소비자 구매행동 예측을 위한 이질적인 모형들의 통합

배 재 권[a] 김 진 화[b]

[a]*서강대학교 경영학과*
서울시 마포구 신수동 1번지, 121-845
Tel: +82-2-705-8860, Fax: +82-2-705-8519, E-mail: baejaekwon@sogang.ac.kr

[b]*서강대학교 경영학과*
서울시 마포구 신수동 1번지, 121-845
Tel: +82-2-705-8860, Fax: +82-2-705-8519, E-mail: jinhwakim@sogang.ac.kr

Abstract

For better predictions and classifications in customer recommendation, this study proposes an integrative model that efficiently combines the currently-in-use statistical and artificial intelligence models. In particular, by integrating the models such as Association Rule, Frequency Matrix, and Rule Induction, this study suggests an integrative prediction model. The data set for the tests is collected from a convenience store G, which is the number one in its brand in S. Korea. This data set contains sales information on customer transactions from September 1, 2005 to December 7, 2005. About 1,000 transactions are selected for a specific item. Using this data set, it suggests an integrated model predicting whether a customer buys or not buys a specific product for target marketing strategy. The performance of integrated model is compared with that of other models. The results from the experiments show that the performance of integrated model is superior to that of all other models such as Association Rule, Frequency Matrix, and Rule Induction.

Keywords:

Customer Recommendation; Artificial Intelligence; Association Rule; Frequency Matrix; Rule Induction; Integrative Prediction Model

1. 서론

최근 디지털 정보기술의 급속한 발전은 다양한 시장공간을 창출시키고 있으며, 특히 인터넷 매체의 빠른 확산은 새로운 경제현상을 만들어낼 뿐만 아니라 기업의 경쟁전략을 변화시키고 있다. 이러한 시장환경의 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 고객의 다양한 정보를 획득함과 동시에 고객과의 밀접한 관계를 유지함으로써 기업의 수익성을 증대시키는 고객관계관리(CRM: Customer Relationship Management)에 대한 관심과 활용에 대한 필요성은 점점 높아지고 있다. 고객관계관리는 데이터베이스 마케팅의 진보된 형태로서 기업이 고객과 상호작용하는 프로세스를 보다 자동화하고 개선시킨 것이다. 고객관계관리에서는 효과적인 고객관리 전략을 개발하고 지속적으로 수행하는 능력이 중요하며, 이를 위해서는 고객정보를 분석하는 도구로서 데이터마이닝의 사용이 요구되고 있다.

고객관계관리의 여러 분야 가운데에서도 제품을 구매한 기존 고객의 정보를 기반으로 고객에게 맞는 새로운 제품이나 서비스를 제안하기 위하여 고객의 구매 패턴을 파악하고 의도를 예측하는 것은 오늘날 실질적인 판매 전략을 수립하는 마케팅 분야에서 상당히 큰 비중을 차지하고 있다. 일반적으로 고객의 구매의도를 파악하고 이를 예측하는 데는 연관성규칙, 의사결정나무, 인공신경망 등의 데이터마이닝 기법들이 주로 사용되어왔다. 그러나 이들 데이터마이닝 기법을 이용한 단일모형은 몇 가지 태생적인 한계점을 가지고 있으며 데이터 특성에 따라 각각의 모형에 대한 예측력 성과가 달라질 수 있기 때문에 어느 모형이 가장 최적의 모형인지 판단하기 어렵다. 따라서, 기존의 데이터마이닝 기법들이 가지고 있는 한계점들을 최소화하기 위하여, 이질적인 단일모형들을 지식 결합을 이용하여 시너지 효과를 생산할 수 있는 통합모형을 제시하고자 한다.

본 연구에서는 보다 효과적인 고객구매예측을 위하여, 매장 내의 상품들과 고객구매패턴과의 연관성을 발견하기 위해 가장 널리 활용되고 있는 연관성규칙(Association Rule)과 인공지능적인 방법으로서 널리 사용되고 있는 빈도행렬(Frequency

Matrix), 규칙유도기법(Rule Induction)의 3가지 모형을 규칙기반으로 통합한 통합모형을 제시하고자 한다.

통합모형은 연관성규칙(A)과 빈도행렬(B)을 결합한 ASFM(ASsociation rule and Frequency Matrix)모형, 연관성규칙(A)과 규칙유도기법(C)을 결합한 ASRI (ASsociation rule and Rule Induction)모형, 빈도행렬 (B)과 규칙유도기법(C)을 결합한 FMRI(Frequency Matrix and Rule Induction)모형, 마지막으로 3가지 모형(A+B+C)을 모두 통합한 ASFMRI(ASsociation rule, Frequency Matrix and Rule Induction)모형으로 구성되어 있다. 이러한 통합모형의 성과를 증명하기 위해 서울시 용산구에 위치한 *G* 편의점으로부터 확보한 1,334건의 거래 내역 데이터를 기초로 분석하였고 그 결과를 기존 단일모형과 성과비교를 통하여 그 유용성을 검증해보고자 한다.

2. 이론적 배경

경영학 분야에서 통합모형으로 가장 많이 연구되고 있는 분야가 바로 부도예측이다. 부도예측에 관한 최근의 연구들은 전통적인 통계기법(판별분석, 로짓, 프로빗)과 인공지능 기법(신경망, 의사결정나무)의 성과를 비교, 분석하는 연구[5, 7, 8, 9, 11, 14, 19, 20]에서 나아가 통합방법론을 통해 모형의 예측력을 향상시키기 위한 방안들을 제시하고 있다. 이에 관련된 연구로 Lee et al.[12, 13]의 연구에서는 인공신경망과 다변량 판별분석, 귀납적 학습방법, SOFM(self organizing feature map) 등의 기법을 통합한 귀납적 학습지원 인공신경망을 제시하였고 실증분석결과 향상된 예측성과를 나타내었다. Shin and Lee[16]의 연구에서는 다수의 인공신경망 모형을 통합한 부도예측모형을 제시하였다. 다수의 신경망 모형의 결과에 따라 데이터를 분류하고 재분류된 데이터를 학습시켜 보다 나은 데이터의 패턴을 신경망에 적용하였다. 실험결과 다수의 신경망 모형을 통해 데이터를 분류 및 재학습시킨 결과 신경망 예측치와 실제 부도 사이의 일치 여부가 크게 개선되었다.

퍼지 기법은 경영 문제에도 두루 적용되어 왔으며 특히 금융공학 문제 및 경영 의사결정분야에 주로 활용되고 있다. 또한 퍼지 이론은 신용평가 문제에도 성공적으로 적용되고 있는데, 특히 사례기반추론 시스템과의 혼합적 적용에 의하여 기존의 연구에 비해 향상된 결과를 보여준 바 있다. 김경재와 한인구[1]의 연구에서는 퍼지신경망을 이용한 기업부실예측모형을 제안하였다. 이것은 기존 신경망에 퍼지집합의 개념을 적용하여 신경망 학습에 사용될 자료를 퍼지화하고 이를 신경망에 학습시키는 것이다. 퍼지신경망을 기업부도예측에 적용한 결과 기존의 신경망보다 우월한 예측성과를 나타내었다.

또한, 유전자 알고리즘이 기업부도예측에 적용된 연구가 많이 진행된 바 있는데, 이들 연구는 기업부도예측 문제에 다양한 분류 기법들을 통합적으로 적용하기 위하여 유전자 알고리즘을 사용한다. Shin and Lee[17]에서는 유전자 알고리즘을 패턴인식과 학습이 뛰어난 인공신경망의 가중치를 훈련시키거나 아키텍처를 설정하는데 통합하여 사용하는 Neuro-Genetic 인공신경망 모형을 제시하였다. 외국의 연구로는 Anandarajan et al.[6]이 유전자 알고리즘에 기반한 인공신경망 모형을 개발하였으며 이 모형을 역전파 인공신경망, 다변량 판별분석 모형과 예측성과를 상호 비교하여 분석하였다. Chen and Huang[10]과 Abdelwashed and Amir[3]의 연구에서도 유전자 알고리즘에 기반한 신경망 모형을 제시하였다. 또한 Pendharkar[15]는 유전자 알고리즘에 기반한 새로운 Threshold-varying 인공신경망을 제시하였으며 이 모형을 역전파 인공신경망, 다변량 판별분석 모형 등과 예측성과를 상호 비교하여 분석하였다.

위 모형들은 알고리즘 관점에서 통합한 실질적인 통합모형이다. 실제 응용분석에 있어서 단순히 한 알고리즘만을 선택하여 모형을 적용시키는 것이 아닌 적용 가능한 알고리즘들의 장점을 통합한 통합모형 구축에 대한 이슈는 많은 실무자들에게 연구의 대상이 되고 있다. 또한 통합모형은 단일모형에 의해 추출된 결과들을 통합하여 하나의 결론으로 도출할 수 있다는 점에서 예측 분야에 유용하게 쓰일 수 있을 것이다.

3. 연구 방법

3.1 자료 수집

고객 구매의도를 다양한 데이터마이닝 알고리즘들을 이용하여 예측하여 보고, 그 성과를 비교 및 분석하기 위하여 본 연구에서는 실제 데이터를 적용하여 해당 결과를 도출하였다.

본 연구에 사용된 데이터는 서울시 용산구에 위치한 *G* 편의점의 판매 자료이다. 1990년 12월에 1호점을 개점한 *G* 편의점은 국내 독자 개발 브랜드로 시작하여 2007년 현재까지 업계 1위를 굳건히 지켜오고 있는 국내 최고의 편의점이다.

G 편의점으로부터 확보한 자료는 지난 2005년 9월 1일부터 12월 7일 사이에 고객들이 구매한 1,334건의 거래 내역 데이터이다. 거래 내역은 *G* 편의점 지점 판매시점 정보관리 시스템(POS system)으로부터 확인하였으며, 이 시스템의 필드는 <표 1>에서 보는 바와 같이 판매일자, 판매시간, 담당자, 영수증번호, 상품명, 수량, 구매금액, 구분/비고의 총 8개로 구성되어 있으며 영수증에 기록된 판매 품목만으로 고객의 구매의도를 예측하겠다는 본 연구의 목적에 따라 '상품명'

필드만 추출하여 표본으로 선정하였다.

<표 1> 판매시점 정보관리 시스템의 테이블 정의서

코 드	필 드	설 명
DP	Date_Purchased	판매일자
TP	Time_Purchased	판매시간
EN	Employee Number	담당자
RN	Receipt Number	영수증번호
IN	Item_Name	상품명
QU	Quantity	수량
PC	Purchase_Cost	구매금액
DE	Description	구분/비고

3.2 변수 선정 및 사전처리

편의점에서 판매되는 제품의 종류가 다양한 관계로 전체 1,334개의 데이터에 포함되어 있는 품목들을 제품이 가지고 있는 성질의 유사성을 기준으로 <표 2>와 같이 총 21개의 카테고리로 분류하였다.

<표 2> 변수 목록

Category	Definition	Category	Definition
1	가공식품	12	생활용품
2	건강음료	13	소주
3	과자	14	신문
4	김밥	15	요구르트
5	냉동식품	16	우유
6	담배	17	주스
7	라면	18	초콜릿
8	맥주	19	캔디
9	빙과류	20	커피
10	빵	21	탄산음료
11	생수	-	-

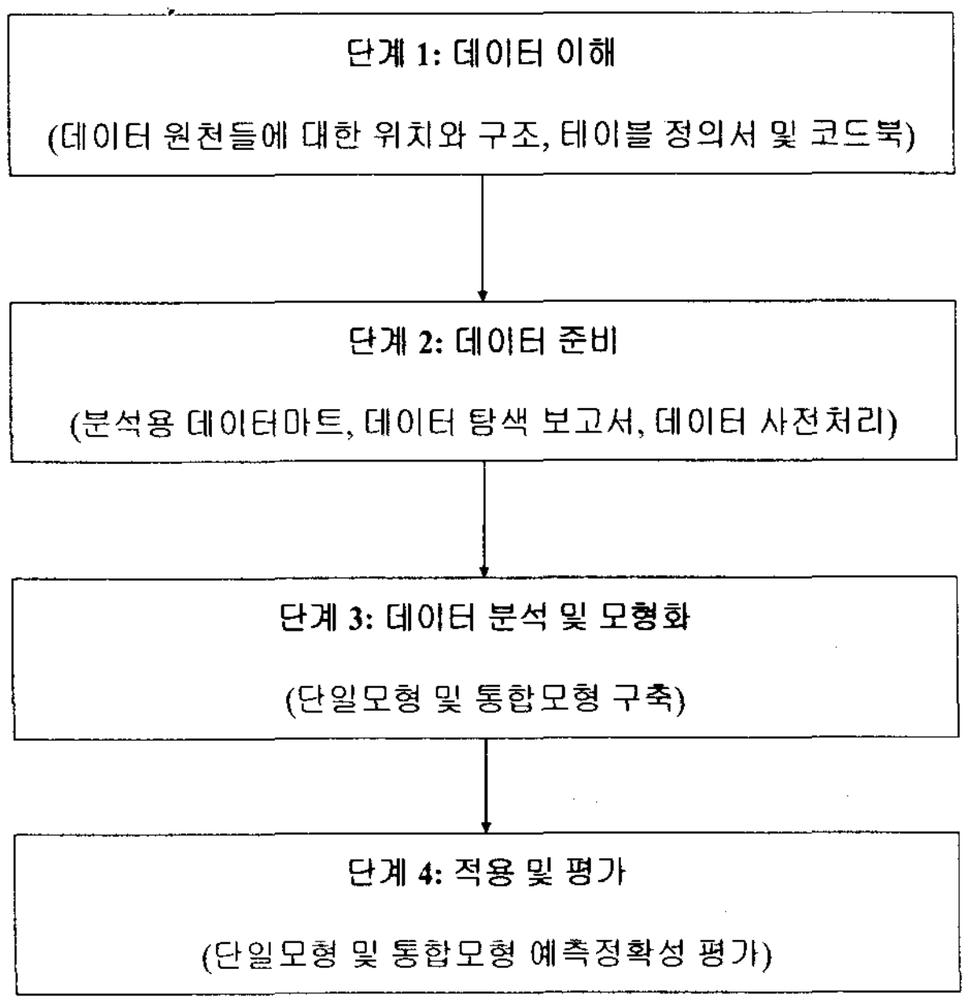
본 연구에서는 카테고리 하나가 실험에서 하나의 변수로 사용되는 것으로, 실험에는 총 21개의 변수가 사용되었다. 한 건의 거래 내역당 일련번호(N=1,2,…,1334)를 지정해줌으로써 훈련용과 검증용 데이터 셋을 추출할 때 중복되는 현상을 방지하였고, 21개의 카테고리 변수에서도 일련번호(W=1,2,…,21)를 부여하였다. 각각의 거래 내역 N에서 고객이 카테고리 변수 W를 구입하였을 경우 1은 구매, 0은 비구매를 나타내도록 지정함으로써 분석이 용이하도록 설계하였다.

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. 실험설계로 종속변수는 21개의 카테고리 중에서 구매거래량이 가장 많은 우유로 지정하였으며 이것은 모든 실험의 목표결과가 고객의 우유구매 여부에 대한 예측정확도가 얼마나 높은지에 관한 것임을 의미한다.

모든 분석은 훈련용과 검증용의 두 가지 데이터 셋으로 구성되었으며 우선적으로 전체 1,334건의 거래 내역 중 우유를 구매한 거래 500건과 우유를 구매하지 않은 거래 500건을 무작위로 추출하여 1,000건의 데이터를 생성하였으며, 전체 데이터의 80%(800/1,000)는 훈련용 데이터 셋으로 사용하고, 나머지 20%(200/1,000)는 검증용 데이터 셋으로 사용하였다. 또한 보다 일반화된 연구결과를 얻기 위하여 본 연구에서는 상호검증방법(cross-validation method)을 사용하였다. 따라서 총 10회에 걸친 상호검증방법을 실시하였다.

3.3 분석절차

본 연구는 <그림 1>과 같은 분석절차에 따라 진행된다.



<그림 1> 분석절차

단계 1에서는 데이터에 대한 이해와 분석이다. 데이터마이닝의 성공여부는 사용 가능한 데이터의 양과 질에 전적으로 의존하며, 사용 가능한 데이터를 검토하고 데이터들의 특징을 이해하는 것이 좋은 모형을 만들기 위한 첫 단계가 된다. 따라서, 사용 가능한 데이터의 파악, 데이터 원천들에 대한 위치와 구조, 데이터 테이블의 필드와 그들의 코드 분석, 마지막으로 데이터들의 신뢰성, 정확성, 유용성을 검토해야 한다. 단계 2에서는 데이터 사전처리를 실시하는 데이터 준비단계이다. 이 단계에서는 제품 또는 고객단위의 레코드가 구성될 수 있도록 재배열(Rearrangement) 해야 한다. 또한 품목들을 제품이 가지고 있는 성질의 유사성을 기준으로 총 21개의 카테고리로 분류하는 그룹화(Grouping) 과정을 실시한다. 단계 3에서는 만들어진 데이터마트를 이용하여 데이터에 대한 분석 및 예측모형의 구축을 수행하는 데이터 분석 및 모형화 단계이다. 본 연구에서는 연관성규칙, 빈도행렬, 규칙유도기법의 단일모형과 이들을 규칙기반(rule-based)으로 통합한 3가지 통합모형을 구축한다. 단계 4에서는 실무데이터를 가지고 단일모형과 통합모형에 적용시킨 후 다양한 평가도구들을 이용하여 이들 모형의 성능을 평가하고, 최종적인 예측모형을 결정한다.

4. 연구 모형

4.1 연관성규칙(Association Rule)

데이터마이닝 기법 중 하나인 연관성규칙은 데이터들의 빈도수와 동시발생확률을 이용하여 한 항목들의 그룹과 다른 항목들의 그룹 사이에 강한 연관성이 있음을 밝혀주는 기술이다.

연관성규칙은 상품 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있는 기법이다. 동시 구매될 가능성이 큰 상품들을 찾아내기 때문에 시장바구니분석을 다루는 문제에 많이 적용된다. 측정의 기본은 얼마나 자주 구매되었는가 하는 빈도이다. 이 빈도를 기반으로 연관 정도를 정량화하기 위해서 다음의 두 가지 기준을 고려한다[4].

(1) 지지도(Support) – 전체 거래 중 항목 *A*와 항목 *B*를 동시에 포함하는 거래가 어느 정도 인가를 나타내주며 전체적 구매의도에 대한 경향을 파악할 수 있다.

(2) 신뢰도(Confidence) – 항목 *A*를 포함하는 거래 항목 *B*가 포함될 확률이 어느 정도 인가를 나타내주며 연관성의 정도를 파악할 수 있다.

지지도는 생성된 연관규칙이 전체 아이템에서 차지하는 비율을 말한다. 즉, 데이터베이스에 속한 전체 트랜잭션 중 *A*와 *B*를 포함하는 트랜잭션의 비율을 의미한다. 신뢰도는 연관규칙의 강도를 의미하며 전제 부를 만족하는 트랜잭션이 결론 부를 만족하는 비율, 즉 *A*를 포함하는 트랜잭션 중에서 *B*가 포함된 트랜잭션의 비율을 의미한다[18].

위 기준을 이용해 연관규칙을 탐사하는 과정에는 기본적으로 다음의 두 단계로 구성된다. 첫 번째 단계는 빈발 항목집합들(large itemsets)을 찾아낸다. 빈발 항목집합이란 최소지지도를 설정하여 사용자가 한꺼번에 구매하는 물품들의 집합(트랜잭션)에서 빈번하게 발생하는 트랜잭션이 그 지지도 이상 발생한다면 이것을 빈발 항목집합이라고 한다. 두 번째 단계는 데이터베이스로부터 연관성규칙을 생성하기 위하여 빈발 항목집합을 사용한다. 연관성규칙의 전체 성능은 첫 번째 단계에서 결정된다. 먼저, 빈발 항목집합을 확인한 후에 해당되는 연관 규칙을 두 번째 단계의 방법으로 쉽게 유도할 수 있다.

연관성 규칙은 일반적으로 구매 행위에 있어서 특정 아이템과 다른 아이템 간에 어떤 연관관계가 있는지를 찾아보는 것이다. 연관성 규칙은 비록 다른 데이터마이닝 기법에 비해 단순하지만, 일반적으로 조건-반응(If-then)으로 결과가 표현되어 이해하기 쉽고 또 이를 바로 실제에 적용하기 용이하다는 장점이 있다.

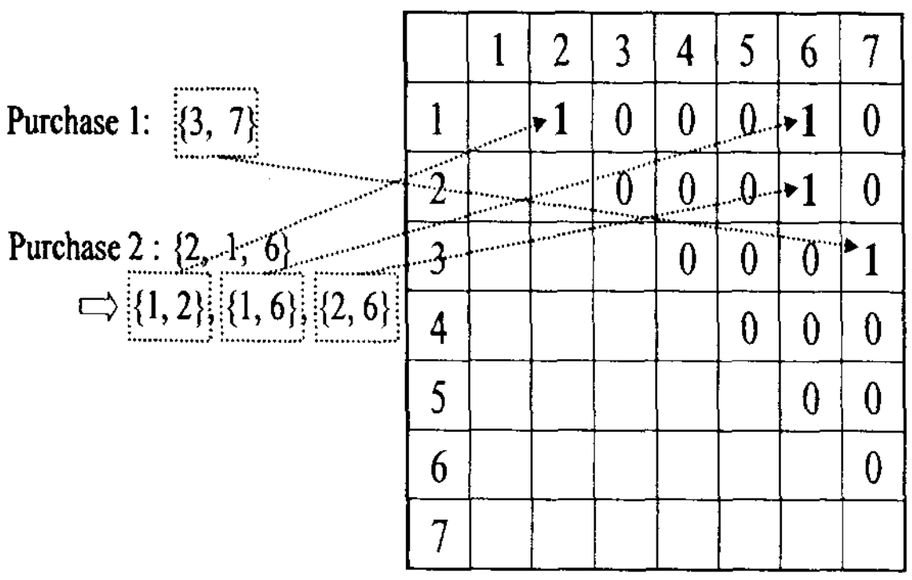
구매예측을 위한 데이터 셋에 적용되는 연관성 규칙은 *SAS E-Miner 4.0* 패키지를 적용해 도출하였다. 규칙 선정을 위한 기준은 지지도를 0.5%, 신뢰도를 10%로 설정하였으며, 그 결과 <표 3>과 같이 총 11개의 규칙이 추출되었다.

<표 3> 연관성규칙 적용 결과(Data Set 1)

지지도(%)	신뢰도(%)	해 석
5.5	82	초콜릿 → 우유
3.5	86	가공식품 → 우유
3.5	57	담배 → 캔디
2	50	건강음료와 빵 → 요구르트
1.5	67	과자와 빵 → 맥주
1.5	67	신문 → 담배
1	100	우유와 캔디 → 초콜릿
1	50	빙과류 → 생활용품
1	50	과자와 냉동식품 → 라면
1	50	건강음료와 냉동식품 → 커피
1	50	빙과류 → 과자

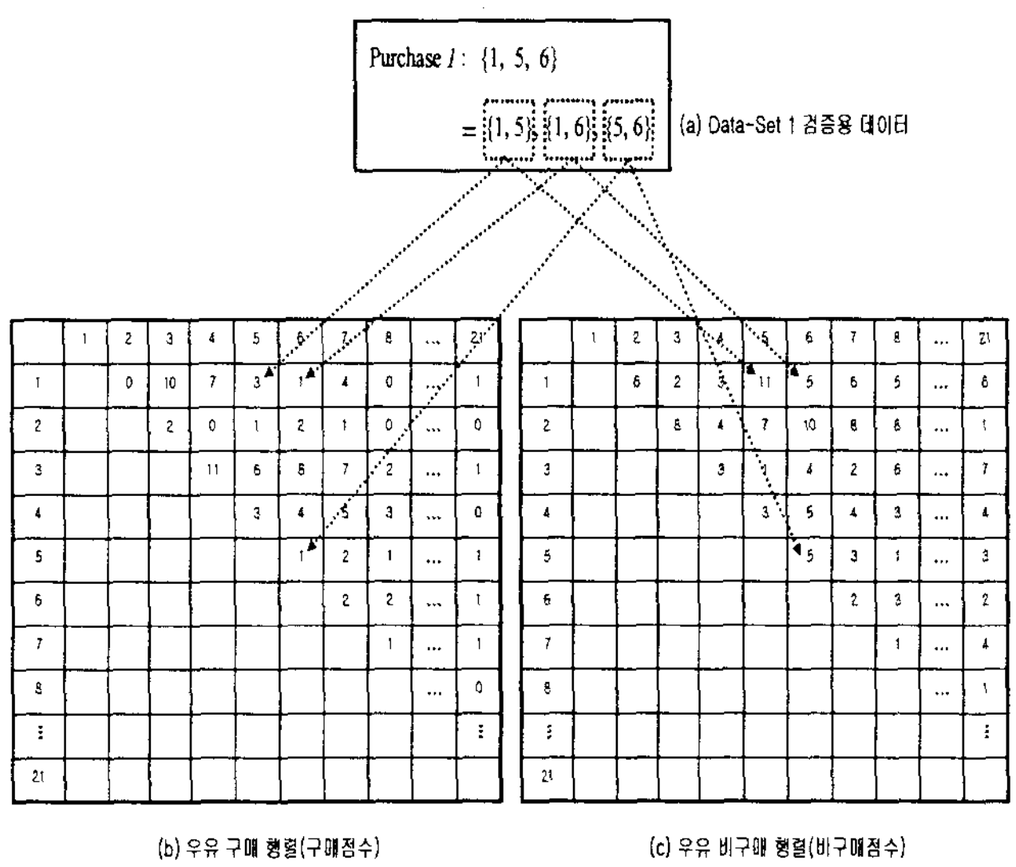
4.2 빈도행렬(Frequency Matrix)

빈도행렬은 인접행렬(Adjacency Matrix)의 개념에서 출발한다. 인접행렬은 데이터의 인접성을 이용하여 의사결정공간에서 유용하게 쓰일 수 있는 개념[2]으로 품목 *A*와 *B*가 존재할 때 품목 *A*와 *B*가 동시에 구매되었는지, 또는 품목 *A*가 *B*의 구매에 영향을 주었는지 그 여부를 확인할 수 있어 데이터마이닝의 기법 중 연관규칙 분석이나, 또는 추천시스템 및 데이터 시각화 등에서 이용되고 있다. 빈도행렬은 상품들의 구매 빈도를 점수화하여 연관 정도를 파악함으로써 규칙을 추출하는 알고리즘이다. 따라서 빈도행렬을 이용하여 품목들간의 연관성 규칙을 발견할 수 있다. 빈도행렬의 원리를 그림으로 구체화하면 <그림 2>와 같은 n × n 행렬로 나타낼 수 있다. <그림 2>에서 보는 바와 같이 거래 *N*에 구매된 상품이 3(과자)과 7(라면)이라면 3번과 7번이 만나는 교차구역에 1점의 빈도를 추가하는 것이다. 또한 구매 상품이 2(건강음료), 1(가공식품), 6(담배)이라면 상품번호를 {1,2,6}과 같이 오름차순으로 정렬한 다음 {1,2},{1,6},{2,6}의 형태로 변형시켜 각각의 해당 교차점에 1점의 빈도를 추가하는 것이다. 이는 빈도행렬은 2차원의 형태로 두 가지 상품간의 연관 정도만을 측정할 수 있기 때문에 거래 구매상품이 3개 이상인 경우는 위와 같은 형태로 변형시켜야 한다. 이와 같은 방법으로 전체 거래 내역에 대한 빈도행렬을 작성하면 거래 빈도가 제일 높은 상품들의 조합과 그들의 연관성 수치를 알 수 있다.



<그림 2> 빈도행렬의 학습/훈련 알고리즘

<그림 3>에서는 첫 번째 검증용 데이터 셋에서 빈도행렬의 예측기법을 설명한 것이다. 우선, (b)와 (c)에서 보는 바와 같이 전체 훈련용 데이터를 바탕으로 각 구매에 해당하는 빈도를 계산하여 우유구매 행렬과 우유비구매 행렬을 생성한다. 위 두 행렬을 바탕으로 검증용 데이터에서 거래 구매상품에 해당하는 빈도수를 대입하여 각각의 구매레코드에 대한 구매점수와 비구매점수를 산출한다. 예를 들면 (a)의 검증용 데이터 구매 1번은 1(가공식품), 5(냉동식품), 6(담배)를 구매한 형태이다. 이에 대한 (b)의 우유구매 행렬 값은 {1,5} = 3, {1,6} = 1, {5,6} = 1의 값을 가지며 이들을 모두 합한 구매점수는 5점이 된다. 또한 (c)의 우유 비구매 행렬 값은 {1,5} = 11, {1,6} = 5, {5,6} = 5 값을 가지며 이 값을 모두 합한 비구매점수는 21점이 된다. 따라서 빈도행렬 예측결과는 구매점수와 비구매점수 중에서 큰 값이 예측결과로 채택되며 예제 1번 레코드에서는 우유비구매로 판별된다. 빈도행렬기법은 Microsoft의 *Visual Basic .NET*을 이용하여 구체화하였다.



<그림 3> 빈도행렬의 예측기법

<표 4>는 첫 번째 데이터 셋에서 빈도행렬 기법을 이용하여 규칙을 추출한 것이다. 첫 번째 데이터 셋에서는 빈도가 최소한 8이상인 규칙 9개가 추출되었다. 9개의 규칙 중에서 빵을 구매하면 동시에 우유를 구매한다는 규칙의 빈도가 65로 가장 높았으며, 냉동식품을 구매하면 동시에 우유를 구매한다는 규칙의 빈도가 57을 나타냈다. 세 번째와 네 번째 규칙으로는 과자를 구매하면 동시에 우유를 구매한다는 규칙의 빈도가 34를 나타냈으며 담배를 구매하면 동시에 우유를 구매한다는 규칙의 빈도는 31을 나타내었다. 이들 4가지의 규칙은 공통적으로 다른 데이터 셋에서도 높은 빈도 값을 기록하였으며 이것이 이들이 가장 영향력 있는 규칙이라는 것을 의미한다.

<표 4> 빈도행렬 적용 결과(Data Set 1)

규칙번호	해 석	빈 도
규칙 1	IF Bread Buy Then Milk Buy (빵을 구매하면 동시에 우유를 구매)	65
규칙 2	IF Frozen food Buy Then Milk Buy (냉동식품을 구매하면 동시에 우유를 구매)	57

규칙 3	IF Sweets Buy Then Milk Buy (과자를 구매하면 동시에 우유를 구매)	34
규칙 4	IF Tobacco Buy Then Milk Buy (담배를 구매하면 동시에 우유를 구매)	31
규칙 5	IF Chocolate Buy Then Milk Buy (초콜릿을 구매하면 동시에 우유를 구매)	20
규칙 6	IF Health beverage Buy Then Milk Buy (건강음료를 구매하면 동시에 우유를 구매)	15
규칙 7	IF Household items Buy Then Milk Buy (생활용품을 구매하면 동시에 우유를 구매)	13
규칙 8	IF Spring water Buy Then Milk Buy (생수를 구매하면 동시에 우유를 구매)	9
규칙 9	IF Processed food Buy Then Milk Buy (가공식품을 구매하면 동시에 우유를 구매)	8

4.3 규칙유도기법(Rule Induction)

본 연구에서는 보다 효과적인 고객구매예측을 위한 방법으로 규칙이라는 형태를 이용한다. 이러한 규칙은 규칙유도기법에 의해 규칙으로 쉽게 전환될 수 있다. 규칙유도기법의 장점은 전체적인 과정에서 불필요한 요소들은 자동적으로 제거가 된다는 것이다. 이것은 사용자 및 분석자에게 데이터에 대해 더 많은 정보를 주게 되고, 사용자가 여타 인공지능 기법을 사용할 때 필드를 선택할 수 있는 기준을 마련하여 더욱 효율적인 인공지능 분석에 도움을 줄 수 있다. 또한 규칙유도기법은 의사결정 나무형식으로 되어 있어 예측 필드에 대해 영향력을 가지고 있는 필드들을 명확히 보여준다. 즉, 규칙이 어떻게 작용하는지 이해하기 위해 웹 분석이나 히스토그램과 같은 방법을 사용할 필요가 없다. 규칙유도기법 시스템은 시스템과 사용자의 상호작용을 가능케 하는 자연어처리 부문, 전문지식을 저장하는 지식베이스, 지식베이스의 내용을 이용하여 사용자의 문제 해결을 도와주는 추론기관의 3부분으로 구성되어 있다.

구매예측에 대한 규칙유도기법 구축과 평가를 위해 *Clementine 8.1* 프로그램을 사용하였다. *Clementine 8.1*은 C5.0 알고리즘과 Build Rule 알고리즘 등의 2가지의 규칙유도 알고리즘이 있으며 이 중에서 C5.0 알고리즘을 사용하였다. 훈련용으로 분류된 데이터를 이용하여 규칙유도기법 모형을 구축한 후 검증용 데이터에 적용하여 분류 예측정확도를 분석하였다.

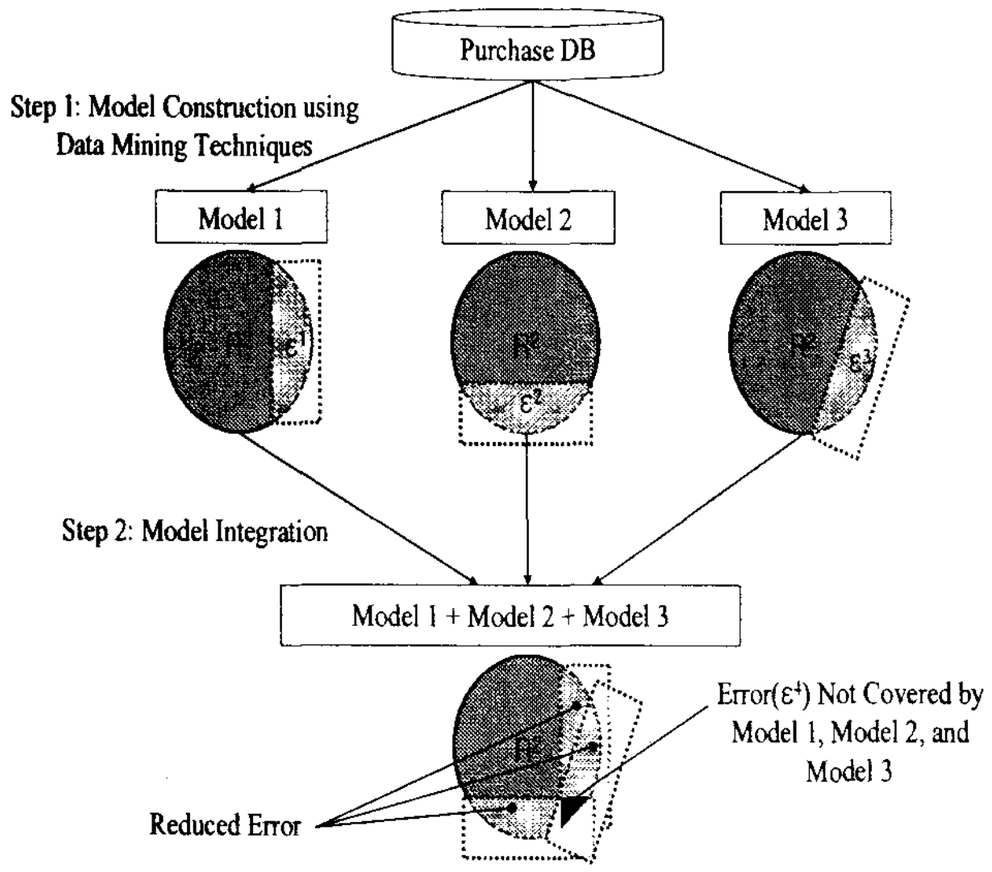
<표 5> 규칙유도기법 적용 결과(Data Set 1)

규칙번호	해 석	신뢰도
규칙 1	IF Chocolate Buy and Candy Buy Then Milk Buy (초콜릿과 캔디를 구매하면 동시에 우유를 구매)	1.0
규칙 2	IF Household items Buy and Ice cream Buy Then Milk Buy (생활용품과 빙과류를 구매하면 동시에 우유를 구매)	1.0
규칙 3	IF Sweets Buy and Bread Buy Then Milk Buy (과자와 빵을 구매하면 동시에 우유를 구매)	1.0
규칙 4	IF Instant noodle Buy and Bread Buy Then Milk Buy (라면과 빵을 구매하면 동시에 우유를 구매)	0.975
규칙 5	IF Tobacco Buy and Carbonated beverage Buy Then Milk Buy (담배와 탄산음료를 구매하면 동시에 우유를 구매)	0.975
규칙 6	IF Rice rolled in dried laver Buy and Frozen food Buy Then Milk Buy (김밥과 냉동식품을 구매하면 동시에 우유를 구매)	0.96
규칙 7	IF Chocolate Buy and Frozen food Buy Then Milk Buy (초콜릿과 냉동식품을 구매하면 동시에 우유를 구매)	0.929
규칙 8	IF Household items Buy and Frozen food Buy Then Milk Buy (생활용품과 냉동식품을 구매하면 동시에 우유를 구매)	0.929
규칙 9	IF Instant noodle Buy and Ice cream Buy Then Milk Buy (라면과 빙과류를 구매하면 동시에 우유를 구매)	0.900
규칙 10	IF Bread Buy Then Milk Buy (빵을 구매하면 동시에 우유를 구매)	0.872
규칙 11	IF Rice rolled in dried laver Buy Then Milk Buy (김밥을 구매하면 동시에 우유를 구매)	0.872
규칙 12	IF Bread Buy and Household items Buy Then Milk Buy (빵과 생활용품을 구매하면 동시에 우유를 구매)	0.845
규칙 13	IF Health beverage Buy and Ice cream Buy Then Milk Buy (건강음료와 빙과류를 구매하면 동시에 우유를 구매)	0.778

<표 5>는 첫 번째 데이터 셋에서 규칙유도기법을 이용하여 규칙을 추출한 것이다. <표 5>에서 제시된 신뢰도(confidence)는 이 규칙에 맞게 분류된 비율을 나타내는 것이다. 신뢰도의 범위는 0과 1사이의 값을 가지며 첫 번째 데이터 셋에서는 신뢰도의 값이 0.778 이상인 13개의 규칙이 도출되었다. 이들 규칙 중에서 초콜릿과 캔디를 구매하면 동시에 우유를 구매한다는 규칙, 생활용품과 빙과류를 구매하면 동시에 우유를 구매한다는 규칙, 과자와 빵을 구매하면 동시에 우유를 구매한다는 규칙들은 1의 신뢰도 값을 나타내었다. 위 3가지 규칙은 공통적으로 다른 데이터 셋에서도 신뢰도 값이 0.95 이상을 기록하였으며 이것은 이들이 가장 영향력 있는 규칙이라는 것을 의미한다.

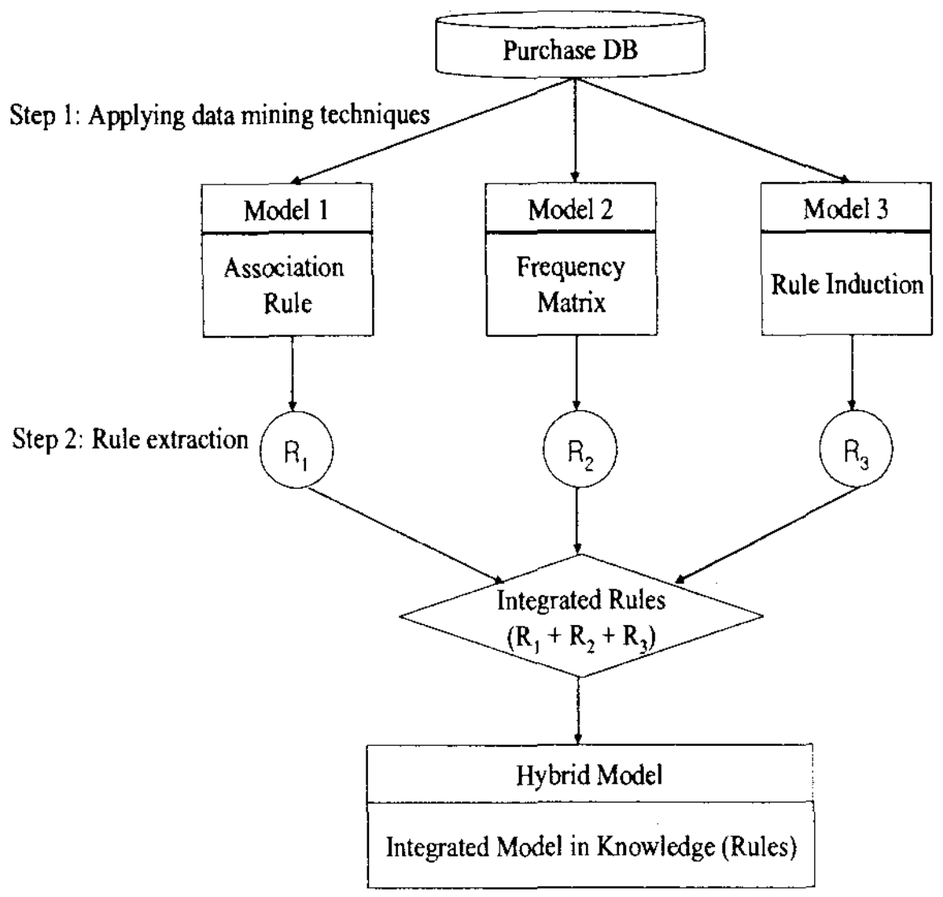
4.4 통합모형

<그림 4>는 통합모형 아키텍처를 설명한 것으로 통합모형을 구축할 경우 오차항(error-term)이 줄어드는 현상을 그림으로 도식화한 것이다. 1단계에서는 구매데이터베이스에서 데이터마이닝 기법을 이용하여 3가지의 구매예측모형을 구축하였다. 각 예측모형에서의 설명력(R-square)과 오차항(ε)이 <그림 4>와 같이 구성되었으며, 여기서 오차항은 각 모형에서 설명할 수 없는 부분을 의미한다. 2단계에서는 위 3가지의 구매예측 모형을 이용하여 규칙기반(rule-based)의 통합모형을 구축할 경우 오차항이 줄어들면서 설명력이 증가하는 좀더 예측력이 뛰어난 모형 구축이 가능하다. 이러한 통합모형 아키텍처의 이론적 배경이 바로 앙상블 접근법(ensemble approaches)이다. 앙상블 접근법은 한 명의 전문가보다 여러 명의 전문가가 한 예측을 종합했을 때 더 나은 결과를 가져올 수 있다는 점에 기초한다. 특히 여러 명의 전문가가 서로 독립적으로 예측결과를 잘못 만들어낼 때에는 그것을 종합했을 때 더욱 정확한 예측을 할 수 있다.



<그림 4> 단일모형을 통합한 통합모형 아키텍처

통합모형 구축과정은 <그림 5>에서 보는 바와 같이 2단계로 구성되어 있다. 첫 번째 단계에서는 연관성규칙, 빈도행렬, 규칙유도기법을 이용하여 구매예측모형을 구축한다. 두 번째 단계에서는 각각의 모형에서 규칙들을 도출한 후 이 규칙들을 결합한 통합모형을 구축한다. 이로써 두 번째 단계에서의 규칙(R_1, R_2, R_3)은 <그림 5>에서 보는 바와 같이 통합된 규칙집합($R_1+R_2+R_3$)으로 전환되어 통합모형에 적용되는 것이다. 즉, 하나의 데이터 셋을 이용하여 각 모형에서 나온 규칙들을 규칙집합(rule set)이라고 정의한다면 통합모형에서의 규칙들은 누적된 규칙집합이라고 말할 수 있다. 이것은 단일모형에서 얻은 규칙집합이 누적되면서 지식으로 축적되기 때문에 단일모형에서의 규칙집합보다 예측력에서 좀더 뛰어난 성능을 가질 수 있다. 즉, 어느 단일모형에서 만들어진 규칙에 의해 분류되느냐에 따라 서로 다른 클래스(비구매, 구매)로 분류될 수 있지만 그것이 누적되면 결국은 실제 분류되어야 할 값과 같은 값으로 분류될 확률을 높일 수 있다는 것이다. 또한 통합모형은 규칙집합 정보만 필요하기 때문에 규칙이 추출되어 집합의 형태로 된 이후에는 원천 데이터 셋이 더 이상 필요하지 않으므로 저장할 필요가 없다. 이는 통계적으로 고정된 데이터가 아닌 시간의 흐름에 따라 무한히 추가되고 그 성격이 변하는 스트림 데이터에 적용할 경우 저장공간, 메모리, 시간 등의 자원 소모를 크게 줄일 수 있다는 장점이 있다. 요약하면, 통합모형을 이루고 있는 누적된 규칙집합은 데이터 셋 대신에 과거 정보를 유지하는 수단으로 저장되는데 그 크기가 현저하게 작으므로 저장공간의 소모가 작고 하나의 데이터 셋을 이용하여 단일모형에서 얻은 규칙집합이 누적되면서 단 하나의 규칙집합 또는 하나의 단일모형으로 예측하는 것보다 예측력에서 좀더 뛰어난 성능을 가질 수 있다.



<그림 5> 통합모형 구축과정

통합모형의 알고리즘과 해석방법을 <그림 6>으로 나타내었다. 우선, 훈련용 데이터에서 연관성규칙, 빈도행렬, 규칙유도기법을 이용하여 예측모형을 구축한 후 각각의 모형에서 규칙들을 도출한다. 이들 단일모형의 모든 규칙들을 누적시킨 규칙집합을 이용하여 200개의 검증용 데이터 셋에 적용한 결과가 바로 통합모형의 예측결과이다. <그림 6>에서 보는 바와 같이 1번 레코드에서는 과자(W_3)를 구매한 고객이 우유를 구매할 것인지에 대한 예측결과를 나타낸 것이다. 단일모형에서 추출된 첫 번째 규칙(Rule 1)은 규칙적용이 불가능한 형태이다. 이것은 과자구매에 따른 우유구매 여부에 관한 1번 레코드에는 첫 번째 규칙이 유용성이 없는 규칙이며 다른 레코드에서 적용이 가능하다는 것을 의미한다. 그러나 두 번째 규칙(Rule 2)과 K 번째 규칙(Rule K)에서는 1번 레코드에 맞는 규칙이 적용되어 우유를 구매한다는 예측결과가 도출되었다. 따라서 이들 누적된 규칙집합의 결과를 합한 값(∑Rule Set)이 양의 값을 가지면 우유구매로, 음의 값을 가지면 우유비구매로 예측을 하며, 1번 레코드에서는 2의 값이 산출되었으므로 통합모형의 예측결과는 우유구매로 판별한다. 즉, 누적된 규칙집합이 검증용 데이터에 적용되어 구매여부를 예측하고 가장 많이 예측된 결과가 통합모형의 예측결과로 판별된다.

ID	W_1	W_2	W_3	…	W_{21}	Rule 1 (R_1)	Rule 2 (R_2)	…	Rule K (R_K)	∑Rule Set	Y
1	0	0	1	…	0	0	1	…	1	2	구매
2	0	0	1	…	1	0	-1	…	0	-1	비구매
3	0	1	1	…	1	-1	-1	…	0	-2	비구매
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	1	0	0	…	0	1	1	…	1	3	구매

<그림 6> 통합모형 결과의 예(Data Set 1)

1) W_i: 21개의 카테고리로 분류된 변수목록 {0 = 비구매, 1 = 구매}
2) Rule K: 연관성규칙, 빈도행렬, 규칙유도기법에서 추출한 K개의 규칙 {- 1 = 규칙적용 우유비구매, 1 = 규칙적용 우유구매, 0 = 규칙적용 불가능}
3) ∑Rule Set: 누적된 규칙(규칙집합)의 결과 합계 {0을 제외한 모든 정수}
4) Y: 통합모형 예측결과 {∑Rule Set < 0이면 우유비구매, ∑Rule Set > 0이면 우유구매}

5. 연구 결과

연구결과 분석을 위하여 10개의 검증데이터의 결과만을 모아 <표 6>을 구성하였다. 평균예측력은 통합모형인 ASFMRI모형(75.35%)이 전체 방법론과 비교하여 가장 예측력이 높았으며 데이터 셋 별로도 ASFMRI모형이 해당 각 데이터 셋에서 예측력이 가장 높다는 것을 알 수 있었다. 또한 2가지의 단일모형을 통합한 ASRI모형(73.45%)과 FMRI모형(69.15%) 또한 단일모형보다 예측력이 더욱 뛰어났으나 3가지 모형을 모두 통합한 ASFMRI 모형(75.35%) 보다는 예측력이 낮다는 것을 알 수 있었다. 또한 단일모형 중에서 빈도행렬은 가장 예측력이 낮은 모형(51.7%)이며 통합에 있어서도 별다른 큰 영향력을 발휘하지 못하였다. 즉, 빈도행렬과 통합한 ASFM모형(65.35%)은 단일모형인 연관성규칙(64.0%)보다는 예측력이 뛰어났으나 규칙유도기법(67.6%)에 비해서는 예측력이 떨어지는 것을 알 수 있었다. 이에 대한 원인으로 다음 2가지를 들 수 있다. 첫 번째 원인은 빈도행렬은 2차원의 행렬형태로 구성되어 있어 두 가지의 물품 $A \rightarrow B$ 형태의 규칙만 생성되기 때문에 다양한 형태의 규칙이 생성되지 못하는 제약(constraint)이 있다.

<표 6> 전체결과

구 분	연관성 규칙(A)	빈도행렬 (B)	규칙유도 기법(C)	ASFM (A + B)	ASRI (A + C)	FMRI (B + C)	ASFMRI (A+B+C)
Set 1	64.0	52.5	68.0	64.0	70.5	68.0	73.5
Set 2	62.5	51.5	69.5	63.5	76.0	69.5	77.5
Set 3	68.5	51.0	73.5	68.5	79.5	73.5	79.5
Set 4	66.5	52.0	73.0	69.5	73.0	75.5	79.0
Set 5	61.5	51.0	64.5	61.5	68.5	65.0	70.0
Set 6	67.0	50.5	70.0	68.0	80.5	70.0	82.5
Set 7	62.0	52.0	68.0	64.5	70.0	69.5	70.0
Set 8	64.5	52.5	64.0	66.5	74.0	69.0	77.0
Set 9	59.0	52.0	64.5	63.0	74.0	66.5	74.0
Set 10	64.5	52.0	61.0	64.5	68.5	65.0	70.5
평 균	**64.0**	**51.7**	**67.6**	**65.35**	**73.45**	**69.15**	**75.35**

두 번째 원인으로는 대부분 빈도행렬에서 추출한 규칙 또는 규칙집합이 연관성규칙이나 규칙유도기법에서 추출한 규칙 또는 규칙집합에 대부분 반영되어 있다는 것을 규칙분석에서 확인할 수 있었다.

<표 7>은 통합모형 중에서 가장 예측력이 뛰어난 ASFMRI모형과 연관성규칙, 빈도행렬, 규칙유도기법의 단일모형과의 McNemar 예측결과를 비교한 것이다. <표 7>에서 보는 바와 같이 ASFMRI모형은 연관성규칙, 빈도행렬과는 5% 수준에서 유의하였고 규칙유도기법과는 10% 수준에서 유의한 차이를 나타내어 이들 단일모형보다 예측성과가 뛰어남을 확인할 수 있었다.

<표 7> McNemar 검정결과

	연관성규칙 (A)	빈도행렬 (B)	규칙유도기법 (C)
ASFMRI (A+B+C)	8.481 (0.004)**	44.180 (0.000)**	2.813 (0.094)*
연관성규칙		24.038 (0.000)**	0.291[1] (0.590)[2]
빈도행렬			10.223 (0.001)**

1. McNemar 통계량 값 / 2. p-값
*: 유의수준 10%에서 통계적으로 유의함.
**: 유의수준 5%에서 통계적으로 유의함.

6. 결론

본 연구에서는 기존의 데이터마이닝 기법들이 가지고 있는 한계점들을 최소화하기 위하여, 이질적인 단일모형들을 지식 결합을 이용하여 시너지 효과를 생산할 수 있는 통합모형을 제시하였다. 따라서 본 연구에서는 보다 효과적인 고객구매예측을 위하여, 연관성규칙, 빈도행렬, 규칙유도기법의 단일모형을 규칙기반(rule-based)에 의한 지식이 축적된 통합모형을 제시하였다. 통합모형은 연관성규칙(A)과 빈도행렬(B)을 결합한 ASFM모형, 연관성규칙(A)과 규칙유도기법(C)을 결합한 ASRI모형, 빈도행렬(B)과 규칙유도기법(C)을 결합한 FMRI모형, 마지막으로 3가지 모형을 모두 통합한 ASFMRI모형(A+B+C)으로 구성되어 있다.

실험결과, 본 연구에서 제안한 통합모형인 ASFMRI모형(75.35%)이 전체 방법론과 비교하여 가장 예측력이 높았으며 데이터 셋 별로도 ASFMRI모형이 해당 각 데이터 셋에서 예측력이 가장 높다는 것을 알 수 있었다. 또한 2가지 단일모형을 통합한 ASRI모형(73.45%)과 FMRI모형(69.15%) 또한 다른 단일모형보다 보다 예측력이 더 뛰어났으나 3가지 모형을 통합한 ASFMRI모형(75.35%) 보다는 예측력이 낮다는 것을 알 수 있었다. 또한 단일모형 중에서 빈도행렬기법은 가장 예측력이 낮은 모형(51.7%)이며 통합에 있어서도 별다른 큰 영향력을 발휘하지 못하였다. 즉, 빈도행렬과 통합한 ASFM모형(65.35%)은 단일모형인 연관성규칙(64.0%) 보다는 예측력이 뛰어났으나 규칙유도기법(67.6%)에 비해서는 예측력이 떨어지는 것을 알 수 있었다. 이에 대한 원인으로 다음 2가지를 들 수 있다. 첫 번째 원인은 빈도행렬은 2차원의 행렬형태로 구성되어 있어 두 가지의 물품 $A \rightarrow B$ 형태의 규칙만 생성되기 때문에 다양한 형태의 규칙이 생성되지 못하는 제약(constraint)이 있다. 두 번째 원인으로는 대부분 빈도행렬에서 추출한 규칙 또는 규칙집합이 연관성규칙이나 규칙유도기법에서 추출한 규칙 또는 규칙집합에 대부분 반영되어 있다는 것을 규칙분석에서 확인할 수 있었다.

실제 응용분석에 있어서 단순히 특정 한가지의 알고리즘만을 사용하겠다고 미리 가정하고 분석하는 것, 즉, 한가지 알고리즘만을 선택하여 모형을 적용시킬 것이 아니라, 적용 가능한 알고리즘 전반을 선택한 다음, 각 알고리즘에서 추출한 규칙을 누적된 규칙집합의 형태로 통합한 통합모형을 구축하는 것이 오차항을 줄이고 설명력을 높일 수 있는 안정적인 결과로 도출될 수 있다는 사실을 실험결과를 통해 알 수 있었다. 이러한 통합모형 아키텍처가 바로 앙상블 접근법(ensemble approaches)이다. 앙상블 접근법은 한 명의 전문가보다 여러 명의 전문가가 한 예측을 종합했을 때 더 나은 결과를 가져올 수 있다는 점에 기초한다. 특히 여러 명의 전문가가 서로 독립적으로 예측결과를 잘못 만들어낼 때에는 그것을 종합했을 때 더욱 정확한 예측을 할 수 있다. 즉, 어느 단일모형에서 만들어진 규칙에 의해 분류되느냐에 따라 서로 다른 클래스(비구매, 구매)로 분류될 수 있지만 그것이 누적되면 결국은 실제 분류되어야 할 값과 같은 값으로 분류될 확률을 높일 수 있다는 것이다. 또한 통합모형은 규칙집합 정보만 필요하기 때문에 규칙이 추출되어 집합의 형태로 된 이후에는 원천 데이터 셋이 더 이상 필요하지 않으므로 저장할 필요가 없다. 이는 통계적으로 고정된 데이터가 아닌 시간의 흐름에 따라 무한히 추가되고 그 성격이 변하는 스트림 데이터에 적용할 경우 저장공간, 메모리, 시간 등의 자원 소모를 크게 줄일 수 있다는 장점이 있다. 요약하면, 통합모형을 이루고 있는 누적된 규칙집합은 데이터 셋 대신에 과거 정보를 유지하는 수단으로 저장되는데 그 크기가 현저하게 작으므로 저장공간의 소모가 작고 하나의 데이터 셋을 이용하여 단일모형에서 얻은 규칙집합이 누적되면서 단 하나의 규칙집합 또는 하나의 단일모형으로 예측하는 것보다 예측력에서 좀더 뛰어난 성능을 보여주고 있다.

그러나 통합모형 알고리즘 상에서도 몇 가지 문제가 발생할 수 있다. 첫째, 누적된 규칙집합의 결과를

합한 값(∑Rule Set)이 양의 값 또는 음의 값을 가지지 않고 0의 값을 가지는 문제이다. 이것은 단일모형에서 추출한 규칙집합이 해당 레코드에 모두 규칙적용이 불가능할 경우와 동일한 수로 구매와 비구매를 예측했을 경우 나타난다. 비록 본 연구에서는 10개의 검증용 데이터 셋 중에서 이와 같은 결과가 약 1.7%(34/2000)로 발생하였으나 데이터의 수가 적거나 누적된 규칙집합이 적을 경우 문제가 발생할 수 있다. 둘째, 단일모형 중 규칙유도기법에서는 규칙도출 시 분리기준과 정지 규칙을 최적으로 설정하는 문제를 한계로 들 수 있다.

마지막으로 본 연구가 가지고 있는 한계점을 정리하면서 앞으로의 관련 연구의 방향을 제시하고자 한다. 첫째, 본 연구에서 사용된 데이터의 유형은 행동특성 데이터(Behavioral Data)이다. 행동특성 데이터는 시간에 따라 빨리 변화하며, 데이터의 구조도 쉽게 변경되고 갱신될 수 있다는 단점이 있다. 따라서 과거의 유용한 행동특성들이 버려지지 않도록 데이터 웨어하우스와 같은 별개의 저장소에 시간에 따라 요약정보의 형태로 보관되어야 한다. 이들 데이터를 가지고 데이터마이닝을 이용한 구매예측 통합모형을 구축한다면 예측력 향상과 더불어 좀더 의미 있는 결과를 가져올 수 있을 것이다. 둘째, 향후 연구에서는 누적된 규칙집합 중에서도 상대적으로 중요한 규칙집합과 중요하지 않은 규칙집합을 구분하여, 중요도가 높은 규칙집합만으로 예측모형을 구성한 경우와 전체 규칙집합을 유지하면서 중요도 높은 규칙집합과 중요하지 않은 규칙집합에 가중치를 달리 적용하는 실험을 통하여 예측정확도를 비교하는 연구가 필요하다. 위 연구는 통합모형의 한계점인 동일한 수의 구매·비구매 예측결과의 한계점을 극복할 수 있을 것이다. 셋째, 누적된 규칙집합을 이용한 통합모형은 스트림 데이터 적용 분야에 가장 유용할 것이다. 따라서 시간의 흐름에 따라 변화가 심한 주식 시장 자료에 적용하는 연구가 필요하다고 본다. 넷째, 규칙집합을 만들 수 있는 여타 인공지능 기법을 이용한 좀더 향상된 통합모형에 관한 연구가 필요할 것이다.

참고문헌

[1] 김경재, 한인구, "퍼지신경망을 이용한 기업부도예측", *한국지능정보시스템학회*, 7권 1호(2001), pp. 135-147.

[2] 김진화, 남기찬, 변현수, "웹 방문 패턴 시각화 및 상품추천 방법에 관한 연구", *한국경영정보학회 추계 학술대회 발표논문집*, 2004, pp. 47-55.

[3] Abdelwashed, T., and Amir E.M., "New Evolutionary Bankruptcy Forecasting Model Based on Genetic Algorithms and Neural Networks," *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, Washington, USA, 2005, pp. 241-245.

[4] Agrawal, R., Imielinski, T., and Swami, A., "Mining association rules between sets of items in large databases," *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207-216.

[5] Altman, E.I., Marco, G. and Varetto, F., "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," *Journal of Banking and Finance*, Vol.18, No.3, 1994, pp. 505-529.

[6] Anandarajan, M., Lee, P., and Anandarajan, A., "Bankruptcy Prediction of Financially Stressed Firms: An Examination of the Predictive Accuracy of Artificially Neural Networks," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.10, No.2, 2001, pp. 69-81.

[7] Atiya, A., "Bankruptcy prediction for credit risk using neural networks: a survey and new results," *IEEE Transactions on Neural Networks*, Vol.12, No.4, 2001, pp. 929-935.

[8] Boritz, J., and Kennedy, D., "Effectiveness of neural networks types for prediction of business failure," *Expert Systems with Applications*, Vol.9, No.4, 1995, pp. 503-512.

[9] Charalambous, C., Chartious, A., and Kaourou, F., "Comparative Analysis of Artificial Neural Network Models: Application in Bankruptcy Prediction," *Annals of Operations Research*, Vol.99, 2000, pp. 403-425.

[10] Chen, M.C., and Huang, S.H., "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques," *Expert Systems with Applications*, Vol.24, No.4, 2003, pp. 433-441.

[11] Grice, S.J., and Dugan, T.M., "The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher," *Review of Quantitative Finance and Accounting*, Vol.17, 2001, pp. 151-166.

[12] Lee, K.C., Kim, M.J., and Kim, H., "An Inductive Learning – Assisted Neural Network Approach to Bankruptcy Prediction: Comparison with MDA, Inductive Learning, and Neural Network Models," *Journal of Management Research*, Vol.23, No.3, 1994, pp. 109-114.

[13] Lee, K.C., Han, I.G, and Kim, M.J., "A Study on the Credit Evaluation Model Integrating Statistical Model and Artificial Intelligence Model," *Journal of Management Science*, Vol.21, No.1, 1996, pp. 81-100.

[14] Ohlson, J.A., "Financial ratios and probabilistic prediction of bankruptcy," *Journal of Accounting Research*, Vol.18, No.1, 1980, pp. 109-131.

[15] Pendharkar, P.C., "A Threshold-varying Artificial Neural Network Approach for Classification and its Application to Bankruptcy Prediction Problem," *Computers and Operations Research*, Vol.32, No.10, 2005, pp. 2561-2582.

[16] Shin, K.S., and Lee, K.J., "Bankruptcy Prediction Modeling Using Multiple Neural Network Models," *Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference(KES'2004)*, Wellington, New Zealand, 2004a.

[17] Shin, K.S., and Lee, K.J., "Neuro-Genetic Approach for Bankruptcy Prediction Modeling," *Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference(KES'2004)*, Wellington, New Zealand, 2004b.

[18] Song, H.S., Kim, S.H., and Kim, J.K., "A Methodology for Detecting the Change of Customer Behavior based on Association Rule Mining," *Proceedings of the PACIS 2001*, Seoul, Korea, 2001, pp. 871-885.

[19] Wilson, R., and Sharda, R., "Bankruptcy prediction using neural networks," *Decision Support Systems*, Vol.11, No.5, 1994, pp. 545-557.

[20] Zhang, G., Hu, Y.M., Patuwo, E.B., and Indro, C.D., "Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis," *European Journal of Operational Research*, Vol.116, 1999, pp. 16-32.