Support Vector Machine 기법을 이용한 고객의 구매의도 예측

Forecasting of Customer's Purchasing Intention Using Support Vector Machine

김 진 화 (Jin Hwa Kim) 서강대학교 경영대학 경영학과 부교수, 교신저자
남 기 찬 (Ki Chan Nam) 서강대학교 경영대학 경영학과 교수
이 상 종 (Sang Jong Lee) 서강대학교 경영전문대학원 박사과정

요 약

기업 경쟁력 강화의 중요한 이슈인 대량 개별화(mass-customization)의 실행을 위하여 통합 고객관계관리 프로세스로서의 CRM(customer relationship management)에 대한 관심과 활용에 대한 필요성은 점점 더 높아지고 있다. 특히, 기존 고객들의 구매 정보를 기반으로 고객의 구매 패턴을 파악하고 의도를 예측하는 것은 오늘날 실질적인 판매 전략을 수립하는 마케팅 분야에서 상당히 큰 비중을 차지하고 있다.

고객의 구매의도 예측에는 대량의 데이터로부터 과거에 인지하지 못했던 의미 있고, 근거 있는 정보를 추출하는 데이터마이닝(datamining)이 주로 사용되고 있다. 기존의 구매의도 예측에 사용된 데이터마이닝 기법들은 주로 신경망(neural networks)과 로지스틱 회귀분석(logistic regression analysis)이었는데, 예측 정확성 및 모형 구축의 어려움으로 인한 다양한 문제점들이 제기되고 있는 실정이다. 따라서, 본 논문에서는 기존의 기법들이 가지고 있는 단점들을 개선하기 위하여 신경망과 로지스틱 회귀분석 외에 연관규칙(association rule), 연관성 매트릭스(association matrix), 의사결정 나무(decision tree), 베이지안 망(bayesian network), SVM(support vector machine) 기법들을 추가로 제안하였다.

본 연구의 목적은 고객의 특정 상품에 대한 구매의도 예측을 위하여 새로운 알고리즘을 제시하기보다는 기존의 다양한 데이터마이닝 기법들을 적용시켜 봄으로써, 가장 우수한 예측 성과를 나타내는 기법을 발견하는 것이다. 연구에 사용된 자료는 기존의 연구에서는 적용되지 않았던 편의점의 영수증 데이터이다. 예측 목표상품은 카테고리화 된 '우유'와 '냉동식품'이며, 제안된 기법들의 신뢰성을 위하여 전체 데이터를 10개의 training과 test 셋으로 중복되지 않게 구분함과 동시에 10번의 교차 검증(cross validation)을 실시하였다.

실험 결과 SVM이 영수증 데이터를 이용한 고객의 특정 상품에 대한 구매의도 예측에서 가장 우수한 성과를 나타내는 것을 확인하였다.

키워드 : 고객관계관리, 추천시스템, 영수증데이터, 데이터마이닝, 구매의도예측

Ⅰ. 서 론

1.1 연구 동기

최근 디지털 정보기술의 급속한 발전은 다양한 시장공간을 창출시키고 있으며, 특히 인터넷 매체의 빠른 확산은 새로운 경제현상을 만들어 낼 뿐만 아니라 기업의 경쟁전략을 변화시키고 있다. 이러한 시장 환경의 변화 속에서 과거와 달리 제품이나 서비스에 대한 고객들의 욕구 또한 더욱 다양화되어 점차적으로 기업에 대한 자신들의 영향력을 증대시키고 있다.

따라서, 기업 경쟁력 강화의 중요한 이슈가 되어버린 대량 개별화(mass-customization)의 실행을 위하여 정보기술을 기반으로 한 고객의 다양한 정보를 획득함과 동시에 그것을 통하여 고객과의 밀접한 관계를 유지함으로써 기업의 수익성을 증대시키는, 통합 고객관계 관리 프로세스로서의 CRM(customer relationship management)에 대한 관심과 활용에 대한 필요성은 점점 더 높아지고 있다. 특히 대량의 데이터로부터 과거에 인지하지 못했던 의미 있고, 근거 있는 정보를 추출하는 데이터마이닝(datamining)의 등장은 고객관련 데이터베이스로부터 보다 정확한 정보를 획득하여 전략적으로 활용해야만 하는 CRM의 요구와 부합하여 그 효과를 더욱 가시화 할 수 있는 기반을 제공하고 있다(장남식, 2000).

CRM의 여러 분야 가운데에서도 제품을 구매한 기존 고객의 정보를 기반으로 그 고객에게 맞는 새로운 제품이나 서비스를 제안하기 위하여 구매 패턴을 파악하고 의도를 예측하는 것은 오늘날 실질적인 판매 전략을 수립하는 마케팅 분야에서 상당히 큰 비중을 차지하고 있다. 일반적으로 고객의 구매 의도를 파악하고 예측하는 데는 연관규칙(association rule), 의사결정나무(decision tree), 신경망(neural networks), 로지스틱 회귀 분석(logistic regression analysis) 등의 데이터마이닝 기법들이 주로 사용되어왔다.

연관규칙의 경우 미시적 뿐만 아니라 거시적 관점의 데이터 분석이 가능하고, 의사결정나무는 적용결과에 대하여 명확하고 쉽게 이해 할 수 있는 장점이 있다. 그리고 신경망은 비선형적인 데이터의 패턴을 잘 식별하고 데이터의 노이즈, 즉 실측 데이터를 잘 처리할 수 있으며, 로지스틱 회귀분석은 통계적 기법에 근간한 모형으로서 각 변수의 영향력을 정확하게 설명할 수 있다는 장점이 있다. 하지만, 연관규칙은 생성되는 많은 양의 규칙 대부분이 실제 활용가치가 적고, 의사결정나무의 경우 새로운 자료의 예측에는 불안정하며, 신경망의 경우에는 모형 구축에 많은 시간이 소요될 뿐만 아니라 모형에 대한 설명력 또한 매우 부족하다는 단점이 있다. 그리고, 로지스틱 회귀분석의 경우에는 예측 성과가 높지 않다는 단점이 있다(Alex Berson 등, 1999; Ganti 등, 1999; 안현철 등, 2005).

이처럼 기존의 데이터마이닝 기법들이 가지고 있는 한계점들을 최소화하기 위하여, 본 연구에서는 최근 화두가 되고 있는 support vector machine(SVM)을 이용하여 고객의 구매 의도를 예측하여 보고, 그 성능을 기존의 기법들과 비교 및 분석해본다. SVM이 관심의 대상이 되는 이유는 명백한 이론적 근거에 기반함으로 결과 해석이 용이하고, 실제 응용에 있어서 높은 성과를 내며, 적은 학습자료만으로 신속하게 분별학습을 수행할 수 있기 때문이다. 따라서, 본 연구를 통하여 고객의 상품 구매의도에 대한 가장 높은 예측력을 보이는 기법을 찾아냄으로써 보다 효율적이고 수익성 있는 CRM 전략수립에 가치 있는 정보를 제공할 수 있을 것으로 기대한다.

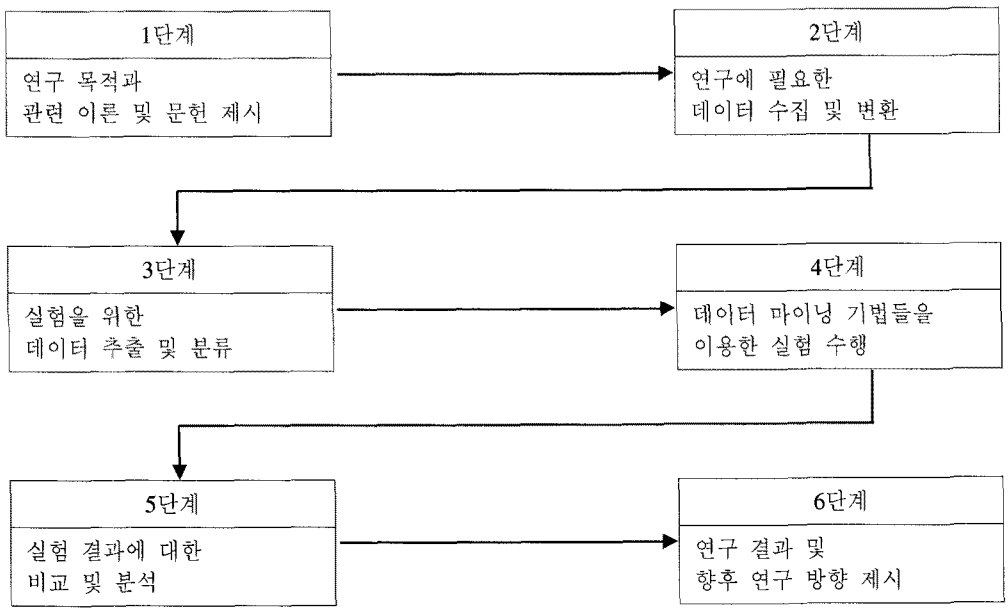
1.2 연구 범위와 방법

본 연구에서는 고객의 상품 구매의도 예측에 대한 새로운 형태의 알고리즘을 제시하기보다는 기존의 다양한 데이터마이닝 기법들을 동일한 데이터로 적용시켜봄으로써 어떠한 기법이 가장

우수한 성능을 보이는가에 대한 실험을 한 후 결과를 비교 및 분석하여 본다.

회원들의 나이, 성별, 직업, 교육 수준 등과 같은 여러 가지 정보들이 포함되어 있는 자료를 기반으로 한 데이터마이닝 기법의 적용에 대한 연구는 많지만, 실질적으로 편의점이나 일반 슈퍼마켓과 같은 소매점에서 단순히 영수증을 이용한 고객의 상품 구매 패턴이나 구매 의도를 예측한 연구는 거의 실행되지 않았다. 따라서, 본 연구에서는 영수증에 표시된 구매 품목들을 데이터로 이용하였으며, 품목 수가 방대한 관계로 카테고리화시켜 데이터의 크기를 축소하였다. 그리고, SVM, 연관규칙, 의사결정나무, 신경망, 베이지안 망, 로지스틱 회귀분석과 같은 데이터마이닝 기법 외에도 연관성 매트릭스(association matrix)를 작성하여 품목간의 연관 정도를 측정하였으며, 총 7가지 기법들의 예측 성능을 측정하여 비교하였다.

연구 절차는 <그림 1>과 같이 6단계로 진행된다. 1단계에서는 연구의 목적 및 관련 이론을 제시한다. 이 단계에서는 연구에 사용되는 다양한 이론들에 대한 개념뿐만 아니라 과거에 진행되었던 연구 결과에 대한 문헌들도 제시한다. 2단계에서는 분석에 필요한 데이터를 구축하기 위하여 편의점에서 거래된 내역에 대한 자료를 수집한 후, 실험에 적용 가능한 형태로 변환을 시킨다. 3단계에서는 정제된 데이터에서 필요한 부분을 추출한 후에 실험에 필요한 훈련용/검증용으로 구분된 중복되지 않은 10개의 데이터 셋을 만든다. 이 단계에서는 데이터 셋의 중복을 방지하기 위하여 전체 데이터 각각에 일련번호를 지정하여 그것을 기준으로 구분한다. 4단계에서는 준비된 데이터를 기반으로 7가지 데이터마이닝 기법들에 대한 실험을 10번씩 반복해서 실행한다. 이 부분에서 각 기법들의 실험 모델이나 결과 화면에 대한 예와 설명, 그리고 간략한 실험 결과를 제시한다. 5단계에서는 모든 실험에 대한 결과들을 종합적으로 정리하고 기법들 간의 성능을 비교 및 분석한다. 이 단계에서는 전체적인 결과를 분석하여 기존에 제시된 이론



〈그림 1〉 연구 절차

및 연구 결과들과 비교해본다. 뿐만 아니라, 가장 우수한 성능을 보인 기법에 대한 이론적 근거 및 특징들을 설명한다. 마지막으로 6단계에서는 전반적인 연구의 개요와 실험 과정을 종합적으로 요약하고, 연구를 통해 얻은 시사점을 제시해본다. 그리고, 본 연구의 과정에서 느꼈던 여러 가지 한계점들과 이러한 부분들을 보완하기 위한 향후 연구 방향에 대해서 제시한다.

Ⅱ. 이론적 배경

본 장에서는 연구의 궁극적 목적인 CRM 및 고객의 상품 구매의도 예측에 대하여 알아보고, 데이터마이닝에 대한 정의와 관련 기법들에 대하여 간략하게 정리하고자 한다.

2.1 CRM과 데이터마이닝

과거와 달리 현대 기업들은 제품뿐만 아니라 고객들의 스타일, 원하는 서비스, 자신들에 대한 이미지 등을 알 필요가 있다. 따라서, 모든 고객들과의 관계를 관리할 필요가 있고, 각각의 관계를 가능한 한 수익성 있게 만들어야 한다. 이러한 사항들을 충분히 고려한 기업은 판매 및 마케팅 비용을 낮춤과 동시에 매출을 늘릴 수 있으며, 고객의 이탈 및 비효과적인 영업 활동에서 비롯되는 비용을 줄일 수가 있다. 이것을 가능하게 만드는 방법론이 고객관계관리, 즉 CRM(customer relationship management)이다(알렉스 버슨 등, 1999).

CRM은 지난 10년 동안 엄청난 관심의 대상이 되었으며, 이와 더불어 CRM과 관련된 다양한 기술적 소프트웨어에 대한 공급사들의 경쟁도 날로 치열해지고 있는 실정이다 (Peppers and Roggers Group(Asia), 2002). CRM 소프트웨어 애플리케이션의 주요 사용자는 고객과 상호작용하는 프로세서가 자동화되기를 바라는 데이터베이스 마케터들이다. 프로세스를 성공적으로 수행하기 위해서, 데이터베이스 마케터들은 먼저 높은 이익 잠재력을 가진 고객들을 포함하고 있는 시장군을 파악하고, 이러한 고객들의 행동에 긍정적인 영향을 미치는 캠페인을 수립하고 집행해야 하는데, 이를 위해서는 잠재고객 및 그들의 구매행태에 대한 많은 데이터가 필요하다. 이론상으로는 데이터가 많으면 많을수록 좋지만, 실제로는 방대한 데이터 자체가 종종 마케팅 담당자들을 혼돈스럽게 만들기도 한다(Alex Berson 등, 1999). 이러한 다양한 고객들의 행동을 이해하고 예측하는데 가장 중요한 역할을 하는 것이 데이터마이닝(datamining)이다(SPSS USA, 2000).

데이터마이닝은 대부분의 사업 조직이 직면하는 문제들에 대한 의사결정을 하는데 필요하며, 주로 방대한 양의 데이터베이스로부터 유용한 정보와 도움이 될 만한 지식을 추출하는데 사용된다(Achok Savasere 등, 1995; M.H.Margahny and A.A.Mitwaly., 2005). 데이터마이닝의 주된 기법으로는 의사결정나무, 신경망, 연관성 규칙, 로지스틱 회귀분석, 베이지안 망, SVM 등이 있다. 특히 CRM에서의 데이터마이닝 애플리케이션은 고객의 구매 패턴과 같은 정보를 발견하기 위해 데이터라는 거대한 산을 검색하는 프로세스를 자동화한 것으로, 데이터를 분석하여 정보를 추출하게 되면 마케팅 담당자는 그 결과를 정의된 시장군의 캠페인 관리를 목표로 한 캠페인 관리 소프트웨어로 보내게 되는 것이다(Alex Berson 등, 1999).

2.2 구매 의도 예측 및 추천 시스템

고객은 구매하고자 하는 상품들을 비교, 평가하여 자신의 지불능력에 비추어 가장 마음에 드는 대안에 대한 구매의도(purchasing intention)을 가지고 구매를 하게 된다. 따라서, 각각의 고객이 자사의 특정 상품 혹은 상품군의 구매와 관련해 관심이나 호응을 갖고 있는지, 아닌지를 분류하는 구매의도에 대한 예측은 오늘날 마케팅

분야에서 매우 중요한 이슈 중 하나로 자리매김하고 있다. 고객의 구매행동을 정확하게 파악할 수 있는 능력과 고객 데이터를 이용한 개인 상품추천시스템을 기업이 보유하고 있는 경우, 그 기업은 이를 이용해 다양한 사업기회를 발굴, 육성할 수 있다(chiu, 2002; 이학식 등, 2001; 안현철 등, 2004).

추천 시스템(recommender systems)은 고객들이 구매 또는 시험해보길 원하는 상품들에 대한 가이드를 제시하는 것으로써, 상품 설명서나 새로운 관련 기사 또는 다른 제품들 등과 같은 다양한 정보를 통하여 추천을 한다. 특히 온라인 정보와 전자상거래의 급격한 발전으로 인하여, 추천 시스템은 더욱더 중요한 도구로써 그 필요성이 급증하고 있다(Buker. R., 2000). 특히 개인화 상품 추천 시스템(personalized product recommendation systems)은 고객의 신상정보와 상점에서의 구매 행위에 대한 정보를 바탕으로 고객 취향을 반영한 상품이나 서비스를 추천한다(김종우 등, 2000).

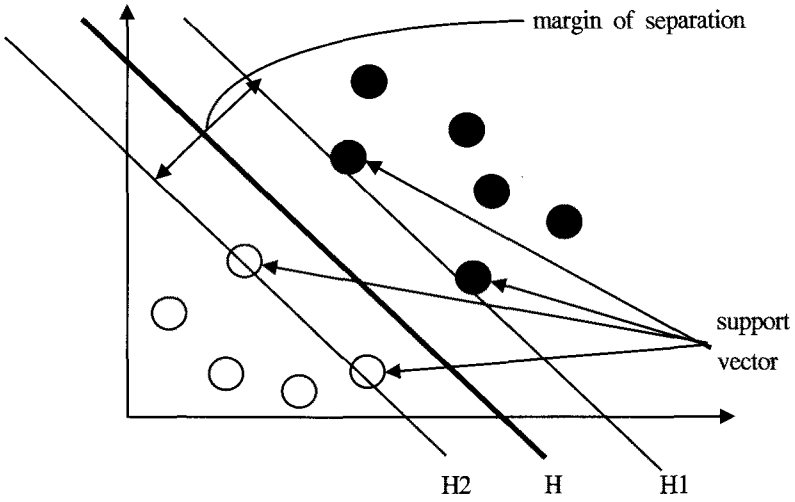
추천 시스템은 초기에는 주로 내용기반 필터링(content-based filtering) 기법을 이용하였으나, 최근에는 협업 필터링(collaborative filtering) 기법이 많이 사용되고 있다. 내용기반 필터링 기법은 정보 검색이나 정보 필터링 연구에서 자연적으로 발전하였다. 따라서, 내용기반 필터링 기법은 상품을 추천하기 위하여 상품의 속성과 고객이 필요로 하는 정보가 일치하는 정도에 따라, 그 결과를 순위화하여 보여준다. 이렇게 상품의 내용을 중심으로 분석하여 이용자에게 추천하는 기법을 내용기반 필터링 기법이라 한다. 협업 필터링 기법은 추천 시스템에서 가장 흔히 쓰이는 기법이다. 협업 필터링 기법을 이용한 추천 시스템은 고객들 사이의 유사함을 근거로 이전에 구매를 하였던 또는 구매를 하고자 하는 제품을 추천하는 것이다. 즉, 고객의 선호도를 수집하여 데이터베이스를 구축하고, 특정 고객과 유사한 취향이나 정보 요구를 갖는 고객들을 데이터베이스에서 찾아내어, 이들이 좋아하는 물건이나 정보를 이용자에게 추천하는 것이다(Francesco Ricci 등, 2003; 정영미 등, 2002).

최근 몇 년 간 협업 필터링 기법을 이용한 개인화된 상품 추천 시스템은 다양한 데이터마이닝 기법들을 응용하여 연구가 진행되었는데, 연관성 규칙이 그 대표적인 예로 볼 수 있다. 또한, 인공 신경망을 마케팅 분야에 적용한 연구도 있는데, 주로 소비자 구매 의도 예측을 통한 상품 추천에 관한 것이다(Tom Brijs 등, 1999; 한상만 등, 2000; 송수섭 등, 2001). 따라서, 본 연구에서는 대부분 연관성 규칙과 인공 신경망을 통해 이루어졌던 기존의 연구에 SVM, 의사결정나무 등과 같은 다양한 데이터마이닝 기법들을 새롭게 적용시켜 보기로 한다.

2.3 SVM(support vector machine)

SVM은 데이터로부터 분류와 규칙을 학습하기 위한 훈련 알고리즘으로써 1960년대에 Vapnik에 의해 처음으로 제안되었으며, 통계적 학습 이론을 기반으로 하고 있다. 그러나, 일반적인 통계적 학습 방법에서의 경험적 리스크 최소화(ERM: empirical risk minimize)와는 다른 구조적 리스크 최소화(SRM: structural risk minimize)를 통해 오류를 최소화 시키는 방법을 이용한다(Robert Burbidge 등, 2001).

SVM의 기본 원리는 훈련데이터들을 서로 다른 두 개의 클래스로 분류 할 때 기준이 되는 분리경계면(hyperplane)을 학습 알고리즘을 이용하여 찾는 것이다. 즉, SVM의 목적은 학습자료로 주어진 n차원의 벡터공간에서 분류 공간 간에 모든 점들 사이의 거리를 최대화하도록 만들어 하나의 평면을 구해내는 것인데, 이 선형 평면 분류 경계면을 OSH(optimal separating hyperplane)라고 하며, OSH에 가장 가까운 점들을 support vector라고 부른다. n차원의 OSH는 n차원 방향 벡터 W와 기준 벡터 b로 $WX + b = 0$를 만족하



〈그림 2〉 선형 공간에서 hyperplane 모델

는 점들의 집합으로 표현되며, 선형 공간에서의 hyperplane 모형은 <그림 2>와 같다.

<그림 2>는 흰색 원과 검은색 원을 구분짓는 hyperplane을 보여주고 있다. 여기서 H는 OHS를 나타내며 H1과 H2는 2개의 벡터 그룹의 영역을 보여주는 hyperplane이 된다. H1과 H2에서 접한 4개의 벡터가 support vector가 되게 된다. H1과 H2는 H를 기준으로 support vector를 최대 마진으로 이분한다. 여기서 2개의 그룹을 1과 -1로 보았을 때, 학습 데이터 {X i, Y i}는 <표 1>과 같이 구분된다.

〈표 1〉 학습 데이터 {X i , Y i }에 대한 구분의 예

H1: W' X i + b = 1(검정색 원 그룹을 1로 보았을 때)
H2: W' X i + b = -1(흰색 원 그룹을 -1로 보았을 때)
W: hyperplane과 직교하는 벡터. 원점에서 hyperplane과의 수직 거리는 |b|/||W||이고, ||W||는 W의 유클리드(Euclidean) 놈(norm).

H1에서 직교 거리가 |1-b|/||W||이고, H2에서 직교거리가 |-1-b|/||W||이므로 H1에서 H2까지의 마진(margin) 거리는 |2|/||W||가 된다. H1과 H2에 대해서 둘은 평행이고 사이에 학습데이터가 존재하지 않으므로, 최대 마진으로부터 hyperplane의 H1과 H2를 찾을 수 있다.

이러한 원리의 SVM은 입력 벡터를 고차원(high dimensional)의 특징공간(feature space)으로 이동시켜(mapping) 분리 경계가 매우 복잡한 문제를 선형판별함수의 사용이 가능한 단순한 문제로 변화시키기 때문에, 수학적 분석이 수월하고 조정해야 할 모수(parameter)의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다는 장점을 가지고 있다. 그리고 구조적 위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 전역적 최적해(global optima)를 구할 수 있다는 점에서 신경망보다 성능이 우수한 기계학습기법으로 주목 받고 있다(Steve R. Gunn, 1998; Marti A. Hearst, 1998; Hsuan-Tien Lin, 2005; 민재형 등, 2004).

2.4 연관성 규칙(association rule)

연관성 규칙은 가장 일반적인 데이터마이닝 기

법 중 하나로, 방대한 양의 데이터에서 규칙들을 발견하는 것으로 간단히 정의 내릴 수 있다. 이러한 연관성 규칙을 분석하기 위한 기법으로는 주로 장바구니 분석(market basket analysis)을 사용한다(M. H. Margahny and A. A. Mitwaly, 2005). 연관성 규칙을 분석하는 목적은 자료에 존재하는 연관 관계들을 찾아내고 이를 확률이나 도표 등을 사용하여 정량화하는데 있으며, 이러한 결과는 일반적으로 제품이나 서비스의 교차판매(cross selling), 매장진열(display), 첨부우편(attached mailing), 금융사기 적발(fraud detection) 등의 다양한 분야에 사용할 수 있다(류문배 등, 2004).

대용량의 데이터로부터 연관성 규칙을 추출할 때는 수량화 된 기준이 필요하고, 그 기준은 근거확률(*support*), 신뢰확률(*confidence*), 리프트(*lift*) 세 가지가 있다. 만약 관심 있는 규칙이 "X라는 상품을 구입한 사람은 Y라는 상품도 구입한다."라고 가정한다면, 연관성 규칙의 세 가지 기준은 다음과 같은 <표 2>로 설명되어진다.

〈표 2〉 *Support, Confidence, Lift*에 대한 식과 의미

1. 근거확률 (support) = $n(X \cap Y)/N$: 전체 거래 N에서 상품 X, Y에 대한 거래를 모두 포함하는 거래의 수. 즉 X와 Y 두 품목이 같이 구매될 확률.
2. 신뢰확률 (confidence) = $Pr(Y|X)$: 상품 A를 구매한 거래가 발생했을 경우 그 거래가 상품 B를 포함하는 조건부 확률.
3. 향상도 (lift) = $Pr(Y|X)/Pr(Y)$: 상품 X를 구매한 경우 그 거래가 상품 Y를 포함하는 경우와 상품 Y가 상품 X에 관계없이 단독으로 구매된 경우의 비율. 상품 X, Y가 *lift* ≈ 1이면 상호 독립적인 관계이고, *lift* > 1이면 양의 상관관계이며, *lift* < 1이면 음의 상관관계.

연관성 규칙은 이들 3가지 지표를 기준으로 추출하게 되는데, 어떤 규칙이 의미 있는가를 판단하기 위한 *support, confidence, lift* 값의 기준이 이론적으로 결정되어 있는 것은 아니며, 일반적으로 분석자의 판단과 경험에 의해 결정되는 경우가 많다(주영진, 2005; 황인수, 2004).

2.5 의사결정나무 (decision tree)

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별분석(discriminant analysis), 회귀분석(regression analysis), 신경망(neural networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(이극노 등, 2003). 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다. 의사결정나무 분석이 활용될 수 있는 응용분야는 <표 3>과 같다.

〈표 3〉 의사결정나무 분석의 응용분야

- 세분화(segmentation): 관측 개체를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하고자 하는 경우.
- 분류(classification): 여러 예측변수(predicated variable)에 근거하여 목표변수(target-variable)의 범주를 몇 개의 등급으로 분류하고자 하는 경우.
- 예측(prediction): 자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자하는 경우.
- 차원축소 및 변수선택(data reduction and variable screening): 매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라내고자 하는 경우.

하지만, 의사결정나무는 비연속성과 비안전성의 문제를 가지고 있다. 비연속성의 경우, 연속

형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에는 예측 오류가 클 가능성이 있으며, 비안정성의 경우에는 의사결정나무가 훈련용 자료에만 의존하기 때문에 새로운 자료의 예측에는 불안정할 가능성이 높다(최종후 등, 2001).

2.6 신경망(neural networks)

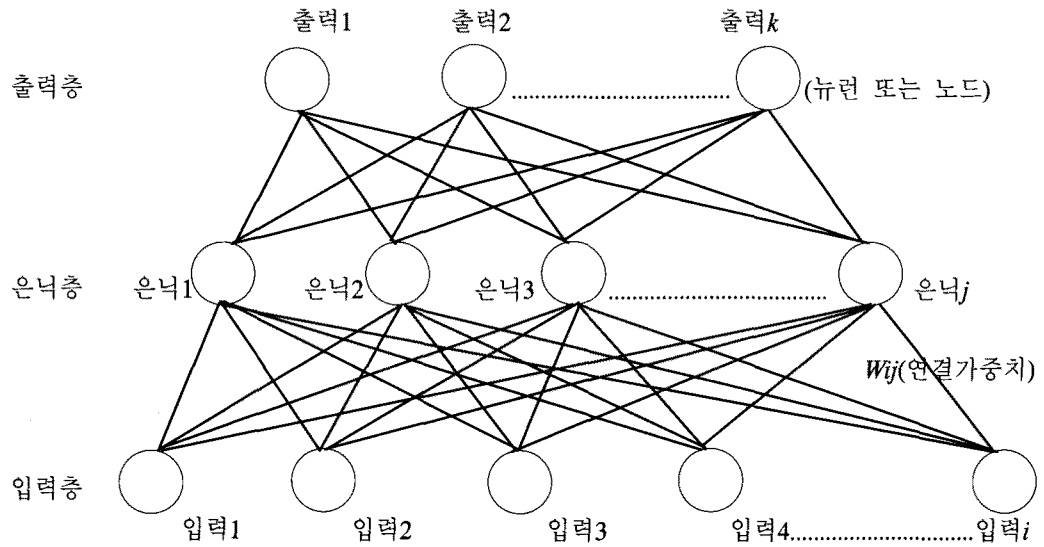
인공신경망은 생물학적 뇌의 작동 원리를 그대로 모방하는 방법으로, 데이터 안의 독특한 패턴이나 구조를 인지하는데 필요한 모델을 구축하는 도구이다(Kate A. Smith., 2002).

인공신경망은 간단한 계산능력을 가진 처리단위, 뉴런(neuron) 또는 노드(node)들이 서로 복잡하게 연결된 컴퓨터 시스템으로서 외부에서 주어진 입력에 대하여 동적인 반응을 할 수있다. 이러한 특징은 결국 인공신경망을 구성하고 있는 다수의 뉴런끼리의 상호연결성에 기인한 것이다. 뉴런은 생체내의 신경세포와 비슷한 것으로써 가중치화 된 상호연결성으로 서로 연결이 있다. 가장 일반적인 인공신경망 모형은 <그림 3>과 같은 다계층 퍼셉트론 모형으로서, 입력층(input layer)에서 은닉층(hidden layer), 은닉층에서 출력층(output layer)으로 각 뉴런이 서로 연결되어 있는 것이 특징이다(Rumelhart 등, 1986).

이러한 신경망 기법은 비선형적인 자료에서 지식이나 패턴을 추출할 수 있고, 입력-출력 맵핑(input-output mapping) 기법이라서 자료에 대한 통계적 분석 없이 결정을 수행할 수 있으며, 상대적으로 적응력(adaptability)이 뛰어나고 견고한(robust) 모델이라는 점에서 다양한 분야에서 적용되고 있다. 하지만 모델이 제시하는 결과에 대해서 왜 그런 결과가 나오는지에 대한 원인을 명쾌하게 설명할 수 없다는 점과 과도하게 학습을 진행할 경우, 전체적인 관점에서의 최적해가 아닌 지역 내 최적해가 선택될 수 있다는 과적합화(overfitting) 문제는 인공 신경망 기법의 치명적인 단점이라고 할 수 있다(Berry, Michael J. A. and Gordon Linoff, 1997; 한상만 등, 2000; 안현철 등, 2004).

2.7 베이지안 망(bayesian network)

베이지안 망은 그래프 이론(graph theory)과 확



〈그림 3〉 다계층 퍼셉트론 모형의 기본 구조(Jain, Bharat and Barin, N. nag., 1997)

률 이론(probability theory)의 결합에 기초한 확률 그래프 모델로서, 모듈성(modularity) 개념 하에 변수나 자질들간의 복잡한 관련성 및 의존관계를 망 구조상에서 보다 간결한 확률적 모듈들의 집합으로 표현한다. 즉, 노드를 이용한 그래픽을 통하여 데이터 내부의 지식을 나타내며, 변수들 간의 관계를 화살표로 표시한다. 이러한 특징에 기반하여, 베이지안 망은 실제 다양한 문제에서의 복잡성(complexity) 및 불확실성(uncertainty)을 효과적으로 다룰 수 있는 틀을 제공하는 것으로써, 의학, 기계, 문서 분석, 의사결정지원 시스템 등에서 지식을 모델링 하는데 사용된다(S. M. Lee and P. A. Abbott, 2003).

베이지안 망은 조건부 독립성(conditional independent)을 나타내는 directed acyclic graph(DAG)를 사용하여 많은 변수들간의 다양한 확률분포를 비교적 축약된 형태로 표현하기 때문에 변수들간의 상관관계를 쉽게 이해하고자 할 때 유용하게 쓰이며, 다른 기술들과 달리 노드와 화살표를 사용함으로써 결과를 더욱더 쉽게 이해할 수 있는 장점이 있다. 하지만, 데이터마이닝에 베이지안 망을 적용할 때 문제가 되는 것들 중에 하나는 일반적으로 데이터마이닝의 데이터들이 차원이 아주 크고 실제로 관심이 있는 속성들과 연관이 없는 정보들이 많기 때문에 데이터 자체의 축소가 필수적으로 실행되어야 한다는 것이다(하선영 등, 2000).

2.8 로지스틱 회귀분석(logistic regression analysis)

로지스틱 회귀분석은 자료가 두 모집단으로 나누어진 상황에서, 연구대상이 어떠한 모집단에 속한 지를 예측하는 분류 목적으로 사용되는 통계적 분석방법이다. 두 모집단의 특성에 대한 차이를 파악하여 연구 대상이 어떠한 모집단에 더 가까운지를 파악하는 목적으로 사용되는 모형을 분류모형(classification model)이라고 하는데, 로지스틱 회귀모형은 이러한 분류모형의 한 형태로써, 종속변수가 두 범주로 구성되어 있는 명목변수 일 때 적절한 통계적 기법이다.

종속변수가 0, 1만의 값을 갖는 가변수(dummy variable)인 경우에 y의 기대값을 나타내는 반응함수의 모형이 S자형 곡선을 그리는 경우가 실제로 많이 나타난다. 이 반응함수는 x가 증가함에 따라 E(y)가 1로 서서히 수렴하는 양상을 보이는데 이와 같은 함수를 로지스틱 함수(logisitic function)라 부른다. 즉, 로지스틱 회귀분석이란 단지 두 개의 값만을 가지는 종속변수(예를 들어 주택유무, 회원가입여부 등)와 독립변수들 간의 인과관계를 로지스틱 함수를 이용하여 추정하는 통계기법이다.

로지스틱 회귀분석은 그 동안 의학분야에서 사용되면서 발전되어 왔다. 예를 들어, 심장병에 대한 위험요인을 파악하기 위하여, 환자들의 신체적 특성이나 증상을 조사하고 정상인들의 특성과 비교함으로써 어떠한 특성을 가진 사람들이 심장병에 많이 노출되어 있는지를 파악하고 예측하는 것이다. 하지만, 지금은 기업의 도산을 예측하는 문제부터 10대 소녀의 임신 가능성에 대한 예측문제까지, 사회과학의 모든 분야에서 분류모형의 대명사로 유용하게 활용되고 있다(이군희, 2004).

Ⅲ. 연구 설계

3.1 자료 수집과 변수 선정

고객 구매 의도를 SVM을 포함한 다양한 데이터마이닝 알고리즘들을 이용하여 예측하여보고, 그 성과를 비교 및 분석하기 위하여 본 연구에서는 실제 데이터를 적용하여 해당 결과를 도출하였다.

본 연구에 사용된 데이터는 서울시 용산구에 위치한 G편의점의 판매 자료이다. 1990년 12월에 1호점을 개점한 G편의점은 국내 독자 개발

브랜드로 시작하여 2006년 현재까지 업계 1위를 굳건히 지켜오고 있는 국내 최고의 편의점이다.

G편의점으로부터 본 연구를 위하여 확보한 자료는 지난 2005년 9월 1일부터 12월 7일 사이에 고객들이 구매한 1,334건의 거래 내역 데이터이다. 거래 내역의 확인은 G편의점 지점 POS 관리 시스템으로부터 하였으며, 이 시스템의 필드는 '판매일자, 판매시간, POS, 담당자, 영수증번호, 객층, 상품명, 수량, 금액, 구분'의 총 10개로 구성되어있는데 영수증에 기록된 판매 품목만으로 고객의 구매 의도를 예측하겠다는 본 연구의 목적에 따라 '상품명' 필드만 추출하여 표본으로 삼았다.

편의점에서 판매되는 제품의 종류가 다양한 관계로 전체 1,334개의 데이터에 포함되어있는 품목들을 제품이 가지고 있는 성질의 유사성을 기준으로 <표 4>와 같이 총 21개의 카테고리로 분류하였다.

본 연구에서는 카테고리 하나가 실험에서 하나의 변수로 사용되는 것이므로, 결국 실험에는 총 21개의 변수가 사용되는 것이다. 그리고, 한 건의 거래 내역당 일련번호 N을 지정해줌으로써 훈련용 데이터 셋(training data set)과 검증용 데이터 셋(test data set)을 추출할 때 중복되는 것을 방지하였고, 변수로 지정된 21개의 모든 카테고리군에도 일련번호(ID) W를 부여하였다. (N = {1, 2, …, 1334}, W = {1, 2, …, 21}) 각각의 거래 내역 N에서 고객이 카테고리군 변수 W를 구입

〈표 4〉 카테고리(변수) 정의

ID(w)	카테고리(변수)	품 목
1	가공식품	동원참치, 천하장사 소시지, 유동 골뱅이, 오뚜기 3분 카레, 햄…
2	건강음료	베지밀, 비타 600, 하늘보리, 녹차를 닮은 마음, 남양 십칠차…
3	과자	스윙칩, 팅클, 초코다이제, 초코파이, 새우깡, 후렌치파이…
4	김밥	참치바베큐&제육, 참치마요네즈 캔디김밥, 불고기참치&치킨…
5	냉동식품	하림 스모크 닭다리, 고향만두, 냉동피자, 볶음밥, 스파게티…
6	담배	레종, 디스, 말보로, 에세, 던힐, 더원, 디스 plus, 인디고…
7	라면	삼양 라면, 안성탕면, 튀김우동 큰사발면, 신라면, 왕뚜껑…
8	맥주	하이트 355ML, 카프리 병맥주 330ML, OB 500ML, 코로나…
9	빙과류	쿠엔크바, 크런치킹, 요맘떼…
10	빵	찐빵, 샤니 스위트 페스츄리, 샌드위치, 치즈케익, 샤니 대보름…
11	생수	퓨리스, 해태 평창 샘물, 에비앙…
12	생활용품	덴탈크리닉 2080치약, 스파크일회용 라이타, 위스퍼클린…
13	소주	진로 참이슬, 두산 산, 백세주, 산사춘…
14	신문	조선일보, 중앙일보, 스포츠신문, 일반 서적, 잡지…
15	요구르트	매일 구트, 덴마크 요구르트, 남양 불가리스, 생크림 요구르트…
16	우유	매일 우유(초코, 딸기), 서울 우유, 남양 진짜 초콜릿 듬뿍…
17	주스	서울 아침에 주스, 델몬트 망고 스카시, 후레쉬 믹스, 쿨피스…
18	초코렛	스니커즈, 트윅스, 크라운 미니쉘, 가나 초코렛, 자유시간…
19	캔디	후라보노 껌, 자이리톨 껌, 츄파춥스, 호올스…
20	커피	레쓰비 마일드, 네스까페, 까페라떼, 프렌치 카페,산타페…
21	탄산음료	코카콜라, 칠성 사이다, 데미소다, 밀키스, 맥콜, 환타…

〈표 5〉 입력 데이터의 형태

W / N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
1334	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

하였을 때는 1, 구입하지 않았을 경우에는 0을 지정해 줌으로써 분석이 용이하도록 하였다. (<표 4>, <표 5>참조)

총 1,334건의 거래 내역에 구매된 제품의 수는 2,965개이고 이중에서 우유가 507개로 전체 17.1%의 가장 큰 비중을 차지하였으며, 그 다음으로는 냉동식품이 276개로 전체 9.3%를 차지하였다.

3.2 실험 설계 및 실행

본 절에서는 먼저 실험에 사용되는 데이터의 분류 및 적용 방법을 설명하고, 그 다음 단계로 데이터를 각각의 데이터마이닝 기법들에 적용시켜 그에 대한 결과를 도출하고자 한다.

본 연구의 목적은 고객의 특정 상품에 대한 구매 의도를 다른 상품들에 대한 구매 패턴에 근거하여 예측하는 것이기 때문에, 21개의 카테고리 변수들 중에서 하나를 선택하여 종속변수로, 나머지 20개의 카테고리 변수들을 독립변수로 지정하였다.

예측을 위한 실험은 결과에 대한 신뢰성을 위하여 크게 2번 실행되는데, 첫 번째 실험의 종속변수는 21개의 카테고리 중에서 가장 거래량이 많은 우유로 지정하였으며, 두 번째 실험의 종속변수는 그 다음으로 거래량이 많은 냉동식품으로 지정하였다. 이것은 모든 실험의 목표 결과가 고객의 우유 및 냉동식품 구매 여부에 대한 예측 정확도가 얼마나 높은지에 관한 것임을 의미한다.

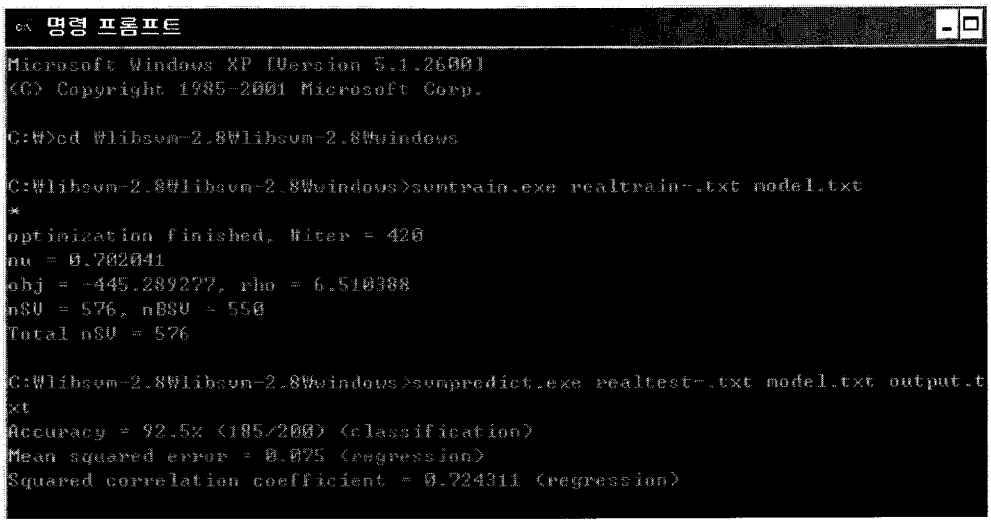
실험에 앞서 학습과 검증을 위한 데이터의 분류가 필요하여 우선적으로 전체 1,334건의 거래 내역 중 우유를 구매한 거래 500건과 우유를 구매하지 않은 거래 500건을 무작위로 추출하여 1,000건의 데이터를 생성하였으며, 이 중에서 학습을 위한 데이터는 새로 생성된 1,000개 데이터의 80%로 우유 구매 400건과 우유를 구매하지 않은 400건을 합하여 800건의 거래 내역을 사용하였고, 검증을 위한 데이터는 나머지 20%를 사용하였다.

냉동식품 또한 구매한 거래 274건과 구매하지 않은 거래 274건을 무작위 추출하여 전체 548건의 데이터를 생성하였으며, 학습 데이터와 검증 데이터의 분류는 우유 데이터의 분류방법과 동일하게 이루어졌다.

본 연구 결과에 대한 신뢰성을 높이기 위하여 각각의 알고리즘에 대하여 10번씩 교차 검증(cross validation)을 하였고, 이 때 총 10개의 데이터 셋을 실험마다 다르게 추출하여 중복되지 않게 적용하였다. 그리고, 각 기법의 실험 방법은 품목에 관계없이 동일함으로 거래량이 가장 많은 우유 품목을 기준으로 설명하였다.

3.2.1 SVM(support vector machine)

SVM 실험은 Chin-Chung Chang and Chin-Jen Lin이 2005년 11월에 발표한 LIBSVM version 2.81을 사용하였다. 입력 데이터의 형태는 LIBSVM에 적용 가능한 텍스트 파일로 변환하였으며, 거래



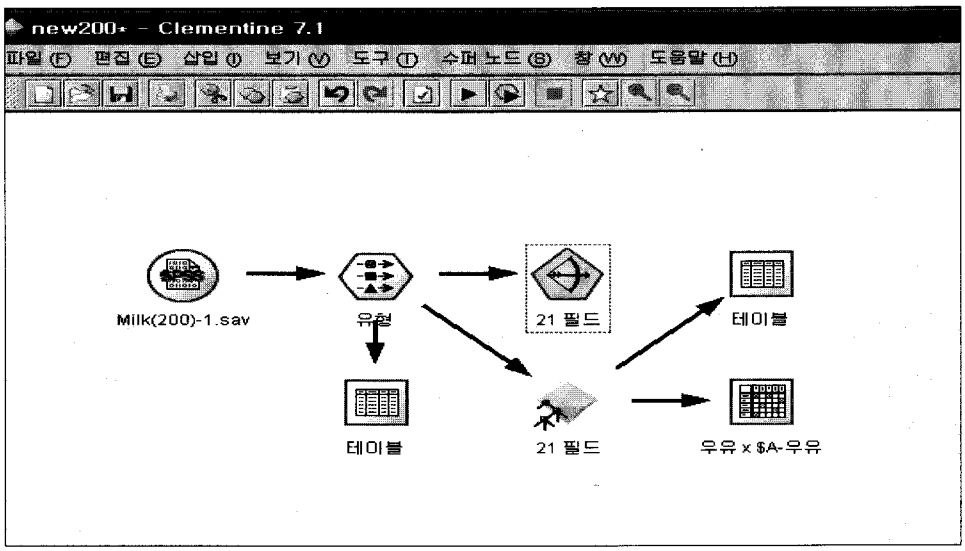
〈그림 4〉 SVM 예측 결과 화면의 예

별 구매 여부에 대한 결과가 맨 앞부분에 나와 있는 데이터 형태의 특성상 입력 변수는 우유를 뺀 20개로 지정되었다. 실험의 목표 결과는 학습용 데이터 셋에 의해 생성된 모델이 검증용 데이터 셋의 우유에 대한 구매 여부를 얼마나 정확하게 분류 및 예측하는 가에 관한 것이다.

첫 번째 데이터 셋의 적용 결과는 <그림 4>와 같이 거래 내역 200건의 검증용 데이터셋에서 우유 구매 여부에 대하여 185건을 예측하는 92.5%의 정확도를 보이고 있다.

3.2.2 연관성 규칙(association rule)

본 연구에서는 연관성 규칙을 이용한 예측 정확도를 측정하기 위하여 두 가지 알고리즘을 사



〈그림 5〉 Clementine 7.1 GRI 알고리즘 모델의 예

용하였다. 첫 번째는 SPSS사의 데이터마이닝 솔루션인 clementine 7.1의 GRI 알고리즘을 이용하였고, 두 번째는 연관성 매트릭스(association matrix)를 작성하여 변수들 간의 연관 정도 및 규칙을 살펴봄과 동시에 우유 구매의 예측 정확도를 측정하였다.

Clementine 7.1의 GRI 알고리즘에서는 최소 규칙 지지도 0%, 최소 규칙신뢰도 50%, 최대 전항값 수 3, 최대 규칙수 100으로 조건을 설정을 하였으며, 본 연구의 목적이 우유를 구매하는 사람들을 예측하는 것이므로 이분형에 대한 참값만 이용하도록 설정하였다.

<그림 5>는 GRI 알고리즘을 이용하여 연관성 규칙을 발견하는 모델을 나타낸 것이다.

그리고, 상품들의 구매 빈도를 점수화하여 연관 정도를 파악함으로써 규칙을 추출하는 프로그램인 연관성 매트릭스(association matrix)를 품목들 간의 연관성 규칙을 발견하는 또 하나의 도구로 활용하였다. 연관성 매트릭스의 원리를 예를 들어 설명하자면, 거래 N에 구매된 상품이 3(과자)와 7(라면)이라면 3번과 7번이 만나는 교차구역에 1점을 추가하는 것이다. 만약 거래 상품이 2(건강음료), 1(가공식품), 6(담배)라면 상품번호를 {1, 2, 6}과 같이 오름차순으로 정렬한 다음 {1, 2}, {1, 6}, {2, 6}의 형태로 변형시켜 각각의 해당 교차점에 1점을 추가하는 것이다. 그 이유는 연관성 매트릭스는 <그림 6>과 같이 두 가지 상품간의 연관 정도만을 측정할 수 있기 때문이다. 이런 방법으로 전체 거래 내역에 대한 연관성 매트릭스를 작성하면 거래 빈도가 제일 높은 상품 2가지가 가장 연관성이 높다는 것을 알 수 있다.

연관성 매트릭스를 이용한 우유 구매 예측에 대한 정확도 측정은 800개의 훈련용 데이터에서 우유를 구매한 거래 내역 400건, 우유를 구매하지 않은 거래 내역 400건의 연관성 매트릭스를 분리해서 작성한다. 그리고, 검증용 데이터 200건에 대한 각각의 거래 내역에 포함된 상품들을 연관성 매트릭스 작성원리와 같은 방법으로 상품 집합들을 나눈다. 이렇게 구분된 상품 집합들을 우유 구매 400건과 비구매 400건에 대한 매트릭스 각각에 대입해 봄으로써 우유 구매 여부를 구분하게 되며, 이러한 원리로 전체 검증용 데이터에서 우유를 구매하는 거래의 비중을 계산하여 그 정확도를 측정하는 것이다. 이러한 연관성 매트릭스를 이용한 예측 정확도 측정 과정의 예는 <그림 7>에서 보여주고 있으며, 정확도를 계산하는 방법은 식 (1)과 같다.

3.2.3 의사결정나무(decision tree)

의사결정나무의 경우에는 clementine 7.1의 C5.0 알고리즘을 사용하여 우유 구매에 대한 예측 정확도를 측정하였다. 가장 많이 알려진 결정트리 알고리즘으로 ID3, C4.5, C5.0 등이 있는데 본 실험에서는 정확도, 속도, 메모리 측면에서 ID3나 C4.5 보다 성능이 많이 향상되고 부스팅 알고리

거래 1 : {3, 7}

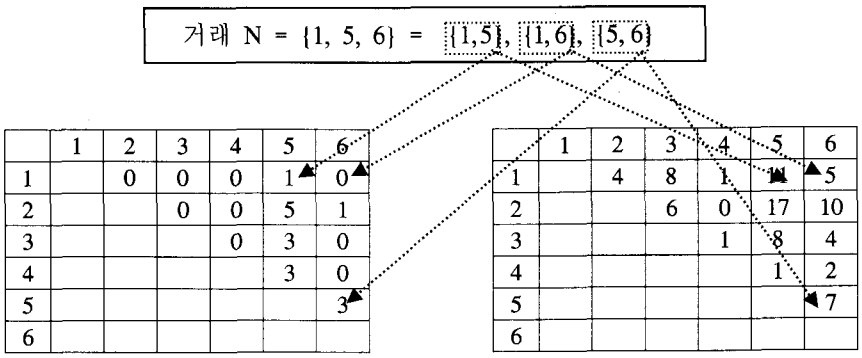
거래 2 : {2, 1, 6}

⇒ {1, 2}, {1, 6}, {2, 6}

	1	2	3	4	5	6	7
1		1	0	0	0	1	0
2			0	0	0	1	0
3				0	0	0	1
4					0	0	0
5						0	0
6							0
7							

주) 중복을 피하기 위하여 matrix 우측에만 점수를 기입하는 것을 원칙으로 함.

〈그림 6〉 연관성 매트릭스 모형의 예



	1	2	3	4	5	6
1		0	0	0	1	0
2			0	0	5	1
3				0	3	0
4					3	0
5						3
6						

〈우유를 구매한 matrix〉

	1	2	3	4	5	6
1		4	8	1	11	5
2			6	0	17	10
3				1	8	4
4					1	2
5						7
6						

〈우유를 비구매한 matrix〉

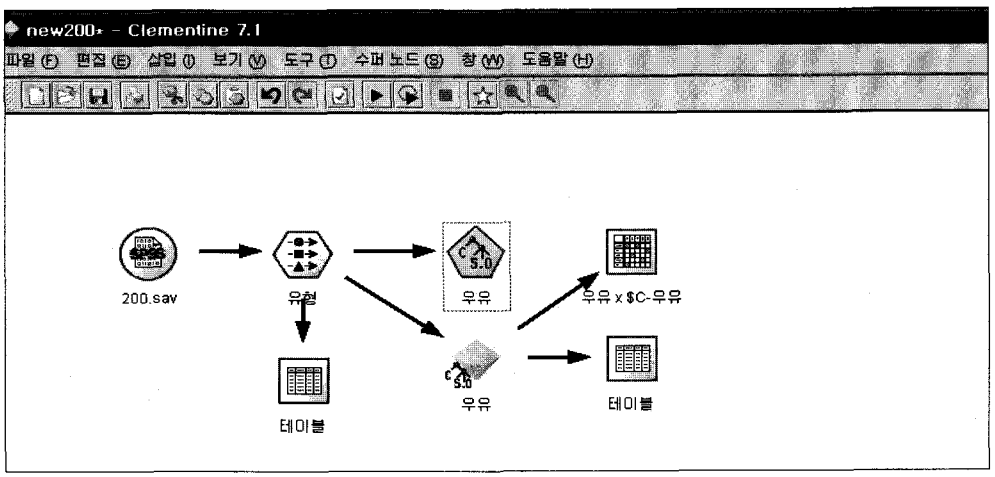
◦ 우유 구매 matrix 대입 결과 = 1 + 0 + 3 = 4(점)
◦ 우유 비구매 matrix 대입 결과 = 11 + 5 + 7 = 23(점)
⇒ 우유 비구매 matrix의 점수가 더 높음으로 거래 N은 우유를 구매하지 않음.

〈그림 7〉 association matrix를 이용한 예측 정확도 측정 과정의 예

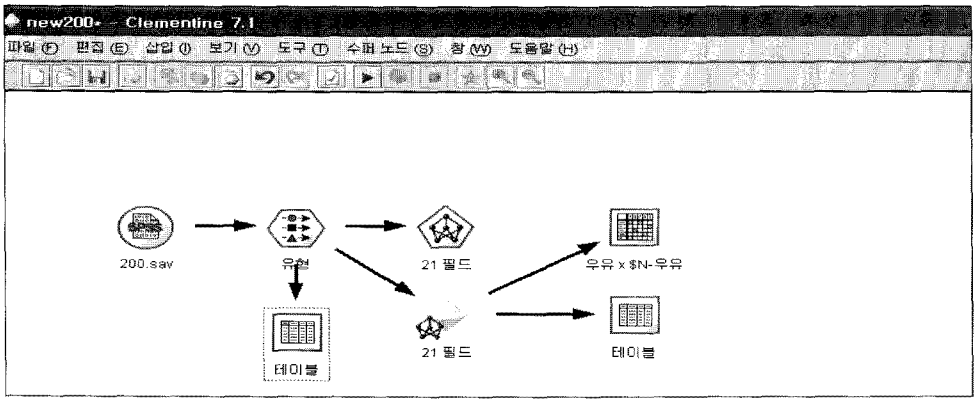
$$\frac{\text{검증용 데이터의 거래 내역을 훈련용 matrix(우유 구매/비구매) 대입 결과 우유 구매 matrix의 점수 값이 더 큰 거래 N의 총 개수}}{\text{200(검증용 데이터의 총 개수)}} * 100 = \text{예측 정확도(\%)} \quad (1)$$

즘이 포함된 C5.0을 이용하였다.

본 연구의 목적이 우유 제품에 대한 구매 예측의 정확도를 측정하는 것이기 때문에 목표필드는 우유, 입력 필드는 나머지 20개의 변수로 지정하였으며, 우선기준은 정확도로 설정하였다. <그림 8>은 clementine 7.1의 C5.0 알고리즘 모델



〈그림 8〉 clementine 7.1의 C5.0 알고리즘 모델



〈그림 9〉 Clementine 7.1의 신경망 알고리즘 모델

을 보여주고 있다.

3.2.4 신경망(neural networks)

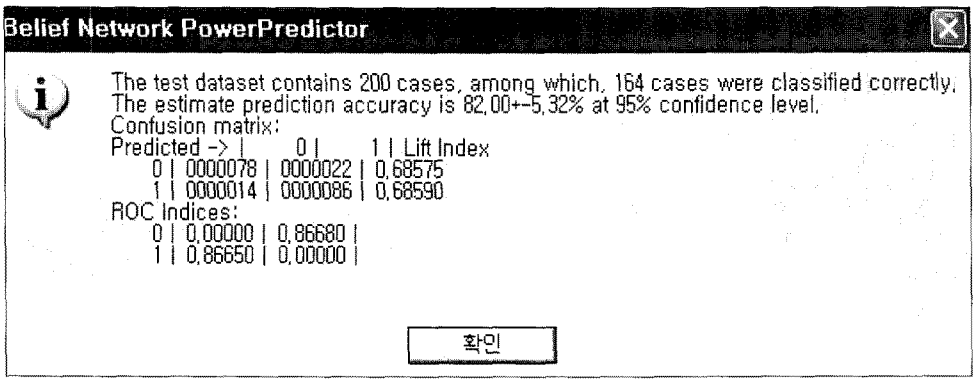
신경망의 실험에서도 연관규칙이나 의사결정나무와 마찬가지로 clementine 7.1의 신경망 알고리즘을 사용하여 우유 구매 예측의 정확도를 측정하였다. 신경망 알고리즘의 예측 정확도 측정을 위하여 <그림 9>와 같은 모델을 작성하였다.

3.2.5 베이지안 망(bayesian network)

베이지안 망 기법의 실험을 위하여 사용한 프로그램은 Jie Chen이 2001년 10월에 발표한 BN PowerSoft Package를 사용하였다. 프로그램의 특성상 기존의 엑셀 파일로 저장되어있던 1,000개의 데이터를 엑세스 파일 형태의 데이터베이스로 변환시켰다. 그리고, BN PowerSoft Package의 구성요소인 Data Preprocessor을 이용하여 훈련용 데이터를 먼저 변환시키고, BN PowerPredictor에서 훈련용 데이터를 학습하여 나온 결과를 토대로 검증데이터의 우유 구매에 관한 예측 정확도를 측정하였다. <그림 10>은 BN PowerPredictor을 실행시켜 나온 결과 화면의 예이다.

3.2.6 로지스틱 회귀분석(logistic regression analysis)

회귀 분석을 통한 우유 구매 예측의 정확도를



〈그림 10〉 BN PowerPredictor 실행 결과 화면의 예

측정하기 위하여 SPSS 12.0의 binary logistic regression analysis을 사용하였다. 종속변수를 우유로 설정하였으며, 독립변수를 우유를 제외한 20개의 카테고리 변수들로 지정하였다. 우유 구매의 예측 정확도에 관한 실험 결과의 예는 <그림 11>과 같이 나타내어진다.

Classification Table[a]

Observed			Predicted		
			우유		Percentage Correct
			0	1	
Step 1	우유	0	98	2	98.0
		1	7	83	93.0
	Overall Percentage				95.5

주) [a] The cut value is .500.

〈그림 11〉 logistic regression analysis 실험 결과 화면의 예

3.3 실험 결과 비교 및 분석

본 절에서는 전체 데이터마이닝 기법들에 대한 실험 결과를 종합해보고, 그것을 토대로 비교 및 분석을 하고자 한다. SVM을 포함한 모든 데이터마이닝 기법들에 대한 실험은 측정되는 결과의 신뢰성을 높이기 위하여 예측하고자 품목인 우유와 냉동식품에 대하여 10번의 반복 실험을 하였으며, 각 실험에 적용된 데이터 셋(훈련용/학습용)은 전체 거래 내역에서 중복되지 않게 구분되었다. <표 6>과 <표 7>은 본 논문에서 실행된 우유와 냉동식품에 대한 실험의 결과들을 종합해서 보여주고 있다.

<표 6>에 나타난 결과를 분석해 보면, 신경망과 로지스틱 회귀분석, SVM이 80% 이상의 높은 예측 성과를 나타내고 있고, 그 중에서도 SVM은 약 90%의 예측력을 보이며 전체 데이터마이닝 기법들 중에서 가장 뛰어난 성과를 보이고 있다. <표 7>의 결과에서도 <표 6>과 마찬가지로 SVM이 약 80%에 가까운 가장 높은 예측력을 보이며, 그 다음으로 신경망과 로지스틱 회귀분석이 우수한 예측 성과를 나타내고 있다. 따라서, <표 6>과 <표 7>에 나타난 우유와 냉동식품에 대한 기법별 예측 정확도를 종합적으로 정리 및 비교해보면 <표 8>과 같다.

<표 8>에서 나타난 결과와 같이 우유와 냉동식품에 대한 고객의 구매 의도 예측 정확도는 support vector machine(SVM)이 두 경우 모두 가

〈표 6〉 데이터마이닝 기법들의 우유 구매 예측 정확도에 대한 실험 결과

	association rule (GRI)(%)	association matrix(%)	bayesian network(%)	decision tree (C5.0)(%)	neural networks(%)	logistic regression(%)	SVM (%)
1	15.0	52.5	66.1	85.0	93.0	93.0	92.5
2	24.0	51.5	68.6	78.0	85.0	97.0	89.0
3	51.0	51.0	66.6	78.0	88.0	89.0	92.0
4	52.0	52.0	66.7	79.0	93.0	94.0	92.0
5	52.0	51.0	67.3	82.0	92.0	93.0	92.0
6	50.5	50.5	66.4	73.0	89.0	87.0	89.5
7	52.0	52.0	65.5	77.0	83.0	83.0	87.5
8	52.5	52.5	65.1	67.0	82.0	81.0	87.5
9	52.0	52.0	64.8	74.0	86.0	83.0	92.0
10	52.0	52.0	66.1	76.0	85.0	86.0	87.5
average	45.3	51.7	66.3	76.9	87.6	88.6	90.2

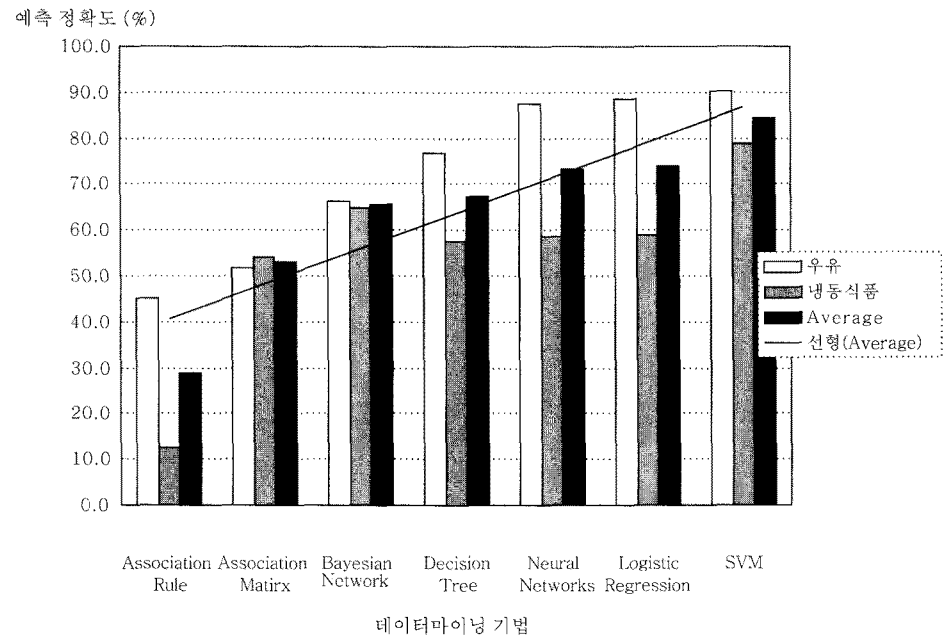
〈표 7〉 데이터마이닝 기법들의 냉동식품 구매 예측 정확도에 대한 실험 결과

	association rule (GRI) (%)	association matrix (%)	bayesian network(%)	decision tree (C5.0) (%)	neural networks(%)	logistic regression(%)	SVM (%)
1	21.0	56.4	68.1	61.0	68.0	67.0	85.7
2	14.0	56.4	68.3	59.0	64.0	62.0	85.0
3	13.0	54.3	68.8	56.0	57.0	59.0	82.1
4	13.0	52.1	67.1	61.0	57.0	58.0	83.5
5	11.0	55.0	64.4	57.0	62.0	62.0	77.1
6	12.0	54.3	61.3	56.0	57.0	61.0	72.8
7	10.0	53.6	62.3	58.0	56.0	59.0	72.8
8	10.0	54.3	63.7	56.0	54.0	54.0	75.7
9	11.0	52.1	62.6	56.0	57.0	56.0	75.7
10	10.0	54.3	62.9	58.0	56.0	52.0	72.8
average	12.5	54.3	64.9	57.8	58.8	59.0	78.3

장 높게 측정되고 있다.

본 실험에서는 고객의 상품 구매 의도 예측에 대한 기존의 연구에서 쓰였던 다양한 변수들(예를 들어 나이, 성별, 교육 등)과 달리 본 연구에서는 영수증 데이터를 자료로 하였기 때문에, 거래당 구매되었던 2~3가지의 상품들만이 한 건의 거래에 영향을 주는 변수로 사용되었다. 뿐만 아니라, 전체 입력 데이터를 구매의도를 예측하고자 하는 품목을 구매한 거래 데이터와 구매하지 않은 거래 데이터로 구분하여 만들었으며, 훈

〈표 8〉 우유와 냉동식품에 대한 구매 의도 예측 정확도 비교



련용 데이터와 검증용 데이터도 정확하게 구매 여부를 기준으로 하여 50%씩 분류하였기 때문에 본 연구의 실험 결과가 기존의 구매의도 예측에 관한 연구 결과보다 전체적으로 우수하게 나타나는 경향이 있었다.

특히, 본 실험의 목표가 과거 판매되었던 상품들의 패턴을 분석하여, 고객이 특정 상품들을 구매할 경우 그 고객의 다른 특정 상품의 구매의도를 파악함으로써, 상품 추천의 실행 유무를 판단하는 것이기 때문에 데이터마이닝 기법들 중에서 이항적인 분류를 목표 결과로 하는 로지스틱 회귀 분석과 SVM이 높은 예측 정확도를 보였다. 그 중에서도 로지스틱 회귀 분석과 같은 통계적 기법의 한계점을 보완한 SVM이 조금 더 정확한 예측율을 보여줌으로써 그 우수성이 입증되었다.

따라서, SVM을 고객구매 예측모형에 적용하고 그 성능을 신경망 및 로지스틱 회귀분석의 성능과 비교 분석하여, SVM의 우수성을 입증한 안현철 등(2004)의 연구 결과와 마찬가지로 영수증 데이터를 이용한 고객의 상품 구매의도를 예측하는데 가장 유용한 데이터마이닝 기법은 SVM이라는 것을 알 수 있다.

Ⅳ. 결 론

4.1 연구 결과 및 시사점

본 연구에서는 고객의 상품 구매의도를 예측하는데 있어서 기존의 다양한 데이터마이닝기법들과 최근 패턴인식 및 분류문제와 관련하여 활발하게 연구되고 있는 SVM을 적용하여 그 결과를 비교 및 분석함으로써 가장 성능이 우수한 기법을 발견하고자 하였다. 실험 과정에서는 데이터마이닝 기법뿐만 아니라 개체간의 연관 정도를 통해 예측의 정확도를 측정할 수 있는 연관성 매트릭스(association matrix)를 추가적으로 작성하였고, 결과의 대한 신뢰도를 높이기 위하여 예측 목표 품목을 우유와 냉동식품으로 바꾸어서 교차 검증하였으며, 각각의 품목에 대하여 모든 적용 기법들을 입력 데이터를 달리하여 10번씩 반복 실험하였다. 실험 결과, 전체 7가지 기법들 중에서 SVM이 가장 우수한 예측 정확도를 보여주고 있다.

이러한 비교연구는 서로 다른 이론적 배경을 가진 다양한 기법들의 활용에 시사점을 제공해 줄 수 있었다. 본 연구의 목적과 같이 고객의 상품 구매 여부를 예측하는 데는 SVM과 같은 기법이 유용하지만 연관성 규칙이나 연관성 매트릭스를 이용하여 고객들이 상품을 구매하는 패턴을 확인할 수 있으며, 구매 고객들에 대한 좀 더 다양한 정보가 있다면 베이지안 망이나 의사결정나무 등을 이용하여 고객의 특정 상품에 대한 구매 행동을 결정하는 요인들을 파악할 수도 있다. 따라서, 우수한 CRM 전략의 수립을 위해서는 전략 수립의 목적과 보유하고 있는 데이터의 형태에 따라 다양한 데이터마이닝 기법들을 적용 할 수 있어야 하며, 그 결과들을 적절하게 조합시킬 수 있는 능력도 필요하다.

4.2 연구의 한계점 및 향후 연구 방향

본 연구는 결과를 도출하기 위한 과정에서 다음과 같은 한계점을 가지고 있다.

첫 번째는 데이터의 크기와 구성에 대한 문제이다. 전체 데이터의 수가 작기 때문에 데이터의 크기가 커짐에 따라 발생하는 데이터마이닝 기법들의 단점들을 파악할 수 없었다. 그리고 데이터의 구성도 예측하고자 하는 품목에 대한 구매자와 비구매자를 50%씩 같은 비율로 맞춤으로써 기법들의 결과가 실체 예측력보다 전체적으로 우수하게 나타나게 되었다. 두 번째는 실험에서 사용된 입력 변수 즉, 거래 건당 구매되는 상품의 수를 2~3개로 제한함에 따라, 본 연구의 결과가 이론상으로는 적용이 가능할 수도 있으나 실제 대량의 상품이 거래되는 대형 할인마트나

백화점에서는 적용하는데 무리가 있을 수 있다.

본 연구의 향후 연구 방향으로는 연구 결과를 분석하였을 때, SVM의 예측력이 로지스틱 회귀분석이나 신경망에 비해 뛰어나긴 하지만 그 차이가 다소 미비하여, 통계적으로 유의한 수준의 우수함을 입증하지 못하였다. 따라서, 좀 더 다양한 형태의 변수를 가지거나 보유하고 있는 정보의 양이 많은 데이터를 적용시켜 그 결과를 비교함으로써 좀더 정확한 성능비교를 해 볼 필요가 있겠다.

그리고, 국내 마케팅 특히 CRM 분야에서 연관성 규칙 기법이나 조금 더 나아가 신경망 기법 외의 SVM를 포함한 다른 데이터마이닝 기법들의 연구나 활용도가 낮은 점을 고려해볼 때, 좀 더 다양한 실험적 연구가 필요하리라고 본다. 특히 아직까지 거의 마케팅에서의 활용사례가 없는 support vector machine(SVM)이 국외에서는 고객 반응이나 패턴 인식과 많은 부분에서 연구 성과들을 내고 있음을 간과해서는 안 될 것이다.

참 고 문 헌

김종우, 이경미, "인터넷 상점에서 개인화 광고를 위한 장바구니 분석 기법의 활용", 한국경영과학회, 제17권, 제3호, 2000.

류문배, 장남식, "의류 판매 자료의 실증적 분석을 통한 연관관계 발견", 한국경영정보학회 춘계학술대회, 1999, pp. 351-360.

민재형, 이영찬, "Support Vector Machine을 이용한 부도예측모형의 개발: 격자탐색을 이용한 커널 함수의 최적 모수 값 선정과 기존 부도예측모형과의 성과 비교", 한국경영과학회, 제30권, 제1호, 2005, pp. 55-74.

송수섭, 이의훈, "인공신경망을 이용한 소비자 선택 예측에 관한 연구", 한국경영과학회, 제26권, 제4호, 2001.

안현철, 김경재, 한인구, "Support Vector Machine을 이용한 고객구매예측모형", 한국지능정보시스템학회, 제11권, 제3호, 2004.

알렉스 버슨, 스테판 스미스, 커트 티어링, Entrue Consulting CRM 그룹, CRM을 위한 데이터마이닝, 대청, 2000.

이군희, 사회과학연구방법론, 법문사, 2004.

이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구", 한국지능정보시스템학회, 제9권, 제1호, 2003, pp. 139-155.

이학식, 안광호, 하영원, 소비자 행동-마케팅전략적 접근, 법문사, 2001.

이학식, 임지훈, SPSS 12.0 매뉴얼: Statistical package for the social science: 통계분석 방법 및 해설, 법문사, 2005.

주영진, "장바구니분석을 이용한 주식투자전략 수립 방안", 한국데이터베이스학회, 정보기술과 데이타베이스 저널, 제9권, 제4호, 2002, pp. 65-78.

정영미, "필터링 기법을 이용한 도서 추천 시스템 구축", 정보관리연구, 제33권, 제1호, 2002, pp. 1-17.

최종후 외, SAS Enterprise Miner 4.0을 이용한 데이터마이닝 방법론 및 활용, 3판, 자유아카데미, 2001.

하선영, 장병탁, "Reversible Jump MCMC와 베이지안망 학습에 의한 데이터마이닝", 한국정보과학회, 2000.

한상만, 박승배, 정남호, "인공신경망과 로짓모형을 이용한 내구재의 구매의도 예측에 관한 비교연구", 한국마케팅학회, 2004.

허명회, 이용구, 데이터마이닝 모델링과 사례, SPSS 아카데미, 2003.

허준, 최병주, 정성원, 클레멘타인을 이용한 데이터마이닝 입문편, SPSS 아카데미, 2001.

황인수, "연관규칙을 이용한 상품선택과 기대수익 예측", 경영정보학연구, 제14권, 제4호, 2004.

Alex Berson, Stephen Smith, Kurt Thearing, Building Data Mining Apllications for CRM, McGraw-

Hill, 1999.

Ashok Savasere, Edward Omiecinski, Shamkant Navathe, "An Efficient Algori-thm for Mining Association Rules in Large Databases", *Proceeding of the 21th International Conference on Very Large Data Bases*, 1995, pp. 432-444.

Berry, J. A. Michael, and Gordon Linoff, Data Mining Techniques: For Marketing, Sales and Customer Support, Wiley Computer Publishing, 1997.

Burke. R, "Knowledge-based recommender systems", *Encyclopedia of Library and Information Systems*, Vol.69, 2000.

Chang, C.-C. and C.-J. Lin, LIBSVM: a library for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2001.

Chiu, C., "A case-based customer classification approach for direct marketing", *Expert Systems with Applications*, Vol.22, 2002. pp. 163-168.

Francesco Ricci, Fabio Del Missier, "Personalized Product Recommendation through Interactive Query Management and Case-Based Reasoning", Proceedings of CHI'03 Workshop on Designing Personalized User Experiences for eCommerce, Fort Lauderdale, USA, 2003.

Ganti, V., J. Gehrke and R. Ramakrishnan, "Mining very large databases", *IEEE Computer*, Vol.32, No.8, 1999, pp. 38-45.

Gunn, S. R., "Support Vector Machines for Classification and Regression", Technical Report, University of Southamton, 1998.

Hsuan-Tien Lin, "Introduction to Support Vector Machines", Learning System Group, Califonia Institute of Technology, 2005.

HyunJung Shin, Sungzoon Cho, "Response Modeling with Support Vector Machines", Preprint submitted to Elsevier Science, 2005.

Jain, Bharat A. and Nag, Barin N., "Performance Evaluation of Neural Network Decision Models", *Journal of Management Information Systems*, Vol.14, No.2, 1997, pp. 201-230.

Kate A. Smith, "Nueral Networks: An Intrduction", Neural Networks for business, 2002.

Marti A. Hearst, "Trends and Controversies: Support Vector Machines", IEEE Intelligent Systems, 1998.

Margahny, M. H. and A. A. Mitwaly, "Fast Algorithm for Mining Association Rules", AIML 2005 Confrence, 2005.

Peppers and Roggers Group(Asia), "Customer Realtionship Management in Asia", 2002.

Robert Burdidge, Bernard Buxton, "An introduction to Support Vector Machines for Data Mining", YOR 12 conference Data Mining Stream, 2001.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, "Leaning International Representations by Error Propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition", Vol.1, Chapter 8, Cambridge, MA: MIT, 1986, pp. 318-364.

Seong-Whan Lee and Alessandro Verri (eds.), "Pattern recognition with support vector machines", first international workshop, SVM 2002, Niagara Falls, Canada, 2002.

SPSS, "Data mining with Clementine for smarter retailing: White paper executive briefing", SPSS USA, 2000.

Sun-Mi, Lee and Patricia A. Abbott, "Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers", *Journal of Biomedical Informatics*, 2003.

Tom Brijs, Gilbert Swinnen, Koen Vanhoof, Geert Wets, "Using Association for Product Assortment Decision: A Case Study", KDD-99 San Diego CA USA, 1999.

J. Cheng, http://www.cs.ualberta.ca/~jcheng/bnsoft.htm., 2001.

Information Systems Review
Volume 10 Number 2
August 2008

Forecasting of Customer's Purchasing Intention Using Support Vector Machine

Jin Hwa Kim* · Ki Chan Nam** · Sang Jong Lee***

Abstract

Rapid development of various information technologies creates new opportunities in online and off-line markets. In this changing market environment, customers have various demands on new products and services. Therefore, their power and influence on the markets grow stronger each year.

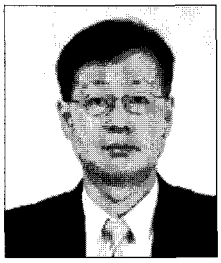
Companies have paid great attention to customer relationship management. Especially, personalized product recommendation systems, which recommend products and services based on customer's private information or purchasing behaviors in stores, is an important asset to most companies. CRM is one of the important business processes where reliable information is mined from customer database. Data mining techniques such as artificial intelligence are popular tools used to extract useful information and knowledge from these customer databases.

In this research, we propose a recommendation system that predicts customer's purchase intention. Then, customer's purchasing intention of specific product is predicted by using data mining techniques using receipt data set. The performance of this suggested method is compared with that of other data mining technologies.

Keywords: Customer Relationship Management, Recommendation System, Receipt Data, Data Mining, Forecasting of Customer's Purchasing Intention

* Associate professor, Sogang University, Dept. of Business Administration
** Professor, Sogang University, Dept. of Business Administration
*** Ph.d course, Sogang University, Graduate School of Business

◐ 저 자 소 개 ◑

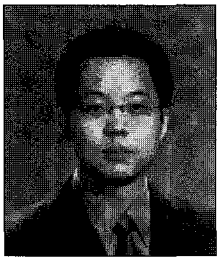


김 진 화 (jinhwakim@sogang.ac.kr)

현재 서강대학교 경영학과 부교수로 재직 중이다. 서강대학교 경영학과와 영문과를 졸업하고 University of Wisconsin-Madison에서 경영학 석사, 전산학 석사, 경영학 박사를 취득하였다. 주요 연구분야는 데이터 마이닝, 의사결정지원 시스템, 미래예측, CRM 등이다. 현재 전자거래학회, 지능정보 시스템 학회 임원으로 활동하고 있다.

남 기 찬 (knam@sogang.ac.kr)

현재 서강대학교 경영학과에서 교수로 재직 중이다. 서강대학교 영문과를 졸업하고, University of Mississippi에서 MBA, 그리고 State University of New York Buffalo에서 MIS로 박사학위를 취득하였다. 주요 연구관심분야는 IT 아웃소싱, SLA, Service Management, ASP, IT 성과평가, Service Science, Service Innovation 등이며 Journal of Management Information Systems, Information Systems Research, Communications of the ACM, International Journal of Electronic Commerce, Decision Support Systems, European Journal of Operational Research, Expert Systems with Applications, International Journal of Information Management, Information Systems Frontier 뿐만 아니라 여러 국외 및 국내 학술지에 논문을 게재하고 있다.



이 상 종 (basur949@sogang.ac.kr)

현재 서강대학교 경영전문대학원 MIS 전공 박사과정에 재학 중이다. 아주대학교 경영학부를 졸업하고, 서강대학교 일반대학원 경영학과에서 MIS로 석사학위를 취득하였다. 동부 CNI와 스킨푸드에서 ERP 및 CRM 관련 업무를 담당하였으며, 주요 연구관심분야는 데이터마이닝, CRM, 인공지능 등이다.

논문접수일 : 2007년 07월 22일 **게재확정일 : 2008년 01월 14일**
1차 수정일 : 2007년 10월 08일