이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구

A Study on the Combined Decision Tree(C4.5) and Neural Network Algorithm for Classification of Mobile Telecommunication Customer

저자
(Authors)

이극노, 이홍철

출처
(Source)

지능정보연구 9(1), 2003.6, 139-155 (17 pages)
Journal of Intelligent Information Systems 9(1), 2003.6, 139-155 (17 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/Article/NODE00409814

APA Style

이극노, 이홍철 (2003). 이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구. 지능정보연구, 9(1), 139-155.

한국지능정보시스템학회논문지 제9권 제1호 2003년 6월(pp.139~155)

이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구

이극노
고려대학교 산업시스템정보공학과
(nick-kno@korea.ac.kr)

이홍철
고려대학교 산업시스템정보공학과
(hclee@korea.ac.kr)

본 논문은 결합된 의사결정 나무(C4.5)와 신경망기법을 적용함으로써 고객의 신용에 대한 예측을 높이기 위하여 이동통신 고객의 패턴을 분류하고, 분석하는 새로운 방법에 대하여 연구하였다. 의사 결정나무(C4.5)를 형성하여 선택된 결정변수와 함께 규칙을 생성함으로써, 신경망의 입력벡터 값을 정의하는 체계적인 방법을 제시하였다. 고객 관리측면에서 본 논문은 이동 통신 회사의 기존고객을 분류하여 패턴을 분석함으로써 우수한 고객의 지속적인 관리와 이탈 가능성이 많은 고객을 차별 관리하여 기업이익을 증대 시킬 수 있을 것이다. 또한 이러한 분류를 통하여 신규 고객에 반영함으로써 고객의 향후 관리에도 기여할 수 있을 것이다. 실제 이동통신 고객데이터를 중심으로 연구의 결과는 예측의 정확도가 기존의 의사결정 트리 모델 (CART, C4.5), 회귀모형, 신경망 접근 방법과 기존에 연구되었던 결합모델(CART & 신경망)보다 훨씬 높게 연구 되었다.

논문접수일 : 2002년 12월 게재확정일 : 2003년 4월 교신저자 : 이홍철

1. 서론

통신업계의 정보시스템 구축은 데이터 웨어하우스를 이용한 고객관리와 통신망 관리로 크게 나눌 수 있다. 고객관리 측면에서 보면 신규 고객 유치를 위한 투자에 박차를 가하는 반면, 한편으로는 기존고객의 이탈로 인한 손해를 감수해야 하는 실정이다. 이러한 상황에서 통신업계는 기업의 이익을 위한 고객의 관리와 신규고객의 유지를 위하여, 신상품개발, 세분화된 요금전략, 유통망 최적관리, 수요예측, 위험관리 등이 연구되고 있으며, 이를 통해 기업의 경쟁력을 확보하기 위한 전략을 수립하고 있다.

기존고객에 대한 해지고객의 특성을 분석함으로써 해지할 가능성이 높은 고객의 특성을 분석하는 연구가 이루어지고 있지만, 기업 내 각 부서별로 분산된 고객데이터의 수집을 위하여 하나의 데이터 웨어하우스를 구축하는 것과 쿼리 앤 리포트, OLAP, 데이터마이닝 등을 이용하여 데이터를 분석 할 수 있는 전략적 데이터 활용이 필요한 실정이다(Hasan and Hyland, 2001). 특히 이러한 데이터를 분석하는데 있어 단일 데이터마이닝 기법만을 이용한 데이터 분석이 아닌, 좀더 깊이 있는 알고리즘 사이에 비교를 통하여 정확한 예측을 할 수 있는 기법이 요구된다.

이러한 신용평가 모델을 위하여 데이터마이닝 기법이 많이 사용되고 있다. 이중에 단일 데이터마이닝 기법을 사용하는 방법보다는 퍼지, 신경

망, 유전자 알고리즘 등의 타 기법들과 그것의 변형된 알고리즘의 결합을 통하여 예측의 정확도(Accuracy)를 향상 시키는 방법들이 소개되고 있다.

특히, 분류 모델에서 훈련 집합(Training Data Set)의 크기 증가는 학습된 분류 모델의 정확성을 증가하게 만든다. 하지만, 많은 데이터의 학습으로 인하여 공간적 복잡성과 수행시간이라는 문제를 발생하게 된다. 이와 같이 훈련 집합의 적정 데이터 크기와 함께 예측의 정확성이라는 학습목표가 알고리즘 형성에 커다란 연구 주제로 자리 잡게 되었다. 이러한 문제 해결을 위하여 훈련 집합의 데이터를 나누거나 관련된 특성을 이용하여 크기를 조절하는 방법론 등이 고려되고 있다(Mitra, Pal and Mitra, 2002).

본 연구는 데이터마이닝 기법 중에 의사결정나무(C4.5)를 이용하여 규칙들을 생성하고, 이 규칙들과 함께 C4.5나무로부터 생성되는 결정변수들을 선택하여 신경망의 새로운 입력 변수로 적용함으로써 분류 모델의 예측 정확도를 향상시키기 위한 결합된 모델을 제안한다. 이 결합된 모델을 통하여 예측의 정확도 향상이라는 관점에서 기존의 단일 데이터마이닝 모델이 가졌던 문제점인 훈련집합 집합의 크기, 적정 변수선정, 모델을 해석할 수 있게 하는 부분적 이해 등을 해결하고자 한다.

결합모델을 검증하기 위하여 이동통신 고객데이터를 적용하였으며, 기존 데이터마이닝 기법인 CART, C4.5등과 Kao과 Chiu에 의해 연구된 CART와 신경망의 결합된 모델을 비교하여 실험하였다.

본 논문은 서론부분과 다음 절에서는 기존연구의 고찰과 이론적 배경을 설명하였고, 이후, 의사결정나무와 신경망의 결합된 모델을 소개하였다. 이를 검증받기 위하여 각 모델을 비교 실험하였으며, 결합모델을 적용한 고객 분류 시스템을 구현하였으며, 끝으로 결론부분으로 구성하였다.

2. 기존연구의 고찰

빠른 기술의 성장으로 많은 정보를 쉽게 습득하고 데이터베이스에 저장 할 수 있게 되었다. 이렇게 저장된 데이터는 무한히 증가하게 되었고, 필요한 데이터를 통하여 의사결정에 필요한 정보를 얻기란 쉬운 일이 아니게 되었다. 데이터베이스 내의 지식발견(Knowledge of Discovery in Database, KDD)을 위하여 데이터 웨어하우스 또는 다른 정보 보관 장소들에 저장된 많은 양의 데이터로부터 구조와 연상, 숨은 패턴을 발견하기 위하여 정교화된 알고리즘을 적용하는 특별한 단계가 데이터마이닝이다. 이를 수행하기 위하여 사용되는 방법론으로는 퍼지집합(Fuzzy Set), 신경망(Neural Network), 유전자 알고리즘(Genetic Algorithm), 러프 집합(Rough Set), 의사결정나무(Decision Tree) 등을 들 수 있다(Mitra, Pal and Mitra, 2002).

본 논문에 사용된 데이터마이닝의 방법론 중에 하나인 의사결정나무는 분류와 예측을 하는데 효과적으로 사용된다. 의사결정나무는 적용결과에 대하여 명확하고 쉽게 이해 할 수 있도록 도와주고, 의사결정나무의 예측 정확도는 다른 분류모델보다 높거나 동등하며, 나무를 만드는데 있어서 분석자로부터 입력매개변수를 요구하지 않기 때문에 많이 사용하는 방법이다(Ganti, Gehrke and Ramakrishnan, 1999).

이러한 의사결정나무의 예측치의 향상을 위하

여 분리 기준(Splitting Criteria)과 전체나무의 크기를 최적화하는 방법에 기존 연구는 많은 관심이 모아졌다. 전체 나무의 크기를 최적으로 하면서 정확성을 최대로 만족하기란 좀처럼 해결되지 않는 문제로 남아있다. 이러한 문제를 해결하기 위하여 의사결정나무를 형성하는 다양한 알고리즘이 있으며, 이 논문에 사용된 C4.5알고리즘은 전체나무의 최적을 위해 많은 연구에 이용되고 있다(Ho, 1998), (Anderieko, 1999), (Ruggieri, 2002).

그러나 의사결정나무의 경우 비연속과 비안전성의 문제를 가지고 있다. 비연속성의 경우, 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에는 예측 오류가 클 가능성이 있다. 더욱이, 훈련용 자료(Training Data)에만 의존하는 의사결정나무는 새로운 자료의 예측에는 불안정할 가능성이 높다. 따라서 이러한 문제를 해결하기 위하여 검증용 자료(Test Data)를 이용한 모델 평가나 신경망을 적용한 예측 기법이 요구된다(최종후, 2001).

이러한 신경망은 자료 분석 분야에서 복잡한 구조를 가지고 있는 자료에 대하여 예측 문제를 해결하기 위한 유연한 비선형 모형의 하나로 분류 될 수 있다. 인간의 신경생리학과의 유사성 때문에 일반적으로 다른 통계적 예측모형에 비해 보다 흥미롭게 연구 되어지고 있다. 특히, 예측 기법으로써 기존의 로지스틱 회귀모형 보다 신경망의 우수함을 비교한 연구들이 고려되고 있다(Paruelo and Toma sel, 1997), (Maher and Sen, 1997).

그러나 신경망은 미래의 목표 값을 예측하는데 있어 입력벡터의 값의 수나 형태를 결정할 수 있는 체계적인 방법의 결여와 모델의 분류가 어떻게 이루어지는지 명확하게 이해 할 수 없는 단점이 제시 되고 있다(Kao and Chiu, 2001). 이러한 단점을 해결하기 위하여 신경망에서 상징적 분류 규칙을 찾거나 의사결정 나무를 통하여 이해 할 수 있는 해석을 얻고자 하는 연구 등이 이루어졌다(Lu, Setiono, and Liu, 1996). 이 논문에서는 이러한 적정 입력변수 선정과 해석의 어려움을 해결하기 위하여 제안하는 결합모델에 의사결정나무를 적용하였다.

현재 각각의 알고리즘의 단점을 개선하고자 많은 연구에서 단일기법보다는 의사결정나무와 신경망 등과 같이 변형된 알고리즘의 결합을 통하여 예측의 정확도를 향상시키는 연구가 이루어지고 있다. Kao과 Chiu는 CART의 결과를 새로운 신경망의 입력변수로 선정하여 결합모델을 형성하였다. 이러한 결합모델이 기존의 CART 또는 신경망보다 정확하게 분류함을 보이고 있다(Kao and Chiu, 2001). 또한, 의사결정나무와 신경망을 이용한 모델이 신경망, 의사결정나무(CART, AMIG)보다 효율적으로 에러율이 감소함을 보이고 있다(Sethi,1990). Sethi의 방법에 분리자로써 C4.5를 사용하여 신경망을 구성함으로써, 실시간 이미지를 인지하는 문제에도 적용되었다(Chung, Wong and Bergmann, 1998). 신경망을 선행학습기로 적용하여 훈련 집합에 보간(Interpolated)된 샘플들의 결과로부터 CART에 접목시키는 ANN-DT를 제안되었으며 기존의 의사결정나무보다 규칙을 잘 설명하는 연구가 이루어졌다(Schmitz, Alsdrich and Gouws, 1999). 또한 지식습득을 위하여 신경망을 사용하여 의사결정나무에 접목시키는 새로운 방법을 제시하여 정확도를 개선시키고자 하였다(Zorman and Peter, 2002). 가지치기된 C4.5를 신경망에 이용하는 방법을 제안함으로써, 가지치기된 나무를 최적화하는 연구가 제시되었으며, C4.5로부터 얻

은 결과를 유전자 알고리즘 운영자로 사용하여 의사결정 트리를 구성함으로써, 의사결정나무의 크기를 줄이는 방법이 제시되었다(Kijsirikul and Kongsak, 2001),(Endou and Zhao, 2002). 이와 같이 많은 분야에 결합모델을 적용하여 우수함을 보이는 연구가 지속적으로 이루어지고 있다.

따라서 이 논문에서는 기존의 데이터마이닝 기법들을 결합하는 방법을 사용하여 예측의 정확성을 향상 시키고자 한다. 전체 분류 모델의 정확성을 향상 시키는 목적에 부합하는 결정변수를 선택하는 방법, 즉, 목표변수에 큰 영향을 주는 변수를 도출해 낼 수 있는 방법과 규칙 생성자를 이용하여 신경망에 새로운 입력변수로 선정함으로써 입력벡터 값의 수나 형태를 결정하기 위한 체계적인 방법을 제안 한다. 나아가, 제시한 결합모델을 통하여 이동통신시장에 적용함으로써, 신규고객과 기존고객을 분류하고 패턴을 분석하여 우수고객의 지속적인 관리 할 수 있고, 연체고객, 이탈 가능성이 많은 고객을 예측 관리하여 기업의 비용을 줄일 수 있을 것이다.

3. 의사결정나무(C4.5)의 형성과 결합모델의 생성

C4.5의사결정나무와 신경망 알고리즘의 결합은 의사결정나무의 장점인 규칙생성과 이를 이용한 변수선정으로 신경망에 결합시킴으로써 보다 좋은 예측의 결과를 얻고자 하는 것이다. 결합모델의 구성하는데 있어 의사결정의 경우, 다양한 분리 기법이 존재한다. 최근에 의사결정나무의 최적을 위하여 C4.5에 관한 많은 연구가 이루어지고 있다. 특히, 분리기법 중에 다지분리(C4.5)의 경우 분류의 가지가 많아짐에 따라 복잡해 질 수 있지만 적은 깊이에서도 예측이 좋아 질 수 있다. 또한 규칙을 쉽게 표현 할 수 있는 장점이 있다. 그러나 이진분리(CART)의 경우에는 정확한 분류를 위하여 나무의 깊이가 깊어지고 복잡해지는 단점이 존재하지만 해석이 용이하다는 장점이 있다. 각 분리기준에는 각 상황에 맞는 장, 단점이 존재하게 된다. 이 논문에서는 의사결정나무로써 C4.5의 다지분리를 이용하여 신경망과 결합한 모델로 이루어진다.

3.1 C4.5의사결정나무의 형성

C4.5는 J Ross Quinlan에 의해 수정 발전된 의사결정 알고리즘이다. 초기 버전이 ID3(Iterative Dichotomizer 3, 1986)는 기계학습 분야에 많은 영향을 주었다. CART가 각 마디에 이원분할을 형성하며 이지분리 나무구조를 만드는데 반하여, C4.5는 각 마디가 다지분리의 구조를 갖는 나무로 구성된다. 연속형 변수의 경우 분산을 이용하지만, C4.5는 범주형 변수의 경우 분할자로 엔트로피지수를 이용하고, CART는 지니 지수를 이용하여 구하게 된다(최국렬, 2001).

C4.5의사결정나무를 형성하기 위하여 처음 수행하는 작업이 분할정복(Divide and Conquer)이다. 입력되는 훈련 집합이 성공적으로 분할 되도록 모든 하부 집합에 하나의 클래스가 속하는 경우들로 구성될 때까지 나무를 형성한다(Quinlan, 1993).

정보이익비율(Information gain ratio)이 노드를 분리하는 기준으로 사용된다. 주어진 예를 분류하기 위하여 요구되는 평균정보(Average Information)를 가장 감소시킬 수 있는 방법으로 현재의 훈련 집합을 분리하기 위한 것이다.

전체 훈련 집합 |S|와 함께, 현재의 훈련 집합

을 S이라하고, 클래스 C_i (i=1,2,...,N)에 속하는 경우(Case)의 수를 Freq(Ci, S)라 하면, 주어진 예의 클래스를 확인하기 위하여 요구되는 평균 정보(Entropy)는 식 1과 같다.

$$Info(S) = -\sum_{i=1}^{n} Freq(C_i, S)/|S| \times \log_2(Freq(C_i, S)/|S|) bits \quad (1)$$

이때, 어떤 시험 X로부터, S가 n개의 하부 집합 $S_1, S_2, \ldots, S_n$으로 분리된다면, 정보이익(Information gain)은 식 2와 같다.

$$gain(X) = Info(S) - \sum_{i=1}^{n} |S_i|/|S| \times Info(S_i) \quad (2)$$

여기서, 정보이익 비율(Information gain ratio)은 식3과 같이 구하게 된다.

$$gian\ ratio(X) = gain(X)/split\ Info(X) \quad (3)$$

where

$$Split\ Info(X) = \sum_{i=1}^{n} |S_i|/|S| \times \log_2(|S_i|/|S|) \quad (4)$$

이러한 분리기준은 이익비율기준(gain ratio)이 이익기준(gain)보다 실험에서 훨씬 좋은 결과를 제시한다. 따라서 분리기준으로 이익비율을 사용한다(Quinlan, 1988).

순환적으로 마디가 나누어지는 과정에 따라 더 이상 개선되지 않을 때까지 나무가 형성된다. 이때 나무구조가 매우 복잡한 경우가 발생하며, 이를 데이터의 과적합(Overfits)이라고 한다. 어떤 경우에는 나무가 복잡해짐에 따라 간단한 나무보다 더 높은 에러율을 가지게 된다. C4.5의 경우 식5와 같이 에러율을 계산한다.

$$Error\ rate = U_{cf}(E, N) \quad (5)$$

(E:Event, N:Trials, U_{cf}:Upper limit is the confidence for the binomial distribution, Cf:Given confidence level)

이러한 에러율을 기초로 하위나무 전체의 예측된 에러수가 상위 잎의 예측된 에러의 수 보다 크다면 상위 잎으로 대체한다. 이렇게 가지치기 함으로써 예측된 에러율을 더 낮게 줄일 수 있다. 이러한 에러율은 규칙 생성에 유용하게 이용된다. 가지치기를 함으로써 크기를 최적으로 하는 C4.5의사결정나무를 형성하게 된다.

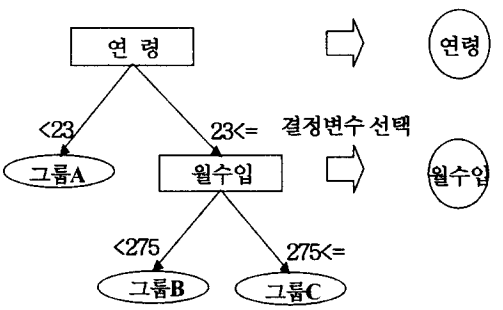
3.2 규칙생성

형성된 의사결정나무가 거대해지면 해석의 어려움이 존재하게 된다. 산술규칙을 생성함으로써 전체나무에 대하여 해석을 용이하게 해주고, 이러한 규칙들은 나무가 분리한 것만큼 정확하게 분리 할 수 있게 한다. 일련의 규칙이 많아지면 혼란할 수 있는데, C4.5는 허용에러를 계산하여 규칙의 정확성에 영향을 주지 않는 조건은 제거하여 간단히 표현할 수 있다. 또한 규칙의 우선순위를 결정하여, 디폴트 클래스를 형성하게 된다. 이러한 규칙 생성과정은 다음과 같다.

1) 교대로 각각의 클래스에 모든 간단히 표현된 규칙들을 전체 규칙 집합의 정확성에 기여하지 않는 규칙은 제거한다.
2) 적은 긍정적 잘못(False Positive Error)을 가지는 규칙을 포함하는 하위 집합의 클래스가 첫 번째 클래스에 놓이게 되고, 디폴트 클래스가 선택되어진다.

3.3 결정변수의 선택

결합모델의 새로운 입력변수를 형성하기 위하여 C4.5알고리즘에 의해 만들어진 의사결정나무로부터 결정변수를 선택하게 된다. 결정변수는 각 마디를 형성하는 변수들로 구성된다. 이는 분할정복에 의해 부모와 자식마디까지 형성되었던 최고의 이익비율을 가지는 각 마디의 변수들을 이용하여 선택하게 된다. <그림 1>은 훈련 집합에 의해 형성된 의사결정나무로부터 결정변수를 선택하는 과정을 보여주고 있다. 입력되는 훈련 집합의 예측변수가 성별, 연령, 월수입, 직업 등이고, 목표변수가 그룹이라면, 이로부터 형성된 의사결정나무의 분리기준이 되었던 변수를 결정변수로 선택하게 된다. <그림 1>의 경우 연령과 월수입이 된다. 즉, 목표변수에 큰 영향을 주지 않는 변수를 제거함으로써, 기존의 고려되었던 변수보다 적은 변수가 새로운 모델의 형성에 필요한 결정변수가 된다.



<그림 1> 결정변수의 선택과정

3.4 결합모델의 구축

결합모델을 구성하기 위하여 학습데이터(Train- ing Dataset)가 입력된다. 입력된 학습데이터는 C4.5 알고리즘에 의해 필요한 결정변수와 규칙 생성을 위하여 의사결정나무를 형성한다. 나무 전체의 예측율을 높게 형성해주는 정지기준과 가지치기를 실행하여 의사결정나무를 형성하고, 규칙을 산출하게 된다. 이렇게 형성된 의사결정나무에 제안된 방법으로 결정 변수를 구하고, 이후 제안할 생성규칙에 의해 분류된 결과를 신경망 구조의 새로운 입력 변수를 선정하기 위하여 적용하게 된다.

의사결정나무로부터 신경망의 새로운 입력변수를 추출해내는 방법으로는 기존에 기술했던 의사결정나무를 형성하는데 사용된 최고의 이익비율을 기준으로 각 마디를 형성하는데 사용된 결정변수들을 적용한다.

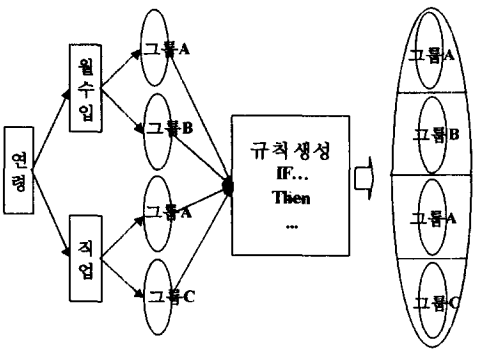
그리고 규칙을 적용하는 방법으로는 다음과 같이 세 가지 방법을 제시하며, 이 세 가지 방법 모두 결합된 모델의 예측 정확도가 향상됨을 실험에서 확인할 수 있었다. 이러한 규칙 생성의 결과를 적용하는 세 가지 방법으로는 다음과 같다.

- 방법1: 의사결정 나무로부터 결정변수와 규칙 산출자에 의해 생성된 결과를 설명변수로써, 하나의 새로운 입력변수 내에 각 클래스의 분류 결과를 적용하는 방법이다(그림 3).
- 방법2: 부모마디로부터 끝마디까지의 깊이에 의해서 분류된 범주를 하나의 구성원으로 보고 이로부터 생성된 규칙을 각각의 새로운 클래스로 재설정하여, 하나의 새로운 입력변수로 분리하는 방법이다.
- 방법3: 위의 두 방법을 모두 입력변수로 적용하는 방법으로 두개의 새로운 입력변수가 생성된다.

<그림 2>는 방법1을 이용하여 입력변수를 구축한 것이다. 방법1에서 규칙 생성과정을 살펴보

면, C4.5의 규칙 생성과정에서 이미 기술했던 것과 같이 산술규칙을 생성하여 모든 잎에 하나의 규칙을 형성하게 된다. 이러한 규칙은 허용에러를 적용하여 규칙의 정확성에 영향을 주지 못하는 조건은 제거하여 규칙을 간단하게 표현하게 된다. 또한 규칙의 우선순위 선정을 위하여 규칙에 정상적인 패턴에 오판정, 즉 긍정적 잘못(False Positive error)을 적게 가지는 규칙을 포함하는 하위 클래가 첫 번째 클래스가 되도록 우선순위를 산정한다. 이렇게 규칙산출자로부터 생성된 각 클래스의 결과를 구하고, 하나의 새로운 변수를 형성하게 된다.

예를 들어, 생성된 규칙이 IF 월평균 체납액 <8380 AND 연령<25 Then 0:100%(우수고객)이라면 규칙의 정확성에 영향을 주지 못하는 조건(연령<25)을 제거하고 규칙을 IF 월평균 체납액 <8380 Then 0:100%로 형성하게 된다. 이러한 규칙으로부터 월평균 체납액이 8380보다 작은 사람들을 Class 0(우수고객)으로 각 규칙들을 분류하여 하나의 새로운 변수에 입력한다. 어떤 규칙도 포함하지 않는 경우, 가장 높은 클래스 빈도를 가지는 클래스를 디폴트 클래스로 선택한다. 이러한 과정으로 규칙을 통해 생성된 결과를 하나



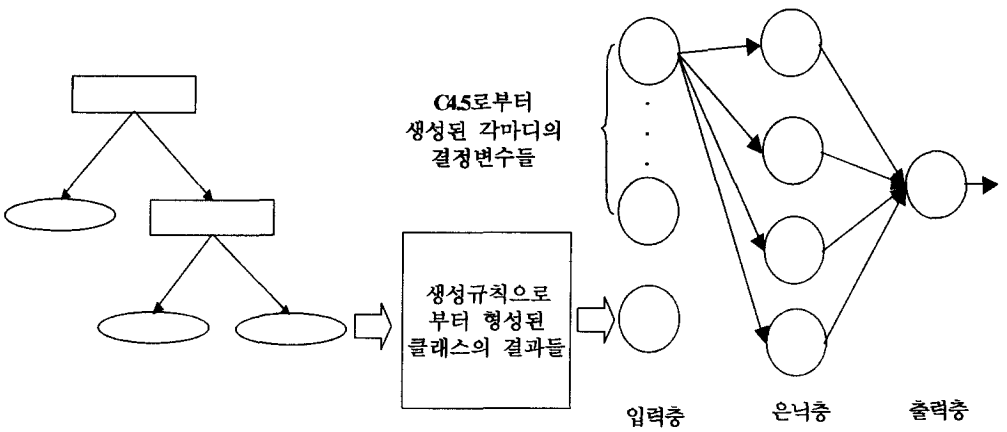
<그림 2> 생성규칙을 이용한 입력변수구축

의 새로운 입력변수에 입력하게 된다.

방법1의 경우, 목표변수의 클래스가 4개라 하면 새로운 입력변수에 들어가는 데이터의 경우 4개의 구분된 클래스가 들어가지만, 방법2의 경우, 부모노드로부터 끝노드까지 형성된 각각의 클래스를 하나의 독립된 패턴 쌍으로 인식하기 때문에, 새로운 입력변수에 들어가는 데이터의 경우 각 가지에서 분리된 규칙의 수와 동일한 클래스의 개수가 재설정되어 들어간다. 예를 들어 규칙이 20개라면 들어가는 입력변수 내의 클래스도 20개가 된다. 방법3은 위의 두 방법을 모두 적용하는 방법으로 규칙에 의해 생성된 결과와 부모노드에서부터 자식노드까지 하나의 패턴으로 묶인 범주들의 쌍을 각각 하나의 새로운 입력변수로 입력하는 과정이다.

이렇게 형성된 결정변수와 생성규칙의 결과들로 이루어진 새로운 입력변수를 신경망의 입력변수로 형성하게 된다. 다음 그림 3은 결합된 C4.5와 신경망 모델의 형성과정을 나타내고 있다. 여기서 C4.5로부터 생성된 각 마디의 결정변수들이란 C4.5의 분리기준이 각 노드에 적용되었던 최고의 이익비율에 의해 선택된 변수들을 의미한다. 규칙 산출자로부터 형성된 새로운 입력변수를 구축하기 위하여 세 가지 방법으로 개선된 알고리즘을 적용한다.

결합모델을 형성하는데 사용되는 신경망 알고리즘은 현재까지 유용하게 활용되는 오류 역전파(Back Propagation)알고리즘을 적용한다. 오류 알고리즘은 학습용 자료가 주어지면 임의로 주어진 연결강도를 이용하여 결과 값을 계산한다. 그리고, 계산된 결과와 실제 값의 차이인 오차를 계산하여 오차신호를 산출하고 이를 통하여 은닉층과 출력층으로 역전파시켜 연결 강도를 조정한다. 또한, 출력층 오차신호를 은닉층에 역전파하



<그림 3> 결합된 C4.5 와 신경망 모델의 형성과정

여 입력층과 은닉층사이의 연결강도를 변경하는 학습방법이다.

<그림 3>의 신경망 모델은 입력층, 은닉층, 출력층으로 나누어지는 다층(Multilayer Perceptron)신경망이다. 출력노드는 하나로 이루어져 있다. 목적함수를 최적화하기위하여 시행착오 방법을 사용하여 은닉층의 수와 노드를 결정하고, 목적함수가 최소일 때, 각 파라미터의 값을 선정한다.

결합모델을 구성하는데 사용된 세 가지 방법 모두 기존의 기법보다 좋은 예측의 정확도를 가져왔다. 특히, 방법1인 규칙 산출자에 의해 생성된 결과가 설명변수로 사용되어 각 클래스의 분류로 이루어진 하나의 새로운 입력변수로 구성하는 방법이 방법2, 방법3보다 좋은 예측치를 가져왔다. 이 논문에서는 방법1에 의해 구성된 결합모델에 대해서 실험 비교하였다. 방법비교 결과는 <표 1>에 제시하였다.

다음 절에는 방법1을 적용한 결합모델을 검증받기 위하여 이동통신 고객데이터를 적용하여 기존의 신경망, 회귀모델, 의사결정나무, 결합모델 등을 비교 실험하였다.

Step1: Construct C4.5 Tree
Step2: Create Production Rules
Use pessimistic error rate
Step3: Ranking Classes and Choosing a default
Compute to minimize false positive errors
Step4: Insert the result of rules to input node
Step5: Insert attributes selected from precedence gain ratio to new nodes
Step6: Compute neural network.

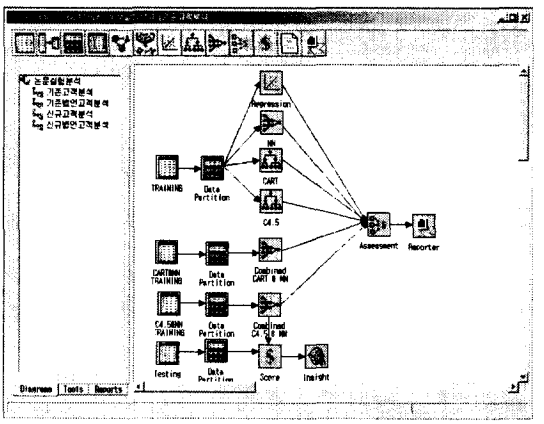
<그림 4> 결합된 C4.5와 신경망알고리즘의 단계

4. 고객 분류 시스템 적용

일반적으로 이동통신 고객데이터와 같이 다양하고 많은 양의 데이터에서 특정한 패턴과 속성을 찾아내고 연체가능성과 이탈 가능성을 정확하게 예측하기란 결코 쉬운 일이 아니다. 예측의 정확도를 향상시키기 위한 해결방안으로 체계적

인 기법을 통하여 고객성향 및 행동을 정확히 분석해 낼 수 있다면 고객에 대한 개별적이고 차별화된 서비스가 가능하게 될 것이다.

따라서 기업 마케팅 측면에서 고객을 이해하고 세분화하여 매출을 증대 시킬 수 있는 계기가 될 것이다. 제안한 결합모델과 타 기법사이의 실험 평가를 위하여 SAS Enterprise Miner 4.1을 이용하여 모델 사이에 예측의 정확도를 비교 분석하였다.



<그림 5> SAS Enterprise Miner 4.1을 이용한 기존고객 분석과정

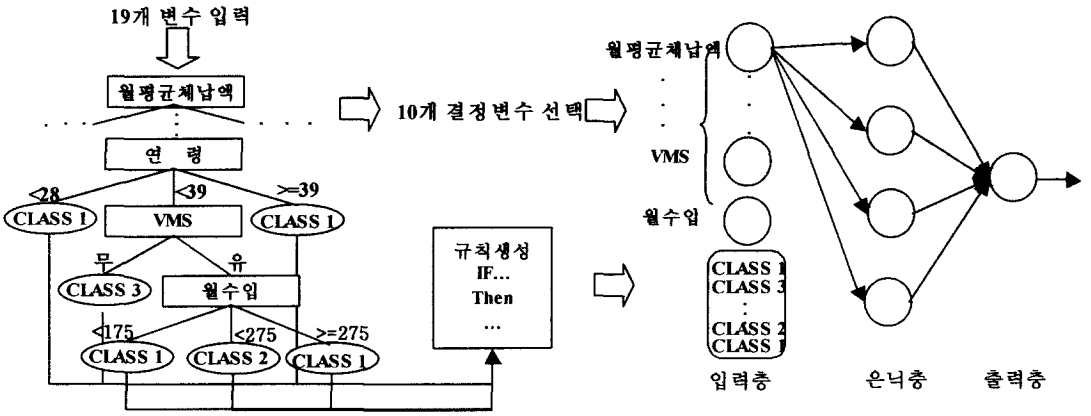
이 논문의 실험은 2001년 10월부터 2002년 3월까지 6개월 동안의 이동통신A사의 실제 고객데이터를 중심으로 적용하였다. 실험에 사용된 36000개 데이터 중 분석용 데이터(Training data) 18000개와 평가용 데이터(Validation data) 11000개, 검증용 데이터 (Testing data) 7000개로 분리하여 실행하였다.

실험은 기존고객에 대한 분석과 신규고객에 대한 분석으로 나누었다. 이 실험의 변수는 전체 33개 변수로 구성되어 있으며 이 중 기존고객을 분류하기 위하여 사용되는 변수로는 연속형 변수 6개(연령, 월수입, 월평균사용액, 월평균 체납액, 체납율, 수납율)로 구성되었으며, 범주형 변수 13개 (고객 분류, 모바일 카드 합, 부가 서비스 합, 통보횟수, 직업, 지급인, 영업장, 주소구분, 이체방법, 확인결과, 성별, 반송물 구분, VMS)등 19개 변수로 구성하였다.

신규 고객에 대한 분류로는 연속형 변수 4개(연령, 월수입, 체납율, 수납율)로 구성되었으며, 범주형 변수 10개(고객 분류, 모바일 카드 합, 부가 서비스 합, 직업, 지급인, 영업장, 주소구분, 이체방법, 성별, VMS)등 14개 변수로 구성하였다. 훈련 집합의 목표변수로써 고객의 분류는 고객의 신용등급(0=좋음(우수고객), 1=보통(보통고객), 2=좋지 않음(불량고객), 3= 매우 좋지 않음(위험고객)을 나타내며 변수는 순서형으로 표현하였다.

결측치를 대체 하는 방법으로 C4.5의 방법을 사용하여 보전하였다. 데이터의 형태가 신경망을 실행하는데 있어 많은 영향을 주므로 모델을 왜곡시키는 이상치 제거와 입력변수를 변환시켜 [0,1]사이의 값을 가지도록 하였다. 훈련집합은 파라미터를 평가하는데 사용된다. 검증용 데이터는 모델의 생성능력을 평가하는데 이용된다.

기존 고객의 실험을 위하여 C4.5 모델을 생성하였다. 한 잎에 위치하는 최소한의 관측개수는 1~100까지 변화시켜 측정하였으며, 엔트로피 지수의 분리기준에 의해 부모마디가 자식마디로 분리되기 위해 요구되는 관측개체의 수는 50~185까지 각 개체의 수를 바꾸어 측정하였다. 자식마디를 형성하기 위하여 고려되는 최대 분리개수는 다지분리를 원칙으로 2~10까지 변화 적용하였다. 또한 뿌리마디로부터 끝마디의 깊이를6~100까지 변화시켜 측정하였다. 이렇게 다양한 파라미터 값들의 변화에 따른 나무 모델의 평가 측도



<그림 6> 기존고객에 대한 규칙생성과 결정변수 선택과정

로써는 평균제곱오차(Average Square Error)가 최소가 되는 나무 구조를 선택하였다.

모델의 비교를 위한 CART의 경우도 위와 동일한 조건하에 분리기준으로 지니지수를 적용하였고, 부모마디로부터 형성되는 자식마디의 개수는 이지분리를 적용하였다. C4.5의 최적의 나무 모델을 형성한 후, 이때의 엔트로피 지수에 의해 선택된 변수를 산출한다. 기존고객에 적용된 19개의 변수 중에 엔트로피지수에 의해 선택된 각 분리 기준이 된 변수는 연속형 변수(연령, 월수입, 월평균 사용액, 월평균 체납액)등 이상 4가지와 범주형 변수(고객 분류, 부가 서비스 합, 직업, 확인결과, 성별, VMS)등 이상 6가지로 산출된 전체 10가지 변수가 선택된다. <그림 6>은 이 과정을 나타낸다. 조합 모델을 방법1에 적용하였으며, 방법에 대한 정확도 비교는 <표 1>과 같다.

<표 1> 기존고객과 신규고객의 방법 비교

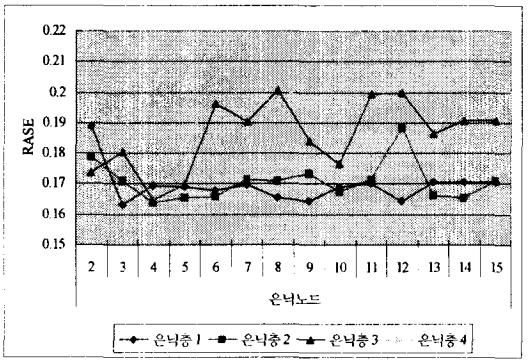
	기존고객		신규고객	
	Overall Predicted Accuracy		Overall Predicted Accuracy	
	Training Set	Testing Set	Training Set	Testing Set
방법1	92.543%	92.272%	74.771%	74.756%
방법2	90.921%	90.845%	73.231%	73.124%
방법3	90.257%	90.143%	72.829%	72.789%

신규 고객의 경우 전체 14가지 변수 중에 연속형 변수(연령, 월수입)등 2가지 변수와 범주형 변수(지급인, 직업, 영업장, 성별, 체납율, 고객 분류)등 6가지 변수로 구성된 전체8가지 변수가 선택된다.

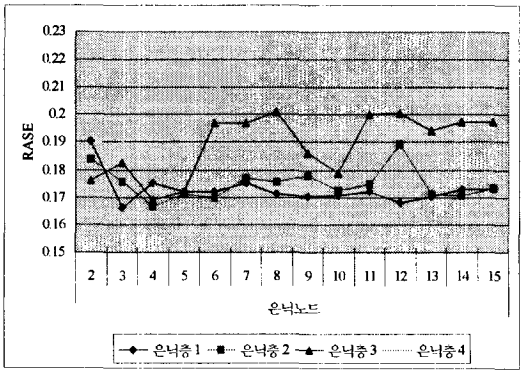
위로부터 선택된 10개의 변수와 C4.5의 규칙에 의하여 생성된 결과를 적용한 1개의 새로운 입력변수를 신경망에 적용하게 되는 것이다. 같은 방법으로 신규 고객에게도 적용하였다. 적용된 신경망의 은닉층의 수는 1~4개 층으로 실험하였으며, 은닉노드의 수는 2~15까지 변경 적용하여 분석하였다. 그림은 훈련집합과 평가집합의 은닉층 1개일 때 최소은닉노드를 기준으로 RASE를 감소시키는 방향으로 은닉층을 추가 하였다.

초기 연결강도는 Uniform Distribution에 따른 난수 발생으로 설정하였다. 이는 연결강도를 얼

마나 빨리 변화 시키느냐를 결정하기 위하여 학습율을 크게 하고 신경망이 학습되어 지면서 점점 작게 하는 것이 좋기 때문이다. 활성화 함수(Activation Function), 결합함수(Combination Function), 절편항(bias), 에러함수(Error Function), 학습율(Learning rate)등 파라미터(Parameter)의 결정을 위하여 평가용 데이터의 오차함수 값이 최소가 되는 반복에서 추정치(Estimate)를 선택하였다. 반복횟수를 100번 적용하여 실험하였다.



<그림 7> 은닉노드와 은닉층의 변화
(기존고객의 훈련집합)



<그림 8> 은닉노드와 은닉층의 변화
(기존고객의 평가집합)

여러 신경망 모형을 검토한 결과 적합한 신경망 모형은 은닉층 1개와 은닉층의 노드 수는 3개이며 다른 신경망 모형보다 작은 ASE와 MSE 그리고 오분류율을 가졌다.

결합된 모델의 평가를 위하여 기존의 다양한 예측 기법들인 회귀모형, 의사결정나무 분석(CART, C4.5), 신경망 등을 비교하였으며, 회귀분석의 경우, 로지스틱(Logistic) 회귀모형을 적용하였다. 또한, 표2의 CART & NN은 CART에 의해 형성된 나무의 분리기준을 신경망의 새로운 입력변수로 사용하는 Kao와 Chiu가 제안한 기존 방법을 말하며, 이 CART & NN의 결합모델과 본 논문의 결합 모델을 비교하였다.

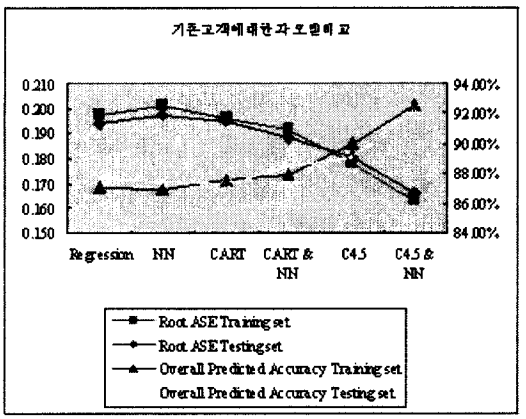
<표 2> 기존고객에 대한 각 모델의 비교

	기 존 고 객			
	Root ASE		Overall Predicted Accuracy	
	Training Set	Testing Set	Training Set	Testing Set
Regression	0.1975	0.1940	87.028%	87.682%
NN	0.2009	0.1977	86.930%	87.428%
CART	0.1960	0.1949	87.476%	87.304%
CART & NN	0.1915	0.1882	87.854%	88.444%
C4.5	0.1775	0.1810	89.998%	89.141%
C4.5 & NN	0.1630	0.1664	92.543%	92.272%

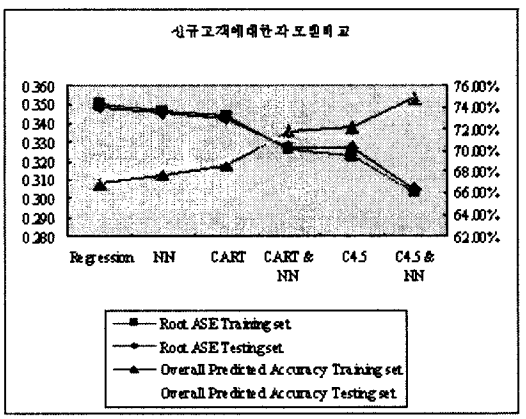
<표 3> 신규고객에 대한 각 모델의 비교

	신 규 고 객			
	Root ASE		Overall Predicted Accuracy	
	Training Set	Testing Set	Training Set	Testing Set
Regression	0.3505	0.3482	66.820%	67.234%
NN	0.3469	0.3456	67.674%	68.200%
CART	0.3438	0.3422	68.613%	69.003%
CART & NN	0.3258	0.3274	71.786%	71.614%
C4.5	0.3225	0.3268	72.058%	71.083%
C4.5 & NN	0.3035	0.3046	74.771%	74.756%

이를 위하여 다른 기법들은 앞에서 기술한 실험의 방법으로 적용하였다. 이러한 다양한 기법들을 적용하여 실험을 하였을 때, 본 논문의 결합 모델이 예측의 정확도가 훨씬 높았으며, ASE와 MSE가 작게 나타났다. 이들 각 모델들에 대한 결과는 위의 <표 2>와 <표 3>에 요약하였다.



<그림 9> 기존고객에 대한 각 모델비교



<그림 10> 신규고객에 대한 각 모델비교

본 논문의 결합모델의 통계적 유의함을 증명하기 위하여 주어진 데이터 집합에 훈련, 검증, 평가집합을 추출 시 시드(Seed)를 랜덤하게 적용하여 5회 반복 실험하였다. 이 때의 정확도의 변화는 다음 <표 4>, <표 5>와 같다.

<표 4> 각 모델의 반복 실행과 정확도
(기존고객의 훈련집합)

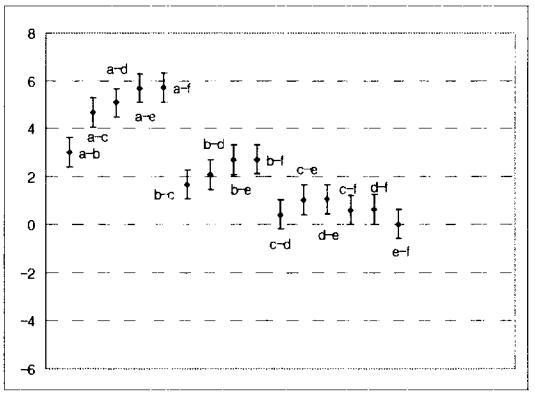
실행 횟수	a	b	c	d	e	f
	C4.5 & NN	C4.5	CART & NN	CART	NN	Regres-sion
1	92.543	89.998	87.854	87.476	86.93	87.028
2	92.124	88.487	87.355	87.321	86.754	87.011
3	91.894	89.485	87.642	87.128	86.543	86.998
4	91.763	89.629	87.807	87.235	86.542	85.432
5	92.345	89.054	87.699	87.128	86.487	86.654

<표 5> 각 모델의 반복 실행과 정확도
(기존고객의 평가집합)

실행 횟수	a	b	c	d	e	f
	C4.5 & NN	C4.5	CART & NN	CART	NN	Regres-sion
1	92.272	89.141	88.444	87.304	87.428	87.682
2	92.018	88.354	88.451	87.213	87.421	87.654
3	91.745	89.301	87.503	87.108	87.328	87.432
4	91.654	89.018	88.213	87.159	87.265	86.172
5	92.128	88.71	87.512	87.023	87.245	86.754

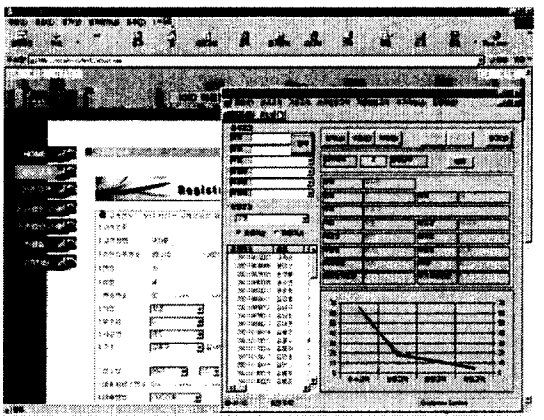
훈련 집합의 모델들에 대하여 F- 검정을 실행하였다. 유의 수준 α=0.05를 적용하였을 때, F비=111.719, F 기각치=2.6206으로 F비가 F 기각치보다 월등히 크므로, 실험에 사용된 모델들은 유의차가 있다고 말할 수 있다. 또한 각 모델 간에 유의함을 보이기 위하여 각 모델간의 모평균 차에 관한 신뢰구간을 구하여 유의함을 검증하였다. 이때 사용된 t 값은 2.064 이며, 이때의 LSD=0.60930643이다. 아래 그림은 각 모델 간의 차를 이용한 신뢰구간을 나타내고 있다. 결합모델과

각 모델의 표본평균의 차가 주어진 LSD보다 크므로 결합모델은 각 모델과 유의함을 알 수 있다. 그러나 훈련집합과 평가집합의 모평균 차에서 공통되게 신경망과 회구 분석 사이에는 유의차가 없었다. 그림 11은 이러한 모델간의 모평균 차에 관한 신뢰 구간을 나타낸 것이다.



<그림 11> 각 모델간의 모평균 차에 관한 신뢰구간(기존고객의 훈련 집합)

고객 분류를 위하여 결합모델을 고객 분류시스템에 적용하였다. 그림 8은 결합알고리즘을 구현된 분류시스템에 적용하여 고객을 세분화하고, 입력된 고객데이터가 최적으로 분류하고 있음을 보여주고 있다. 분류시스템은 Active Server Page(ASP)를 통한 고객등록 웹 페이지에 입력된 고객데이터를 MS-SQL Server 2000에 의해 형성된 고객데이터 베이스에 입력되며, 이렇게 입력된 고객데이터는 Visual Basic 6.0으로 구현된 분류시스템에 의해 세분화된 고객 분류가 이루어진다. 이 분류시스템은 입력된 신규고객에 대하여 분류를 나타내며, 시스템 내 그래프는 전체 입력된 고객 분포의 백분율과 예측된 신규고객의 분포를 백분율로 비교하여 표현한 것이다.



<그림 12> 웹을 통한 고객입력과 분류시스템을 이용한 고객 분류

5. 결론

의사결정 나무와 신경망의 결합모델이 특별한 데이터의 경우를 제외하고는 일반적으로 단일 기법보다 우수하다는 것은 많은 연구를 통하여 보여 지고 있다.

본 논문은 C4.5와 신경망의 결합 모델이 회귀, 신경망, 의사결정나무, CART와 신경망의 결합 모델 등 기존모델 보다 예측의 정확도가 우수함을 설명하였다. C4.5의 생성규칙에 의해 형성된 예측 결과와 결정변수를 통하여 신경망에 결합함으로써 분류(Classification)의 능력을 개선시켰다.

이러한 결합 모델을 통하여 예측의 정확도 향상, 체계적인 변수선정 및 모델의 해석에 부분적 이해를 도울 수 있다. 이 결합 모델의 검증을 위하여 이동통신 고객 데이터가 적용되었으며, 기존고객과 신규고객으로 분류하여 실험에 사용되었다.

실험결과 결합모델이 기존의 각 모델보다 좋

은 예측결과를 가져 왔다. 고객 데이터의 분석으로 연체자 및 신용불량자에 가장 큰 영향을 미치는 변수는 월평균 사용액과 월평균 체납액임을 알 수 있었으며, 이러한 데이터의 패턴을 분석함으로써 우수고객의 지속적인 관리와 연체고객을 통한 이탈가능고객을 분류 예측하여 기업의 손실을 줄일 수 있을 것이다. 나아가 고객의 신용평가에 도움이 될 것이다. 끝으로 이동통신 고객데이터를 결합모델에 적용하여 고객 분류 모형을 구현하였다.

참고문헌

최국렬 외 8인 공저, *데이터마이닝 이론과 실습*, 청구문화사, 2001

최종후 외 5인 공저, *SAS Enterprise Miner 4.0을 이용한 데이터마이닝 방법론 및 활용*, 3판, 자유아카데미, 2001

Andrienko, G.L. and N.V. Andrienko, "Data mining with C4.5 and interactive cartographic visualization", *User Interfaces to Data Intensive Systems*(1999), 162-165

Chung, Y. Y, M. T. Wong and N.W.Bergmann, "High speed neural network based classifier for real-time application", *Signal Processing Proceedings, 1998. ICSP '98. 1998 Fourth International Conference on*, Vol.,1(1998), 506-509

Endou, T.and Q. Zhao, "Generation of Comprehensible Decision Trees Through Evolution of Training Data", *2002. CEC '02. Proceedings of the 2002 Congress on Evolutionary Computation*, Vol.2 (2002), 1221-1225

Ganti, V., J. Gehrke and R Ramakrishnan, "Mining very large databases", *IEEE - Computer*, Vol., 32Issue: 8(1999), 38-45

Hasan, H. and P. Hyland, "Using OLAP and multidimensional data for decision making", *IT Professional* , Vol., 3Issue: 5, (2001), 44-50

Ho,T.K., "C4.5 Decision Forests",*Proceedings of the International Conference in Pattern Recognition*, Vol.,1 (1998). 545-549

Kijsirikul, B. and K. Chongkasemwongse, "Decision Tree Pruning Using Backpropagation Neural Networks", *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, Vol.,3(2001), 1876-1880

Kao, L. J. and C.C. Chiu, " Mining the customer credit by using the neural network model with classification and regression tree approach", *IFSA World Congress and 20th NAFIPS International Conference*, Vol.,2 (2001), 923-928

Lu,H., R.Setiono and H. Liu, "Effective data mining using neural networks", *IEEE Transactions on Knowledge and Data Engineering* , Vol., 8 Issue: 6 (1996), 957-961

Maher, J. J. and T.K. Sen, "Predicting Bond Ratings Using Neural Networks: a Comparison with Logistic Regression", *Intelligent Systems in Accounting, Finance and Management*, Vol.,6(1997), 59-72

Mitra, S., S.K. Pal and P. Mitra, "Data mining in soft computing framework: a survey", *IEEE Transactions on Neural Networks*, Vol.,13 Issue: 1(2002), 3-14

Paruelo, J.M. and F.Tomasel, "Prediction of Functional Characteristics of ecosystems: a comparison of artificial neural networks and regression models", *Ecol. Modell*, Vol.,98 (1997), 173-186

Quinlan J.R.,*C4.5: Programs for Machine Learning*. Morgan Kaufmann,1993

Quinlan, J.R., "Decision Trees and multi-valued attributes", *Machine Intelligence 11*, (1988), 305-318

Ruggieri, S. "Efficient C4.5", *IEEE Transactions on Knowledge and Data Engineering*, Vol.,14 Issue: 2(2002), 438-444

Schmitz, G. P. J., C. Aldrich and F. S. Gouws, "ANN-DT: An Algorithm for Extraction of Decision Trees from Artificial Neural Networks", *IEE Transactions on Neural Networks*, Vol.,10 No. 6(1999), 1392-1401

Sethi, I.K., "Entropy nets: from decision trees to neural networks", *Proceedings of the IEEE*, Vol., 78Issue: 10(1990), 1605-1613

Zorman, M. and P. Kokol, "Hybrid NN-DT Cascade Method For Generating Decision Trees From Backpropagation Neural Networks", *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, Vol.,4(2002)

Abstract

A Study on the Combined Decision Tree(C4.5) and Neural Network Algorithm for Classification of Mobile Telecommunication Customer

Keukno Lee* · Hongchul Lee*

This paper presents the new methodology of analyzing and classifying patterns of customers in mobile telecommunication market to enhance the performance of predicting the credit information based on the decision tree and neural network. With the application of variance selection process from decision tree, the systemic process of defining input vector's value and the rule generation were developed. In point of customer management, this research analyzes current customers and produces the patterns of them so that the company can maintain good customer relationship and makes special management on the customer who has high potential of getting out of contract in advance. The real implementation of proposed method shows that the predicted accuracy is higher than existing methods such as decision tree(CART, C4.5), regression, neural network and combined model(CART and NN).

Key words : Combined C4.5 and Neural Network, Classification, Prediction, Decision Tree, Data Mining

* Department of Industrial Systems and Information Engineering, Korea University