한인구와 신경식, 1999).

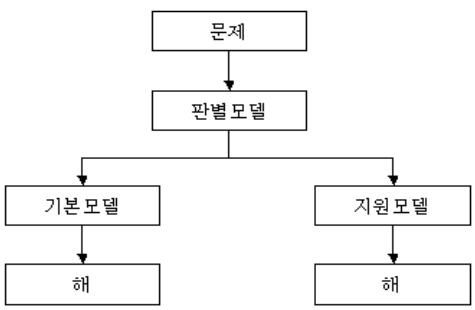
유형1은 하나의 모델이 다른 모델의 효율성이나 부분적인 최적화를 지원하기 위해 전처리 모델로 사용되는 형태이다(Yang and Honavar, 1998; Last et al., 2001; Hsieh, 2005; Zhang et al., 2005). 예를 들면, 의사결정나무를 통해 속성을 선정한 후 인공신경망을 통해 최종해를 예측하는 형태가 유형 1에 해당한다. 유형 2는 하나의 모델이 다른 모델의 최적화를 위한 구조형성에 참여하는 형태이다(이재식과 차봉근, 1999; Sexton and Sikander, 2001; 이재식과 전용준, 2001). 이재식과 차봉근(1999)은 기업도산예측을 위해 인공신경망의 입력속성 및 은닉노드의 수를 최적화하기 위한 방법으로 유전자알고리즘의 사용을 제시하였다. 또한 이재식과 전용준(2001)은 사례기반추론 모델에서 입력사례에 따라 속성들의 가중치를 동적으로 결정하는 별도의 사례기반추론 모델의 사용을 제시하였다. 유형 3은 문제영역을 세분화한 후 세분화된 문제영역에 대해 개별적인 모델을 생성하여 예측을 수행하는 형태이다(Kwon and Lee, 1997; Wei and Chiu, 2002). 마지막으로 유형 4는 각각의 단위모델들이 독립적으로 예측을 수행한 후, 그 결과를 결합하여 최종해를 도출하는 형태이다(Opitz and Maclin, 1999; Zhou et al., 2001; Daskalaki et al., 2003; Hothorn and Lausen, 2003; Kim et al., 2006; Wang and Wang, 2006).

[그림 1]에서 유형 1과 유형 2는 다중모델이 예측에 직접적으로 사용되는 것이 아니라 예측에 사용되는 단일모델의 최적화를 위해 사용된다. 그에 반해 유형 3과 유형 4는 예측에 직접적으로 다중모델이 사용된다. 유형 3에서는 먼저 예측대상 문제를 세부문제로 분할하여야 하는데, 분할에 대한 이론적인 방법론이 없기 때문에 분석가의 직관에 의존해야하는 문제점이 있다. 유형 4의 경우에도 단일모델의 예측결과를 어떻게 결합하는 것이 좋은지에 대한 이론적 판단기준이 없기 때문에 이 또한 분석가의 경험과 직관에 의존할 수밖에 없다. 특히 유형 4에서 다중모델을 구성하는 단일모델들의 예측성능이 낮을 경우에는 잘못된 예측결과가 최종해로 선택될 수 있는 위험성도 존재한다. 이에, 본 연구에서는 이러한 문제점들을 완화시킬 수 있는 새로운 방식의 다중모델을 제시하고자 한다.

2.2. DyMoS 다중모델의 구조

본 연구에서 제시하는 다중모델은 입력사례의 패턴분석결과에 따라 동적으로 예측모델을 선정하는 방식이고, 그 구조는 [그림 2]와 같다. 우리는 동적으로 예측모델을 선정한다는 의미에서 이 모델을 DyMoS(Dynamic Model Selection)로 명명하였다. DyMoS 다중모델을 구성하는 요소모델들의 정의는 다음과 같다.

- **판별모델 :** 입력사례의 특성에 적합한 예측모델을 선정하는 모델.
- **기본모델 :** 학습데이터 전체를 사용하여 생성되며, 해당 문제영역의 입력사례 대부분을 예측하기 위한 모델.
- **지원모델 :** 기본모델이 예측에 실패한 사례만을 사용하여 생성되며, 기본모델이 예측에 실패할 것으로 판별된 사례들에 대해서만 예측을 수행하는 보조모델.



[그림 2] DyMoS 다중모델의 구조

본 연구에서 제시한 DyMoS 다중모델은 다음의 4 단계를 거쳐 완성된다.

- **제 1단계 : 기본모델의 생성**

기본모델의 생성을 위해 먼저 대상 문제영역에 대해 독립적인 단일모델들을 구축한다. 구축된 모델들 중 가장 좋은 성능을 보인 모델을 기본모델로 선정한다. 단일모델의 구축에는 인공신경망, 사례기반추론, 그리고 의사결정나무 등의 다양한 데이터마이닝 기법이 사용될 수 있다.

- **제 2단계 : 판별모델의 생성**

제 1 단계에서 선정된 기본모델의 예측결과로부터 예측에 성공한 사례와 실패한 사례를 분류한다. 이렇게 분류한 사례들을 이용하여 예측에 성공할 사례를 판별하는 모델을 구축한다. 판별모델의 구축에도 다양한 데이터마이닝 기법이 적용될 수 있다.

- **제 3단계 : 지원모델의 생성**

지원모델은 기본모델이 예측에 실패한 사례들만을 가지고 생성한다. 즉, 기본모델로는 패턴 파악을 할 수 없는 사례들의 영역을 처리하기 위한 보조모델이다. 지원모델 또한 모든 데이터마이닝 기법들이 적용될 수 있다.

- **제 4단계 : DyMoS 다중모델의 구축**

제 1, 2, 3 단계에서 생성된 기본모델, 판별모델, 그리고 지원모델을 결합하여 DyMoS 다중모델을 완성한다.

본 논문에서 제안하는 DyMoS 다중모델이 [그림 1]의 유형 4와 유사하게 보이지만, 다음과 같은 점에서 차이가 있다. 첫째, 모델을 구축할 때 학습과정이 다르다. 유형 4의 경우 다중모델에 사용된 단일모델들([그림 1]의 유형 4에서 모델 A와 모델 B)은 대부분 동일한 학습데이터를 사용하여 구축된다. Boosting과 같이 동일한 학습데이터를 사용하지 않고 선행 모델의 결과에 따라 학습데이터를 재구성하여 다음 모델을 구축하는 방식도 있으나, 이러한 경우에도 재구성된 데이터의 특성이 원래의 학습데이터 특성과 큰 차이가 없고 데이터의

크기 또한 동일하다. 그러나 본 논문에서 제안하는 DyMoS 다중모델은 기본모델과 지원모델을 서로 다른 특성의 학습데이터를 사용하여 구축한다. 제5절에서 자세히 언급하겠지만, 기본모델은 학습데이터와 검증데이터 전체를 사용하여 구축하지만, 지원모델은 기본모델이 예측오류를 일으킨 사례들만을 사용하여 구축한다. 둘째, 새로운 사례에 대한 예측 수행과정이 다르다. 유형 4는 새로운 입력사례가 주어지면, 개별 단일모델로 각각 예측을 수행한 후에 그 결과를 결합하여 최종해를 제시하는 방식이다. 반면에 DyMoS 다중모델은 새로운 입력사례가 주어지면, 판별모델을 통하여 기본모델과 지원모델 중 하나만을 선정한 후에 예측을 수행하여 그 결과를 제시하는 방식이다.

3. 자동차보험 가입고객의 이탈예측

최근 금융 산업에서는 고객이탈문제가 심각한 이슈로 대두되고 있다. 특히 보험업의 경우 정부의 보호 하에서 높은 성장률을 보여 왔으나, 금융과 보험간의 규제완화, 보험요율의 자율화, 외국보험업체의 국내진출 등에 의해 경쟁이 심화되고 있다. 보험업체간의 경쟁은 신규고객 확보에 대한 비용증가와 기업부담을 가중시켰고, 이로 인해 기존고객의 이탈문제가 보험업체의 중요한 이슈로 부각되었다. 이러한 상황에서 기업들은 고객의 이탈방지를 위해 다양한 노력을 하고 있으며, 이탈고객 예측모델은 이러한 노력을 더 효과적이고 효율적으로 지원하게 된다. 본 연구에서는 국내 자동차보험사의 고객이탈 예측문제를 다루었고, '보험가입고객 중 보험만료일 시점으로부터 3개월 이내에 재가입을 하지 않은 고객'을 이탈고객으로 정의하였다.

실험데이터를 구성하기 위해 먼저 데이터베이스로부터 고객이탈과 관련 있을 것으로 판단되는 후보속성들을 추출하여 분석용 데이터를 구성하였다. 분석용 데이터에는 <표 1>과 같이 총 52개의 속성이 포함되었으며, 이 중 이탈여부 속성은 앞서 정의한 이탈고객의 정의에 따라 생성되었고 모델 설계과정에서 목표속성으로 사용되었다. 분석용 데이터는 약 110만 건의 사례를 포함하고 있으며, 이 중 이탈고객의 비율은 약 45.5%에 해당한다.

<표 1> 고객테이블의 속성들

1	피보험자구분코드	14	표준할인할증	27	영업소	40	대물담보
2	피보험자	15	가입경력	28	취급자	41	자손담보
3	나이	16	보험시기	29	물건구분	42	무보험담보
4	성별	17	보험종기	30	연식	43	차량담보
5	직업코드	18	전계약사	31	차종	44	납입방법
6	단체적용율대상	19	보험기간내사고여부	32	운행용도	45	수납형태
7	우편번호1	20	취급자구분	33	차명코드	46	영수보험료
8	우편번호2	21	지점	34	대인담보	47	손해율
9	특약1	22	특약2	35	특약3	48	특약4
10	특약5	23	특약6	36	취급자 주민번호	49	위촉/계약일
11	입사일	24	학력	37	활동지역	50	직책
12	3년누적손해율평균	25	평균수익	38	마감횟수	51	소득범위
13	해촉/해지구분	26	해촉/해지일자	39	근무년수	52	이탈여부

예측에 사용되는 데이터의 질(Quality)은 예측 모델의 성능에 많은 영향을 미친다. 따라서본 연구에서는 모델 구축에 앞서 연구에 사용될 데이터에 대해 다음과 같은 데이터정제(Data Cleaning) 작업을 수행하였다.

- 작업 1 : 각 속성들 간의 상충관계를 조사한 후, Noise를 적합한 값으로 변환한다. Noise란, 예를 들어, 어떤 설계사의 해촉일자가 2005년 8월인데, 이 설계사가 2005년 9월에 계약된 고객의 담당설계자로 등록되어있는 경우를 말한다.
- 작업 2 : 각 속성들을 대상으로 Noise 및 Null값을 조사한 후, 그 비율이 전체의 40% 이상을 차지하는 속성들은 분석용 데이터에서 제거한다. 단, 업무적인 처리를 목적으로 Null값이 사용된 경우에는 해당 Null값을 업무 목적에 부합되는 코드값으로 변환한다.
- 작업 3 : 각 속성들에 대해 Noise 및 Null값이 차지하는 비율이 40% 미만일 경우에는 이 값들을 대표값으로 변환한다. 해당 속성이 수치형일 경우 평균값을 대표값으로 사용했으며, 범주형일 경우 최빈값을 사용하였다. 그리고 한 사례에 포함된 속성 중 4개 이상이 Noise 및 Null값인 경우에는 분석용 데이터에서 제외하였다.

데이터정제 작업의 결과, 6번(단체적용율대상), 8번(우편번호2) 그리고 32번(운행용도) 속성을 제거하였고, 나머지 48개의 속성들을 이용하여 새로운 파생속성(Derived Variables) 21개를 생성하였다. 이렇게 구성된 총 69개의 후보속성들로부터 다음과 같은 방식으로 예측에 사용할 최종입력속성을 선정하였다.

- 제 1단계 : 69개의 후보속성과 목표속성의 단변량 통계분석을 통해 목표속성과 연관성이 높은 속성들을 제 1차 입력속성으로 선정한다. 단변량 분석에는 Chi-square 검정과 ANOVA 분석이 사용되었고, 제 1차 입력속성은 p-value가 0.05 이하인 것들이 선정되었다.
- 제 2단계 : 제 1 단계에서 선정된 입력속성들 사이에 다중공선성이 존재하는 속성들을 제거한다. 다중공선성 분석에는 상관분석을 사용하였고, 두 속성간 상관계수 r의 절대값이 0.8 이상인 경우 그들 중에 하나의 속성만을 선정하였다.

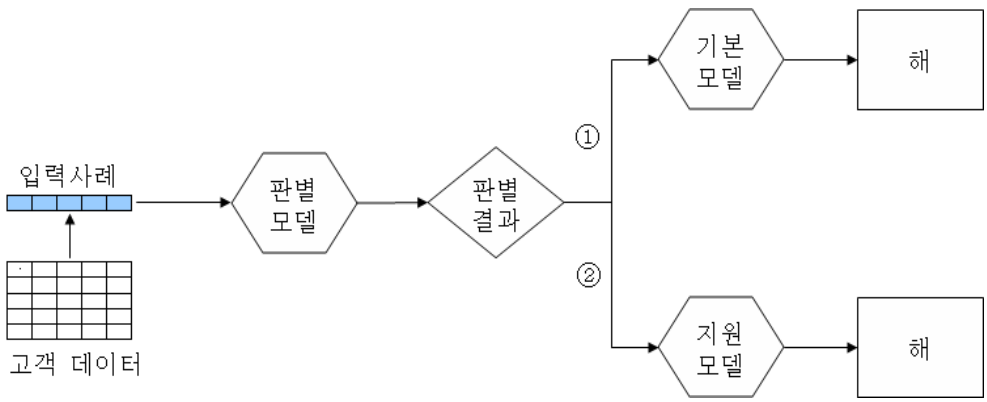
속성선정 작업을 통해 최종적으로 46개의 속성이 선정되었고, 선정된 속성들의 일부는 <표 2>와 같다.

<표 2> 이탈고객 예측에 사용된 입력속성

속성	속성설명	속성유형
1	고객 성별	범주형
2	고객 연령	수치형
3	보험 가입 경력	수치형
4	전계약사	범주형
5	차량 연식	수치형
6	납입 방법	범주형
7	수납 형태	범주형
8	고객 직업 대분류	범주형
9	보험 종목	범주형
10	차종 약칭	범주형
11	보험기간 내 사고여부	범주형
12	대인담보 가입여부	범주형
...	...	...

본 연구에서는 예측모델의 구축을 위해 110만건의 고객데이터로부터 30,000건을 추출하였다. 이

렇게 추출한 데이터는 전체데이터와 통계적으로 동일한 분포를 가지는데, 이를 실험데이터로 명명하였다. 실험데이터의 이탈고객 비율은 45.6%이다. 실험데이터는 모델의 학습과 검증을 위해 학습데이터(Training Data), 검증데이터(Validation Data), 그리고 평가데이터(Test Data)로 분할하였다. 모델구축을 위한 데이터분할에 특별히 정해진 방법은 없으나 학습데이터, 검증데이터, 그리고 평가데이터의 비율을 5 : 3 : 2 또는 7 : 2 : 1로 나누는 경우가 보편적이다. 그러나 본 연구에서는 데이터의 양이 충분하고, 모델의 성능을 좀 더 신중히 평가하기 위해 각 데이터의 비율을 동일하게 하였다. 학습데이터는 모델의 학습을 위해 사용하였으며, 검증데이터는 모델의 과잉적합(Overfitting) 방지와 최적모델 선정을 위해 사용하였다. 그리고 평가데이터는 선정된 최적모델의 일반화 성능을 평가하기 위한 목적으로 사용하였다.



[그림 3] DyMoS 다중모델의 예측과정

3.2 이탈고객 예측모델의 설계

본 연구에서 구축한 DyMoS 다중모델의 이탈고객 예측과정은 [그림 3]과 같다.

먼저 입력사례가 들어오면 판별모델로 기본모델과 지원모델 중 어떤 것을 사용하여 예측할지를 결정한다. 만약 기본모델을 사용하는 것이 더 적합하다고 판별되면 [그림 3]에서 ①번 방향으로 예측이 진행된다. 만약 지원모델을 사용하는 것이 더 적합하다고 판별하면, ②번 방향으로 예측이 진행된다.

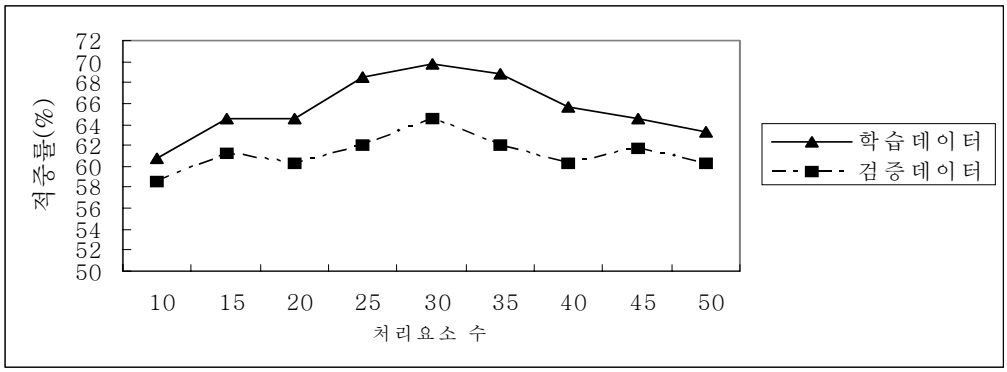
4. 단일모델에 의한 이탈고객 예측모델의 설계

이 절에서는 자동차보험 가입고객의 이탈예측을 위한 단일모델 생성과정과 그 결과를 설명한다. 단일모델의 생성에는 인공신경망, C5.0 의사결정나무, 그리고 사례기반추론 기법이 사용되었다. 본 논문에서 사용한 인공신경망 모델과 C5.0 의사결정나무 모델은 SPSS의 Clementine 5.2로 개발하였고, 사례기반추론 모델은 Microsoft Visual Basic 6.0으로 개발하였다.

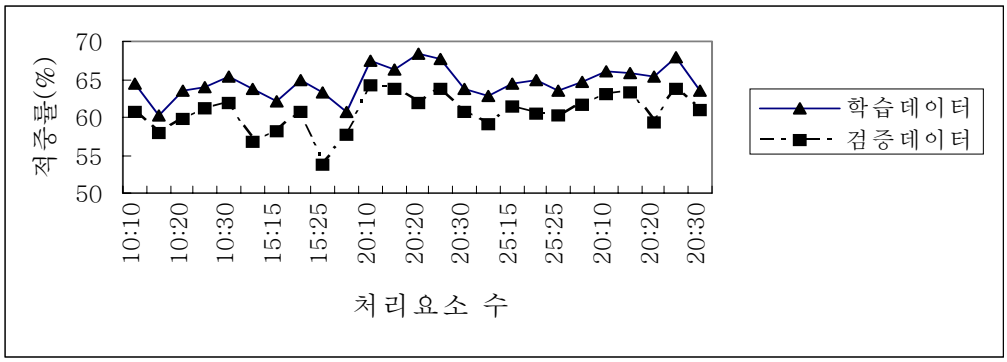
4.1 인공신경망에 의한 이탈고객 예측모델의 설계

본 연구에서는 다계층퍼셉트론(Multi-Layer Perceptron)의 구조로 인공신경망 모델을 설계하였다(Nelson and Illingworth, 1991). 입력층(Input Layer)에는 제 3절에서 선별된 46개의 속성을 사용하였고, 출력층(Output Layer)은 이탈여부를 판별하는 하나의 노드로 구성되었다. 은닉층(Hidden Layer)의 개수는 이론적으로 제한은 없으나, 하나의 층만으로도 모든 형태의 비선형문제해결이 가능하기 때문에 최대 2개의 범위 안에서 실험을 진행하였다. 학습알고리즘으로는 Backpropagation 알고리즘을 사용하였다.

최적의 인공신경망 모델을 찾기 위해, 먼저 한 개의 은닉층 내에서 처리요소의 수를 10개에서 시작하여 5개씩 증가시키면서 실험을 수행하였다. [그림 4]의 실험결과를 보면 은닉층의 처리요소 수가 30개에 가까워질수록 학습데이터 적중률과 검증데이터 적중률이 좋아지다가 30개 이후부터는 점점 감소하는 현상이 나타났다. 따라서 은닉층의 처리요소를 50개 이상으로 설정하여도 모델의 성능이 개선되지 않을 것으로 판단되어 그 이상의 실험은 진행하지 않았다. 그 결과, 단일 은닉층 구조에서는 처리요소를 30개로 하였을 때 검증데이터에 대해 가장 좋은 적중률을 보였고, 이때의 적중률은 64.5%였다.



[그림 4] 인공신경망의 단일 은닉층 구조에서의 실험결과



[그림 5] 인공신경망의 은닉층이 두개일 때의 실험결과

두 개의 은닉층을 사용하는 실험에서는 처리요소 수가 지나치게 클 경우 과잉적합 문제가 발생할 수 있으므로 각 층의 처리요소를 30개 이내로 제한하였다. [그림 5]의 결과를 보면 첫 번째 은닉층에 20개의 처리요소를 두고, 두 번째 은닉층에 10개의 처리요소를 둔 모델이 검증데이터에 대해 가장 높은 적중률(64.2%)을 보여주었다.

결과적으로, 인공신경망을 이용한 자동차보험 가입고객의 이탈예측은 은닉층의 처리요소 수가 30개인 단일 은닉층 구조에서 가장 좋은 적중률(64.5%)을 보여주었다.

4.2 의사결정나무에 의한 이탈고객 예측모델의 설계

의사결정나무는 대상이 되는 집단을 몇 개의 소집단으로 구분하여 분류 및 예측을 수행하는 기법이다. 이것은 나무의 깊이, 최종노드 안에 포함되는 사례의 개수, 가지의 분리방법, 그리고 가지치기 등과 같은 기준에 의해 생성된다. 의사결정나무는 가지를 분리하는 방법에 따라 여러 종류의 알고리즘이 있는데, 본 연구에서는 가지의 분리기준으로 이득률(Gain Ratio)을 사용하는 C5.0 알고리즘(Quinlan, 1998)을 사용하였다.

의사결정나무 모델의 과잉적합을 방지하기 위해서는 적절한 가지치기와 최종노드(Leaf node)에 포함될 사례의 개수에 대한 적절한 설정이 필요하다. 그러므로 본 연구에서는 최적의 의사결정나무 모델을 찾기 위해 C5.0 알고리즘에서 제공하는 가지치기 엄격도와 최종노드에 포함되는 최소 사례수를 조정하면서 실험을 수행하였다. 가지치기 엄격도의 경우 값을 크게 설정할수록 의사결정나무는 단순하게 되고 값이 작을수록 복잡한 형태를 가지게 된다. <표 3>은 가지치기 엄격도 값과 최종노드에 포함될 최소 사례수의 변경에 따른 실험결과이다. C5.0 의사결정나무에 의한 이탈고객 예측모델의 경우, 가지치기 엄격도 값이 85%이고 최종노드에 포함되는 최소 사례수가 10개일 때 가장 좋은 성능을 보였다.

<표 3> C5.0 의사결정나무 실험결과

가지치기 엄격도(%)	최소포함 사례 수	학습데이터 적중률(%)	검증데이터 적중률(%)	가지치기 엄격도(%)	최소포함 사례 수	학습데이터 적중률(%)	검증데이터 적중률(%)
75	10	64.4	57.6	85	30	77.1	62.7
75	20	67.1	59.7	85	40	75.6	61.0
75	30	66.4	57.3	90	10	73.8	60.4
75	40	72.9	62.3	90	20	76.7	61.2
80	10	65.1	58.8	90	30	73.8	59.8
80	20	78.7	62.8	90	40	75.9	60.1
80	30	75.3	60.6	95	10	80.2	57.0
80	40	75.0	60.8	95	20	82.3	58.1
85	10	79.8	63.2	95	30	79.2	60.1
85	20	80.1	61.1	95	40	78.5	58.9

4.3 사례기반추론에 의한 이탈고객 예측모델의 설계

사례기반추론은 새로운 문제가 주어지면 그 문제와 유사한 과거의 사례를 검색하여 그 사례의 해법(Solution)을 주어진 문제해결에 활용하는 기계학습 기법이다. 사례기반추론 모델을 구축하기 위해서는 먼저 해법이 알려진 과거사례들로 사례베이스(Case Base)를 구성해야 한다. 본 연구에서는 인공신경망 및 의사결정나무 모델의 학습에 사용된 학습데이터를 사례베이스로 사용하였다.

사례기반추론은 유사도 측정함수를 이용하여 새로운 입력사례와 사례베이스에 있는 과거사례 간의 유사도(Similarity)를 측정한다. 두 사례의 동일 속성 간의 유사도 점수 산정에는 유클리디안 거리(Euclidean Distance)가 가장 일반적으로 사용된다. 그러나 이것은 사례의 속성이 수치형일 경우에만 적용 가능하다는 제약을 가지고 있다. 따라서 범주형 속성의 경우 다른 형태의 유사도 점수 산정방법이 필요하다. 대표적으로 영역지식(Domain Knowledge)에 의한 부분일치(Partial Matching) 방법과 확률적 접근에 의한 유사도 점수부여방법이 있다(Wilson and Martinez, 1997). 본 연구에서는 해당 기업의 영역지식을 이용하여 <표 4>와 같이 유사도 점수를 산정하였다.

<표 4> 사례기반추론 모델의 유사도 점수 산정기준 (일부)

구분	속성명	속성형	유사도 점수 부여방법
1	고객 성별	범주형	If n = c Then S=10 Else S=0
2	고객 연령	수치형	If |n-c| <= 9 Then S= 10-|n-c| Else S = 0
3	보험 가입 경력	수치형	If |n-c| <= 9 Then S= 10-|n-c| Else S = 0
4	전계약사	범주형	If n = c Then S=10 Else S=0
5	차량 연식	수치형	If |n-c| <= 4 Then S= 10-(2*|n-c|) Else S = 0
6	납입 방법	범주형	If n = c Then S=10 Else S=0
7	수납 형태	범주형	If n = c Then S=10 Else S=0
8	고객 직업 대분류	범주형	If n = c Then S=10 Else S=0
9	보험 종목	범주형	If n = c Then S=10 Else S=0
10	차종 약칭	범주형	If n = c Then S=10 Else S=0
11	보험기간 내 사고여부	범주형	If n = c Then S=10 Else S=0
12	대인담보 가입여부	범주형	If n = c Then S=10 Else S=0
13	대물담보 가입여부	범주형	If n = c Then S=10 Else S=0
14	자손담보 가입여부	범주형	If n = c Then S=10 Else S=0
15	무보험담보 가입여부	범주형	If n = c Then S=10 Else S=0
…	…	…	…
46	가족운전 한정특약	범주형	If n = c Then S=10 Else S=0

(S = 유사도 점수, n = 입력사례의 속성값, c = 과거사례의 속성값)

본 연구에서는 <표 4>의 기준에 따라 사례의 속성간 유사도 점수를 부여했다. 각 속성간의 유사도 점수는 0~10점 사이로 표준화 하였으며, 한 사례의 총 유사도 점수도 0~10점 사이가 되도록 조정하였다. 즉, 총 유사도 점수는 각 속성의 유사도 점수의 합을 전체 속성의 수로 나눈 값이 된다.

식 (4-1)은 본 연구에서 사용된 두 사례간의 유사도 측정함수이다.

$$Similarity(N, C) = \frac{\sum_{i=1}^{l} f(n_i, c_i)}{l} \quad (4\text{-}1)$$

N : 새로운 입력사례.
C : 사례베이스에서 검색된 사례.
$f()$: 두 속성 사이의 유사도 측정함수.
n_i : 새로운 사례의 i번째 속성값.
c_i : 검색된 사례의 i번째 속성값.
l : 사례에 포함된 속성의 수.

사례기반추론 모델의 예측성능은 유사도 측정함수와 최종해를 구하는데 사용되는 최근접 이웃(Nearest Neighbor)의 수 k에 민감한 영향을 받는다. 본 연구에서는 최적의 사례기반추론 모델을 찾기 위해서 k를 1, 5, 10, 15, 20, 25, 30으로 변화시키며 실험을 진행하였다. 최종해의 생성에는 일반적으로 많이 사용되는 다수결 선택방식(Majority Voting Method)을 사용하였다.

<표 5> k값 변화에 따른 사례기반추론 모델의 실험결과

	k						
	1	5	10	15	20	25	30
학습데이터 적중률(%)	68.8	69.3	71.4	72.1	68.5	69.6	70.5
검증데이터 적중률(%)	60.9	62.3	65.6	64.3	62.7	63.0	64.5

사례기반추론에 의한 이탈고객 예측모델의 경우 <표 5>에 나타나 있듯이 k를 10으로 했을 때 가장 좋은 성능을 보여주었다. 이때 학습데이터와 검증데이터의 적중률은 각각 71.4%와 65.6%였다. 여기서 학습데이터 적중률이란 사례베이스 10,000건에 대해 leave-one-out 방식을 통해 산출된 결과이다.

5. DyMoS 다중모델의 설계

5.1 기본모델의 선정

DyMoS 다중모델을 구성하기 위한 첫 단계는 기본모델을 선정하는 것이다. 기본모델은 인공신경망, 의사결정나무, 그리고 사례기반추론 기법에 의해 생성된 단일모델들 중 가장 좋은 예측성능을 보이는 모델이 선정된다. <표 6>은 각각의 단일모델들에 대해 가장 좋은 예측성능을 보인 모델들의 예측적중률을 나타낸다.

<표 6> 단일모델들의 이탈고객 예측적중률

	검증데이터 적중률(%)	평가데이터 적중률(%)
인공신경망	64.5	64.2
의사결정나무	63.2	63.3
사례기반추론	65.6	65.5

단일모델들 중 가장 좋은 예측성능을 보인 모델은 사례기반추론 모델로 검증데이터와 평가데이터 적중률이 각각 65.6%와 65.5%였다. 따라서 사례기반추론 모델을 자동차보험 가입고객의 이탈예측을 위한 DyMoS 다중모델의 기본모델로 선정하였다.

5.2 판별모델의 설계

DyMoS 다중모델에 있어서 판별모델은 새로운 입력사례가 주어졌을 때 이 사례를 기본모델로 예측하는 것이 적합한지를 판별하는 메타모델(Meta Model)이다. 본 연구에서는 다음과 같은 단계를 통해 판별모델을 구축하였다.

• **제 1 단계 : 목표속성의 생성.**

기본모델인 사례기반추론 모델이 학습데이터(10,000건)와 검증데이터(10,000건)를 대상으로 예측을 수행한 예측값과 실제값을 비교하여 예측적중여부 속성을 생성한다. 예측적중여부의 속성값은 <표 7>에 의해 생성된다.

<표 7> 예측적중여부 속성값 생성규칙

실제 이탈여부	사례기반추론에 의한 예측값	예측적중여부
Yes	Yes	T
Yes	No	F
No	Yes	F
No	No	T

• **제 2 단계 : 학습데이터와 검증데이터의 구성.**

제 1 단계에 사용된 총 20,000건의 사례로부터 판별모델 생성에 사용될 학습데이터와 검증데이터를 구성하였다. 본 연구에서는 무작위 방식을 사용하여 학습데이터와 검증데이터를 7 : 3의 비율로 구성하였다. 따라서 판별모델의 생성에 사용된 학습데이터는 14,000건의 사례로 구성되었고, 검증데이터는 6,000건의 사례로 구성되었다.

• **제 3 단계 : 판별모델의 생성.**

제 3 단계에서는 예측적중여부를 목표속성(Target Feature)으로 하는 최적의 판별모델을 찾는다. 본 연구에서는 판별모델의 생성을 위해 C5.0 의사결정나무 알고리즘을 사용하였다. 최적의 판별모델을 찾기 위해 제 4.2절에서 단일모델을 생성하기위해 수행했던 실험방식을 동일하게 적용하였다. 그 실험결과는 <표 8>과 같다. 그 결과, 가지치기 엄격도 값이 80%이고 최종노드에 포함되는 최소 사례수가 20개일 때 가장 좋은 판별모델이 도출되었으며, 이때의 검증데이터 적중률은 79.8%였다.

<표 8> 의사결정나무를 이용한 판별모델실험결과

가지치기 엄격도(%)	최소포함 사례수	학습데이터 적중률(%)	검증데이터 적중률(%)	가지치기 엄격도(%)	최소포함 사례수	학습데이터 적중률(%)	검증데이터 적중률(%)
75	10	78.9	75.7	85	10	79.4	74.5
75	20	80.3	76.1	85	20	78.3	70.2
75	30	79.9	71.7	85	30	80.5	73.7
75	40	84.5	73.6	85	40	76.8	69.8
80	10	87.3	78.1	90	10	77.8	74.1
80	20	87.1	79.8	90	20	79.3	73.5
80	30	84.9	75.8	90	30	80.6	69.5
80	40	85.1	70.6	90	40	81.3	68.9

5.3 지원모델의 설계

지원모델은 판별모델에 의해서 기본모델로 예측하는 것이 부적합하다고 판별된 사례들만을 예측하는 모델이다. 즉, 지원모델은 전체 문제영역에서 기본모델이 처리하지 못하는 문제영역만을 처리하기 위한 모델이다. 따라서 지원모델의 생성에는 기본모델이 잘못 예측한 사례들만이 사용된다. 본 연구에서는 다음과 같은 단계를 통해 지원모델을 생성하였다.

• **제 1 단계 : 대상사례의 추출.**

단일모델 생성에 사용된 학습데이터와 검증데이터에 대해 기본모델로 선정된 사례기반추론 모델이 잘못 예측한 사례들만을 추출하여 지원모델 생성을 위한 데이터를 구성한다. 그 결과 총 20,000건의 사례 중 6,300건의 사례가 추출되었다.

• **제 2 단계 : 학습데이터와 검증데이터의 구성.**

제 1 단계에서 추출한 총 6,300건의 사례를 7 : 3의 비율로 나누어 학습데이터(4,410건)와 검증데이터(1,890건)를 구성하였다.

• **제 3 단계 : 지원모델의 생성.**

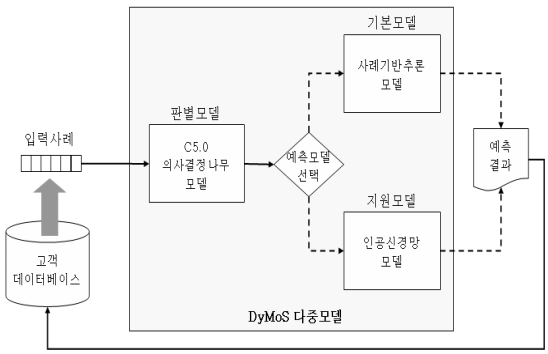
지원모델의 생성에는 인공신경망을 사용하였고, 제 4.1절에서 사용한 실험방식을 동일하게 적용하였다. 단, 은닉층은 한 개만 설정하였다. 그 결과가 <표 9>에 제시되어 있다. 실험결과에서 알 수 있듯이, 인공신경망에 의한 지원모델은 은닉층의 처리요소 수를 35개로 했을 때 검증데이터에 대해 가장 좋은 적중률(64.4%)을 보여주었다. 따라서 이 모델을 DyMoS 다중모델의 지원모델로 선정하였다.

<표 9> 인공신경망을 이용한 지원모델의 실험결과

은닉층 처리요소 수	학습데이터 적중률(%)	검증데이터 적중률(%)
10	65.3	60.4
15	66.2	59.5
20	65.3	62.8
25	68.4	61.2
30	67.8	62.6
35	69.6	64.4
40	68.7	61.6
45	65.3	59.1
50	65.7	58.5
55	62.3	59.7

5.4 DyMoS 다중모델의 완성

이 절에서는 기본모델, 판별모델, 그리고 지원모델을 결합하여 DyMoS 다중모델을 완성하고 그 결과를 단일모델과 비교 평가하였다. 본 연구에서 구축된 자동차보험 가입고객의 이탈예측을 위한 DyMoS 다중모델은 [그림 6]과 같다.



[그림 6] 자동차보험 가입고객 이탈예측을 위한 DyMoS 다중모델

DyMoS 다중모델의 예측성능평가에는 단일모델의 예측성능평가에 사용된 10,000건의 평가데이터를 동일하게 사용하였다. 실험결과 평가데이터에 대한 DyMoS 다중모델의 예측 적중률은 71.7%로 나타났으며, 그 결과가 <표 10>에 나타나 있다.

<표 10> DyMoS 다중모델에 의한 이탈고객 예측결과

		예측		합계
		갱신	이탈	
실제	갱신	3,619 (36.2%)	1,761 (17.6%)	5,380 (53.8%)
	이탈	1,069 (10.7%)	3,551 (35.5%)	4,620 (46.2%)
합계		4,688 (46.9%)	5,321 (53.1%)	10,000 (100%)

<표 11>은 DyMoS 다중모델과 단일모델의 평가데이터 적중률을 보여준다. 평가데이터에 대한 DyMoS 다중모델의 적중률은 71.7%로서 단일모델에서 가장 좋은 성능을 보여준 사례기반추론 모델의 적중률 65.5% 보다 6.1% 포인트 높았다.

<표 11> 단일모델과 DyMoS 다중모델의 평가데이터 적중률

		평가데이터 적중률(%)
단일 모델	인공신경망	64.2%
	의사결정나무	63.3%
	사례기반추론	65.5%
DyMoS 다중모델		71.7%

이러한 성능차이가 통계적으로 유의한 것인지를 확인하기 위해 McNemar 검정을 수행하였다. McNemar 검정은 실험전의 측정치와 실험후의 측정치에 차이가 있는지를 검정하는 통계적 기법으로, 분류문제에 있어서 두 기법간의 성능에 차이가 있는지를 평가하는데 주로 사용된다. <표 12>은 평가데이터에 대한 DyMoS 다중모델과 인공신경망, C5.0 의사결정나무, 그리고 사례기반추론에 의해 생성된 단일모델과의 McNemar 검정결과를 보여준다.

<표 12> 평가데이터에 대한 McNemar 검정결과

	인공신경망	C5.0	사례기반추론
DyMoS 다중모델	2270.4*	2171.9*	2440.6*

* 1% 수준에서 유의함

<표 12>에 나타나있듯이 DyMoS 다중모델이 1% 수준에서 모든 단일모델보다 이탈고객 예측에 있어서 더 우수한 성능을 보였다. DyMoS 다중모델의 이러한 성능향상은 사례기반추론 모델만을 사용했을 때 놓칠 수 있는 문제영역의 숨은 패턴을 지원모델인 인공신경망 모델이 보완해 줌으로써 얻어진 결과라고 볼 수 있다.

6 결론

본 연구에서는 분류문제에 있어서 예측성능을 향상시킬 수 있는 새로운 형태의 다중모델 구축방법을 제시하였다. 본 연구에서 제시한 DyMoS 다중모델의 기본사상은 입력사례의 특성에 따라 적절한 예측모델을 사용함으로써 단일모델의 사용에서 발생하는 예측오류를 최소화하는데 있다. 본 연구에서 제시된 DyMoS 다중모델은 판별모델, 기본모델, 그리고 지원모델로 명명된 3개의 단위모델로 구성되었다. 판별모델은 새로운 입력사례가 주어졌을 때, 이 사례를 기본모델과 지원모델 중 어떤 모델을 사용해서 예측하는 것이 적합한지를 결정해 주는 모델이다. 따라서 하나의 입력사례에 대해서, 판별모델의 판별결과에 따라 예측을 수행할 모델이 결정된다. 기본모델은 해당 문제영역에 있어서 대부분의 문제들을 처리하기 위한 모델로 본 연구에서는 사전에 생성한 단일모델들 중 가장 좋은 성능을 보인 모델을 기본모델로 선정하였다. 지원모델은 기본모델이 잘못 예측하는 영역에 대해 예측을 수행하는 모델로 기본모델이 예측에 실패한 사례만을 가지고 만들어진다.

본 연구에서 제시한 DyMoS 다중모델의 실증적인 평가를 위해서, 국내 자동차보험사의 고객이탈 예측문제에 적용한 후 그 결과를 인공신경망, 의사결정나무, 그리고 사례기반추론 기법을 사용한 단일모델들과 비교하였다. 본 연구에서 이탈고객 예측을 위해 구축된 DyMoS 다중모델에서는

판별모델로 C5.0 의사결정나무 모델이 사용되었고, 기본모델로 사례기반추론 모델이 사용되었으며, 지원모델로 인공신경망 모델이 사용되었다. 실험결과, 평가데이터에 대해서 단일모델의 예측성능은 사례기반추론 모델이 65.6%로 가장 높았고 인공신경망 모델이 64.5%, C5.0 의사결정나무 모델이 63.2%이었다. 본 연구에서 제시한 DyMoS 다중모델의 적중률은 71.7%로서, 단일모델에서 가장 좋은 성능을 보여준 사례기반추론 모델의 적중률 보다 6.1% 포인트 높았다. 물론 본 연구에서 제시한 DyMoS 다중모델의 결합방법이 일반적으로 모든 문제영역에 대하여 적합하지는 않을 수 있다. 하지만 문제영역의 복잡성이 큰 분류문제에 있어서 단일모델의 사용에서 오는 예측성능저하 문제를 해결할 수 있는 새로운 대안을 제시하였다는 점에서 본 연구의 의의가 있다.

본 연구의 한계점 및 향후 연구로 다음의 몇 가지를 들 수 있다. 첫째, 본 연구에서 제시한 DyMoS 다중모델의 성능평가에 다양한 모델이 반영되지 못했다는 점이다. 즉, 기존 문헌에 소개된 다른 형태의 다중모델과의 성능비교 실험을 추가적으로 수행할 필요가 있다. 둘째, DyMoS 다중모델을 구성하는 요소모델의 생성에 적용되는 모델링 기법에는 기본 원리상 아무런 제약을 둘 필요가 없다. 본 연구에서는 세 가지의 기계학습 기법에 국한하여 모델링 기법을 사용하였으나 향후에는 다양한 기법을 사용한 연구를 할 필요가 있다. 마지막으로 다양한 문제영역에 DyMoS 다중모델을 적용해 봄으로써 그 유용성을 검증하는 추가적인 연구가 필요하다.

참고문헌

[1] 이재식, 전용준, “사례기반 추론을 위한 동적 속성 가중치 부여 방법”, *한국지능정보시스템학회논문지*, 7권 1호(2001), 47-61.

[2] 이재식, 차봉근, “유전적 알고리즘을 이용한 인공신경망의 구조 설계”, *한국경영과학회지*, 24권 3호(1999), 49-62.

[3] 한인구, 신경식, “지능형 중기업 신용평가시스템이 개발 및 활용 : 보람은행의 사례를 중심으로,” *Information Systems Review*, 1권 1호(1999), 51-61.

[4] Berry, M. and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, 2nd Ed., Wiley Pub., Inc., 2004.

[5] Daskalaki, S., I. Kopanas, M. Goudara, and N. Avouris, “Data Mining for Decision Support on Customer Insolvency in Telecommunications Business”, *European Journal of Operational Research*, Vol.145(2003), 239-255.

[6] Giacinto, G. and F. Roli, “Dynamic Classifier Selection based on Multiple Classifier Behaviour”, *Pattern Recognition,* Vol.34(2001), 1879-1881.

[7] Hothorn, T. and B. Lausen, “Bagging Tree Classifiers for Laser Scanning Images : A Data- and Simulation-Based Strategy”, Artificial Intelligence in Medicine, Vol.27(2003), 65-79.

[8] Hsieh, N. C., “Hybrid Mining Approach in the Design of Credit Scoring Models”, *Expert Systems with Applications*, Vol.28(2005), 655-665.

[9] Kim, E. J., W. J. Kim, and Y. B. Lee, “Combination of Multiple Classifiers for the Customer’s Purchase Behavior Prediction”, *Decision Support Systems*, Vol.34(2002), 167-175.

[10] Kim, Y. S., W. N. Street, and F. Menczer, “Optimal Ensemble Construction via Meta-

evolutionary Ensembles", *Expert Systems with Applications*, Vol.30, No.4(2006), 705-714.

[11] Kuncheva, L. I., "Using Measures of Similarity and Inclusion for Multiple Classifier Fusion by Decision Templates", *Fuzzy Sets and Systems*, Vol.122(2001), 401-407.

[12] Kwon, Y. S. and K. C. Lee, "Ordinal Pairwise Partitioning (OPP) Approach to Neural Networks Training in Bond rating", *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol.6(1997), 23-40.

[13] Last, L., A. Kandel, and O. Maimon, "Information-Theoretic Algorithm for Feature Selection", *Pattern Recognition Letters*, Vol.22(2001), 799-811.

[14] Nelson, M. M. and W. T. Illingworth, *A Practical Guide to Neural Nets*, Addison-Wesley, 1991.

[15] Opitz, D. and R. Maclin, "Popular Ensemble Methods : An Empirical Study", *Journal of Artificial Intelligence Research*, Vol.11 (1999), 169-198.

[16] Quinlan, R., C5.0 : An Information Tutorial, RuleQuest, http://www.rulequest.com/see5-unix.html, 1998.

[17] Sexton, R. S. and N. A. Sikander, "Data Mining Using a Genetic Algorithm Trained Neural Network," *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol.10(2001), 201-210.

[18] Wang, X. and H. Wang, "Classification by Evolutionary Ensembles", Pattern Recognition, Vol.39 No.4(2006), 595-607.

[19] Wei, C. P. and I. T. Chiu, "Turning Telecommunications Call Details to Churn Prediction : a Data Mining Approach", *Expert Systems with Applications*, Vol.23 (2002), 103-112.

[20] Wilson, D. R. and. T. R. Martinez, "Improved Heterogeneous Distance Functions", *Journal of Artificial Intelligence Research*, Vol.6(1997), 1-34.

[21] Yang J. and V. Honavar, "Feature Subset Selection Using a Genetic Algorithm", *IEEE Intelligent Systems and their Applications*, Vol.13(1998), 44-49.

[22] Zhang, P., B. Verma, and K. Kumar, "Neural vs. Statistical Classifier in Conjunction with Genetic Algorithm based Feature Selection", *Pattern Recognition Letters*, Vol.26(2005), 909-919.

[23] Zhou, Z. H., J. X. Wu, Y. Jiang, and S. F. Chen, "Genetic Algorithm based Selective Neural Network Ensemble", In Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, WA, Vol.2(2001), 797-802.

Abstract

Customer Churn Prediction of Automobile Insurance by Multiple Models

Jae Sik Lee* · Jin Chun Lee*

Since data mining attempts to find unknown facts or rules by dealing with also vaguely-known data sets, it always suffers from high error rate. In order to reduce the error rate, many researchers have employed multiple models in solving a problem. In this research, we present a new type of multiple models, called DyMoS, whose unique feature is that it classifies the input data and applies the different model developed appropriately for each class of data. In order to evaluate the performance of DyMoS, we applied it to a real customer churn problem of an automobile insurance company. The result shows that the DyMoS outperformed any model which employed only one data mining technique such as artificial neural network, decision tree and case-based reasoning.

Key words : Multiple Models, Hybrid Model, Customer Churn Prediction, Data Mining, Case-based Reasoning, Artificial Neural Networks, Decision Tree

* School of e-business, Ajou University