퍼지 및 지능시스템 학회 2003, Vol. 13, No. 5, pp. 530-538

신경망과 연관규칙을 이용한 구매패턴 분류시스템의 구현

Implementation of Purchasing Pattern Classification System Using Neural Network and Association Rules

이종민 · 정 홍 · 김진상

Jong Min Lee, Hong Chung, Jin Sang Kim

계명대학교 정보통신대학 컴퓨터공학과

요 약

최근 마케팅 업계의 동향을 보면 기존 고객 유지에 대한 필요성을 중요시하면서, 타깃 마케팅의 개념에 의한 고객집단의 세분화된 분류와 각각의 세분화된 고객집단에 대한 차별적인 대응이 요구되고 있다. 본 논문에서는 신경망과 연관규칙의 Cumulate 알고리즘을 이용하여 고객집단을 분류하고 고객집단간의 구매패턴을 분류하는 시스템을 구현하였다. 실제 특정 두 집단간의 연관규칙을 조사한 결과 서로 간에 비슷한 연관규칙이 있음을 알 수 있었고, 마케팅 의사결정을 위해 우량/일반 고객집단으로 분류해야 할 필요성이 있음을 밝혔다. 따라서 고객집단의 분류에 있어 예측율의 정확성을 높임으로써 차별적인 마케팅의 효율을 극대화 할 수 있음을 보였다.

Abstract

Recently, the needs for keeping existing customers is increasing in the field of marketing. So, the customers needs to be classified by groups and the differentiated responses to the specified customer groups are demanded. In this paper, we implemented a system that classifies the customer groups using the neural network, and classified the purchasing patterns among customer groups. Empirically examining the association rules between two groups, we could find out that similar rules exist between them. So, it is important that customers should be classified into the excellent customer group and the general group for the decision making of marketing. This paper shows that the efficiency of the differentiated marketing can be maximized by raising the correctness of the expectation in the classification of customer groups.

Key Words : purchasing pattern classification, neural network, association rule

1. 서 론

최근 마케팅 업계의 동향을 보면 새로운 고객의 유치보다 기존 고객의 유지에 대한 필요성을 중요시하고 있다. 따라서 타깃 마케팅(target marketing)의 개념에 의한 고객집단의 세분화된 분류와 각각의 세분화된 고객집단에 대한 차별적인 대응이 요구되고 있다[3]. 이를 위해서는 특정 기준에 따라 고객집단을 분류해야 하고, 상품 판매 트랜잭션(transaction)의 상관관계를 찾는 상품 구매 패턴을 알아야 한다.

고객 집단의 분류를 위해서는 신경망(neural network) 모형이 가장 적합한데, 이는 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내 내어 자신이 가진 데이터로부터 반복적인 학습 과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써 향후를 예측하고자 하는 문제에 유용할 수 있기 때문이다[1,9,11]. 이러한 신경망 모형은 고객의 신용평가, 불량거래의 색출, 의료진단예측, 우량고객의 선정, 타깃 마케팅 등 여러 분야에 응용될 수가 있는데, 주로 교사학습(supervised learning)에 적용되어 목적변수에 대한 예측이나 분류를 목적으로 감춰진 패턴을 찾고 이를 일반화하는데 이용된다[4,9].

접수일자 : 2003년 2월 14일
완료일자 : 2003년 4월 25일

고객의 구매패턴 분류를 위한 연관규칙(association rule)에 대한 연구는 IBM Almaden 연구소에서 슈퍼마켓 데이터를 분석하여 고객들이 특정 상품을 구입했을 때 동시에 어떤 상품을 구입하는 경향이 있는가를 분석하기 위해 시작되었다[6]. 일반적으로 데이터에서 숨겨진 패턴을 탐사하는 연구 중에서 연관규칙 탐사분야에 많은 연구가 이루어지고 있는데, 그 이유는 연관규칙이 항목들의 집합들로 이루어진 데이터베이스에서 항목들 간의 존재 의존성을 반영하는 지식이기 때문이다[7]. 연관규칙 탐사는 주어진 데이터베이스에서 사용자가 미리 정의한 최소 신뢰도와 최소 지지도를 넘는 모든 연관규칙을 찾아내는 것이다.

본 논문에서는 신경망을 이용해 고객집단을 분류하고 분류된 고객의 특성에 따라 세분화된 고객들에 대해 일반화된 연관규칙(generalized association rule)을 적용해서 고객의 상품 구매 패턴을 찾아줌으로써 마케팅 전략 결정을 지원하는 구매 패턴 분류 시스템을 구현한다. 시스템 구현 환경은 Visual C++ 6.0과 SQL 6.0, MS Access 2000을 사용한다.

2. 연관규칙의 탐사

임의의 단위에 발생한 사건들의 묶음을 트랜잭션이라 하고, 대용량의 트랜잭션들이 데이터베이스에 누적된 환경에서

사건 부류 혹은 트랜잭션간의 상호 관계를 발견하는 작업을 데이터베이스상의 연관규칙 탐사로 정의할 수 있다[2,6,7]. 연관규칙 탐사는 관계형 데이터베이스에서 함수적 종속성(functional dependency)을 추출하는 기법과는 달리 통계적 방법에 의해 연관성이 있는 항목들 사이의 규칙성을 추출하는 과정이다[13].

연관규칙 탐사에는 사용자에 의해 주어지는 임계치인 최소 지지도(minimum support : minsup)와 최소 신뢰도(minimum confidence : minconf)가 적용된다[7]. 지지도는 특정 항목 집합의 통계적 중요성을 나타내는 수치 값으로, 예를 들면, '전체 트랜잭션에 대해 커피와 설탕을 함께 구매한 트랜잭션 수의 비율'로 정의된다. 신뢰도는 연관규칙의 강도를 나타내는 척도이다. 예를 들면, '커피와 크림을 구매한 고객들 중에 설탕을 함께 구매한 트랜잭션의 비율'로 측정된다.

X를 규칙의 가정, Y를 규칙의 결과라 할 때, 가정과 결과로 이루어지는 연관규칙을 갖는 항목들의 집합을 항목집합(itemset)이라 한다. 항목집합에 있는 항목들의 수를 항목집합의 길이라 하고 k 길이를 가지는 항목집합은 k-항목집합이라 한다[6,7]. N개의 트랜잭션 T로 구성된 집합을 D라 표기하면 연관규칙은 다음과 같은 정의를 가진다.

기본 가정

- 항목집합 I의 부분집합 X에 대해, X⊆T이면 T는 X를 만족한다고 정의한다.
- 항목집합 X(X⊆I)를 만족시키는 D의 트랜잭션 수를 |X|로 표기한다.
- X, Y⊆I에 대한 연관규칙 X→Y는 X∩Y=∅의 특성을 갖는다.

용어 설명

- 지지도 : 연관규칙 X→Y는 지지도 S를 갖는다.

$$S = \frac{|X \cup Y|}{N}$$

- 신뢰도 : 연관규칙 X→Y는 신뢰도 C를 갖는다.

$$C = \frac{|X \cup Y|}{|X|}$$

일반적으로 연관규칙 탐사는 두 단계로 나누어질 수 있다. 첫 번째는 데이터베이스에서 추출될 수 있는 모든 항목집합 중 최소 지지도보다 높은 지지도를 갖는 모든 항목집합을 찾는 단계로서 추출된 항목집합은 빈발 항목집합(large itemset)이라고 한다. 두 번째 단계는 첫 단계에서 얻는 빈발 항목집합을 이용하고 데이터베이스를 참조하여 연관규칙을 찾는 단계이다.

2.1 빈발 항목집합 찾기

X ⊆ T일 때 트랜잭션 T는 X를 포함한다고 말하고 T가 집합 X를 지지한다고 한다. X의 지지도를 supp(X)로 정의하면 이것은 X를 지지하는 T에 있는 모든 트랜잭션 수를 의미한다. 만약 주어진 최소 지지도 minsup에 대하여 supp(X) ≥ minsup 이라면 집합 X는 빈발하다고 한다.

다음은 빈발 항목집합을 효과적으로 찾는 과정에서 얻어진 특성들이다[14].

특성 1(부분집합 지지도) : 항목 집합 A, B에 대해 A ⊆ B 라면 B를 지지하는 D의 모든 트랜잭션들이 A를 지지하므로 supp(A) ≥ supp(B)이다.

특성 2(빈발하지 않은 집합의 상위집합은 빈발하지 않다) : 항목집합 A가 D에서 최소 지지도 보다 낮다면, 특성 1에 의해 supp(B) ≤ supp(A) ≤ minsup 이므로 A의 모든 상위집합은 빈발하지 않다.

특성 3(빈발 항목집합의 부분집합은 빈발하다) : 항목집합 B가 D에서 빈발하다면 특성 1에 따라 supp(A) ≥ supp(B) ≥ minsup 이므로 B의 모든 부분집합 A는 D에서 또한 빈 발할 것이다. 특히 A={i1, i2,…, ik }가 빈발하면 이것의 모든 k개의 (k-1)-부분집합들도 빈발하다. 그 역은 성립하지 않는다.

일반적으로 빈발 항목을 찾는 알고리즘은 데이터에 대해 여러 단계를 가진다. 첫 번째 단계에서는 각 항목의 지지도를 계산하고 최소 지지도 이상의 빈발 항목을 결정한다. 다음 단계에서는 앞에서 찾은 빈발 항목을 시드 집합(seed set)으로 후보 항목집합들로 불리는 새로운 잠재적 빈발 항목집합들을 생성하는데 사용하게 된다. 그리고 이 후보 항목집합들이 빈발한지 실제로 지지도를 계산한다. 마지막으로 후보 항목집합 중에서 빈발 항목집합을 결정하고 이것은 다음 단계에 시드 집합으로 사용한다. 이러한 프로세스는 더 이상 새로운 빈발 항목집합이 생성되지 않을 때까지 지속된다.

연관규칙 탐사에서 문제점인 많은 수의 후보집합의 생성은 데이터베이스와의 많은 비교연산을 초래하는데, 후보 항목집합의 수를 줄이는데 성공적인 Apriori-gen이라는 새로운 후보 항목집합 생성 전략이 만들어지면서 대부분의 알고리즘에서 사용하게 되었다.

2.2 Apriori 알고리즘

Apriori 알고리즘[6,7]은 그림 1과 같이 첫 번째 단계에서는 1-항목집합을 결정하기 위해 단순히 항목이 발생한 수를 계산한다. 다음 k-단계는 두개의 단계로 이루어진다. 첫 번째는 (k-1)-단계에서 발견된 빈발 항목집합 Lk-1을 apriori-gen 함수를 이용해서 후보 항목집합 Ck를 생성한다. 다음으로 데이터베이스를 읽어 Ck에 후보의 지지도가 계산된다.

```
L1 = 빈발 1-항목집합들;
For ( k = 2 ; Lk-1 ; k++ ) do
begin
   Ck = apriori-gen 함수에 의해 Lk-1의 새로운 후보
        항목집합 생성;
   forall 데이터베이스에 있는 모든 트랜잭션 t do
   begin
        Ct = subset (Ck , t) ;
        t에 포함된 Ck의 모든 후보항목들의 count 증가;
        end
        Lk = 최소지지도를 가진 Ck의 모든 후보항목;
end
Answer = ∪k Lk ;
```

그림 1. Apriori 알고리즘
Fig. 1. Apriori Algorithm

Aprioir-gen 함수는 모든 빈발 (k-1)-항목집합들의 집합인 Lk-1을 인자로 가지고 모든 빈발 k-항목집합의 상위집합을 돌려준다. 이 함수는 첫 번째로 조인(join) 단계에서 Lk-1과 Lk-1을 조인하게 된다.

insert into Ck
select p.item1, p.item2, ··· , p.itemk-1, q.itemk-1 **from** Lk-1 p, Lk-1 q
where p.item1 = q.item1, ··· , p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1;

다음으로 Lk-1에 없는 c의 (k-1)-부분집합인 모든 항목집합 c∈Ck를 잘라낸다.

forall itemsets c∈Ck **do**
  forall (k-1)-subsets s of c **do**
    if (s∉Lk-1) **then**
      delete c from Ck ;

그림 2는 Apriori에서 빈발 항목집합을 찾는 과정을 그림으로 설명한다.

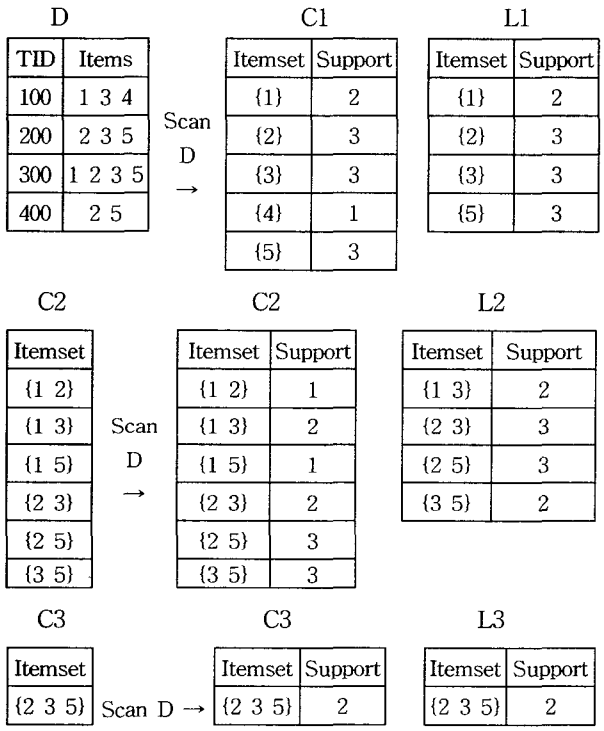

그림 2. Apriori에서 빈발 항목집합 추출의 예
Fig. 2. Example of Extracting Large-Itemsets in Apriori

각 단계에서 Apriori는 빈발 항목집합의 후보집합을 구축하고, 각 후보 항목집합의 발생 횟수를 계산하고, 사전에 정한 최소 지지도를 기초로 빈발 항목집합을 결정한다. 첫 번째 단계에서 Apriori는 각 항목의 발생 횟수를 세기 위해 단순히 모든 트랜잭션을 스캔한다. 얻어진 1-후보 항목집합의 집합 C1로부터 요구된 최소 트랜잭션 지지도가 2라 가정하면 요구된 최소 지지도를 갖는 1-후보 항목집합으로 구성된 1-빈발 항목집합 L1이 결정된다. 빈발 항목집합의 부분집합은 또한 최소 지지도를 가져야 한다는 사실의 관점에서 2-빈발 항목집합을 발견하기 위해 Apriori는 항목집합 C2의 후보집합을 생성하기 위해 L1*L1을 이용한다. 여기서 *는 연결연산자이다. C2는 $\left(\frac{|L1|}{2}\right)$개의 2-항목집합으로 이루어진다. 다음에 D에 있는 4개의 트랜잭션을 조사하고 C2에 있는 각 후보 항목집합의 지지도를 계산한다. 그림 2의 두 번째 행 중간 테이블은 C2에서 계산한 결과를 보여준다. 2-빈발 항목집합 L2는 C2에 있는 각 2-후보 항목집합의 지지도를 기초로 하여 결정된다. 후보 항목집합 C3는 다음과 같이 L2로부터 생성된다. L2로부터 동일한 첫 항목을 갖는 두 개의 2-빈발 항목집합 {2 3}과 {2 5}가 먼저 확인된다. 다음에 Apriori는 이들의 두 번째 항목으로 구성한 2-항목집합 {3 5}가 2-빈발 항목집합의 구성 요소인가를 검사한다. {3 5}는 그 자체가 빈발 항목집합이므로 {2 3 5}의 모든 부분집합은 빈발하고 {2 3 5}는 후보 3-항목집합이 된다. L2에서 더 이상의 3-후보 항목집합은 없다. Apriori는 모든 트랜잭션을 조사하고 3-빈발 항목집합 L3을 발견한다. L3으로부터 구성되는 4-후보 항목집합은 더 이상 존재하지 않는다. Apriori는 빈발 항목집합 발견 과정을 종료하게 된다.

2.3 일반화된 연관규칙(Generalized Association Rule)

많은 응용 분야에서 데이터 항목 사이의 흥미로운 연관은 상대적으로 상위 수준의 개념에서 발생한다[10]. 예를 들어 거래 데이터베이스에서 구매패턴은 바코드 수준과 같은 원시적인 데이터 수준에서는 흥미있는 규칙성을 보이지 않지만 빵과 우유와 같은 상위개념 수준에서는 어떤 흥미있는 규칙성을 보이는 경우가 많다.

주어진 트랜잭션 데이터베이스에서 각 트랜잭션이 항목들의 집합, 항목간의 분류(taxonomy)로 이루어져 있을 때 어떤 수준의 분류에서 항목간의 관계를 발견할 수 있다. 대부분의 경우 데이터는 계층적으로 분류되고 분류계층을 고려하여 유용한 정보를 추출할 수 있다. 이때 각 항목을 일반화시키기 위해 그림 3의 예와 같은 개념계층(concept hierarchy)을 이용한다[16].

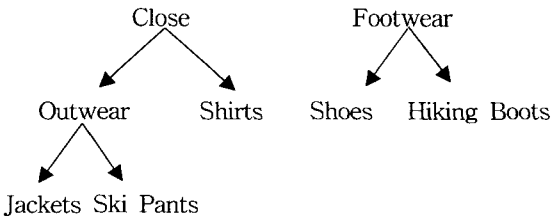

그림 3. 개념계층의 예
Fig. 3. Examples of Concept hierarchy

개념계층은 저수준의 데이터베이스를 고수준의 데이터베이스로 일반화하기 위해 사용되며, 데이터베이스의 속성에 있어서 일반화 관계의 집합이다. 일반화 관계는 속성값의 전체집합과 고수준으로 일반화된 단일값 간의 관계이다. 일반화 관계는 {a1,a2,...,ak}⊂A로 표현되는데, A는 각 ai(1≤i≤k)의 일반화이다[8].

2.4 Cumulate 알고리즘

Cumulate 알고리즘은 전처리로서 데이터베이스의 각 트랜잭션에 포함된 항목의 모든 상위 항목을 추가하고, 데이터베이스를 재편성한 다음 Apriori 알고리즘을 적용하여 연관규칙을 추출하는 기본 알고리즘에 다음과 같은 몇 가지 최적

화를 추가한 알고리즘이다[16].

현 단계에서 계산된 후보 항목들의 상위 항목을 추가하고 포함되지 않은 항목은 트랜잭션에서 제거한다. 개념 계층을 순회하여 각 항목의 상위항목을 찾지 않고 상위항목을 미리 계산하면서 동시에 후보 항목들에 없는 상위 항목들은 제거한다. 항목과 상위항목을 모두 포함하는 항목집합은 가지치기한다. 그림 4는 전체적인 Cumulate 알고리즘이다.

```
데이터베이스 T*의 각 항목들의 상위 항목집합 T를 계산;
L1 = 빈발 1-항목집합들;
k = 2;
while (Lk-1 ≠ ∅ ) do
begin
    Ck = Lk-1로부터 생성된 크기가 k인 새로운 후보 항목;
    if (k=2) then
        C2내에 항목과 상위항목으로 이루어진 후보 항목 전지;
        T에서 Ck에 존재하지 않는 항목의 모든 상위 항목 전지;
    forall 트랜잭션 t ∈ D do
    begin
    foreach 항목 x ∈ t do
            T에 있는 모든 상위항목을 t에 추가;
            t에서 중복 제거;
            t에 있는 Ck의 모든 후보항목의 count 증가;
    end
    Lk = 최소 지지도를 가지는 Ck의 모든 후보항목;
    k = k + 1;
end
```

그림 4. Cumulate 알고리즘
Fig. 4. Cumulate Algorithm

3. 시스템의 구현

3.1 시스템 구성도

그림 5는 본 연구에서 구현하고자 하는 구매패턴 분류시스템의 개략적인 구성도이다.

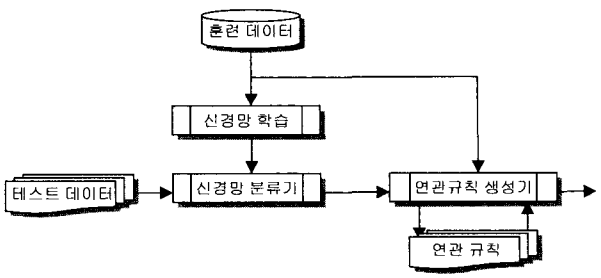

그림 5. 시스템 구성도
Fig. 5. System Diagram

데이터베이스로부터 고객에 대한 정보와 구매 물품의 목록을 입력을 받아 다층 퍼셉트론의 역전파 알고리즘을 이용해 신경망 학습을 하고, 연관규칙 생성기로부터 미리 주어진 개념계층과 Cumulate 알고리즘을 이용하여 일반화된 연관규칙을 생성한다. 그리고 테스트 데이터로부터 신경망 분류기를 통해 고객들의 구매패턴을 분류하기 위한 목표 고객집단을 분류하고 분류된 각각의 고객집단에 대해 구매패턴을 예측한다.

본 연구에서는 훈련 데이터와 테스트 데이터에 대한 비율은 2:1로 설정했으며 데이터는 정제되어 있다고 가정한다.

3.2 고객집단 분류

3.2.1 데이터의 준비

고객집단의 분류를 위해 신경망에 사용하기 위해 먼저 모 전자대리점 판매 데이터베이스를 변화시킨다. 신경망 입력에 불필요한 요소인 주소, 이름 등을 제거하고 선택된 요소는 표 1과 같다.

표 1. 신경망에 사용하기 위해 선택된 요소
Table 1. Selected Factors for Neural Network

선택된 요소	설 명
구매금액	4가지로 분류 : 200만원이하, 200만원~250만원, 250만원~300만원, 300만원이상
구매횟수	3가지로 분류 : 1회, 2회, 3회 이상
고객구분	2가지로 분류 : 임시, 고정
상권	13가지로 분류 : CB2, G01, G02, G05, G06, G07, G08, G09, G11, G12, G13, G14, none

표 1에서 선택된 요소의 구간에 따라 Field ID를 정하고 데이터베이스 입력을 위한 테이블을 설정한다.

3.2.2 신경망 분류기 모델링

그림 6은 고객집단의 분류를 위한 신경망 모델이다.

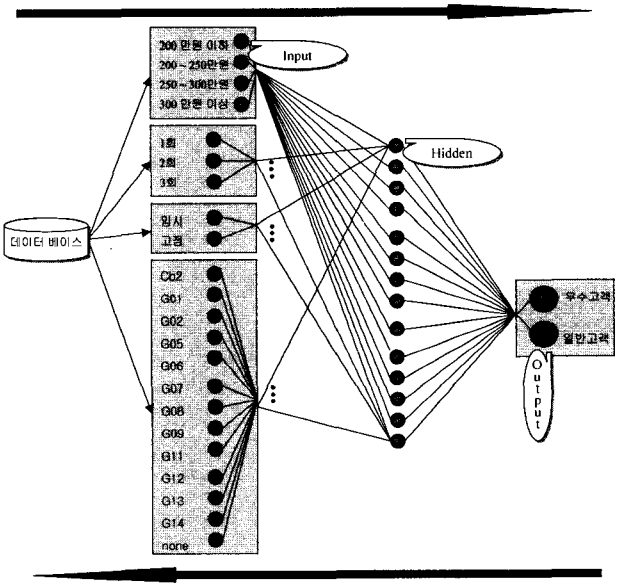

그림 6. 고객집단 분류를 위한 신경망 모델
Fig. 6. Neural Network Model for Customer Group Classification

입력층의 입력 노드 수는 22개이고 출력층의 처리 요소 수는 2개이다. 은닉층은 하나이며, 처리 요소 개수는 14개이다. 최소 $\log_2 N = 22$ 보다 많은 처리 요소를 가진 몇 가지 후보 모델 중에서 여러 가지 가능한 아키텍처로 모델링을 시도해 검출 효율과 검출 능력 측면에서 모델의 성능을 비교해 최종 예측 모형을 선택했다. 그리고 역전파 알고리즘을 구현하고 학습시키는 과정에서 지역 최소점에 빠지거나 진동하는 경우에는 학습 계수나 초기값(weight, offset), 은닉 노드 수 등을 조정했다.

연결 강도의 초기값은 -0.5~0.5 사이의 임의의 수를 주었으며, 진동을 적게 하기 위해서 모멘텀 항은 아주 작은 값인 0.1로 설정했다. 이것은 모멘텀 항을 0.2 이상으로 했을 때보다 오차가 크고 학습시간이 길어지는 경향이 있지만 진동을 적게 해 예측 적중률이 높아지는 결과를 얻었다. 이 과정에서 오차를 줄이기 위해 이득항을 0.5로 설정하였다.

3.3 고객패턴 분류

3.3.1 개념계층

구매 물품의 일반화를 위해 개념 계층을 구성한다. 그림 7은 본 연구에서 사용한 모 전자대리점의 데이터베이스로부터 추출한 개념계층의 일부이다.

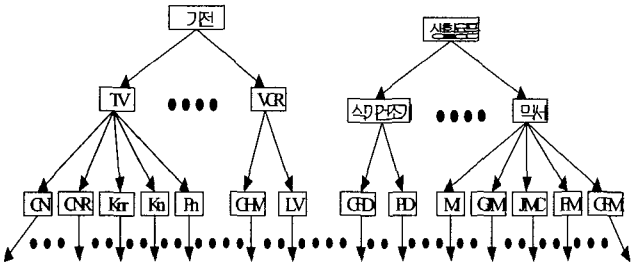

그림 7. 전자 제품의 개념계층
Fig. 7. Concept Hierarchy of Electronic Products

모든 물품 트랜잭션들은 모두 코드화 되어 있으며 최하위 계층의 트랜잭션 물품들은 CN-1430, CN-14F7 등으로 나타나므로 '-'이하를 제거시킴으로써 상위계층을 준비한다.

3.3.2 연관규칙 생성기

그림 8은 연관규칙 생성기의 구성도 이다.

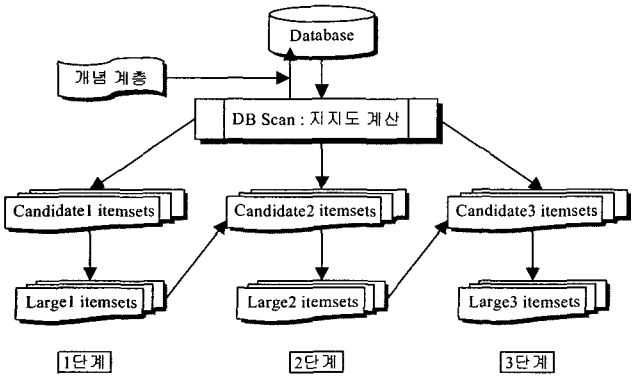

그림 8. 연관규칙 생성기의 구성도
Fig. 8. Diagram of Association Rules Generator

먼저 데이터베이스를 읽어 모든 항목들의 지지도를 계산하고 최소 지지도 20%를 넘지 못하는 항목들을 제외한 나머지 항목들을 개념 계층을 이용해 개념상승을 시도해서 상위 개념 항목들의 새로운 트랜잭션 데이터베이스를 만든다. 이때 하나의 레코드에서 항목을 개념 상승시킬 경우 중복이 발생하게 되는데 이를 제거한 후 데이터베이스를 생성하게 된다. 예를 들어 구매자가 CN, CNR의 물품을 구입했을 때 개념계층을 이용해 개념 상승을 시키면 두 개 모두 TV를 구매한 결과가 나타난다.

새로운 데이터베이스 생성 후 데이터베이스에 있는 모든 항목들로 1-후보 항목집합을 생성한다. 후보 항목집합으로부터 데이터베이스를 스캔해서 지지도를 구한 후 최소지지도 이상의 빈발 항목들로만 1-빈발 항목집합 테이블을 생성한다. 그림 9는 일반고객과 우량고객의 1-빈발 항목집합 생성 과정이다.

D

TID	Items
1	125
2	2
3	24578
4	4
5	
6	123456
7	1347
8	
9	:

Scan D →

C1

Itemset	Support
{1}	69
{2}	59
{3}	51
{4}	44
{5}	45
{6}	25
{7}	28
{8}	24
{9}	:

L1

Itemset	Support
{1}	69
{2}	59
{3}	51
{4}	44
{5}	45
{6}	25
{7}	28
{8}	24

(a) 일반 고객

D

TID	Items
1	125
2	2
3	24578
4	4
5	
6	123456
7	1347
8	
9	:

Scan D →

C1

Itemset	Support
{1}	68
{2}	72
{3}	60
{4}	43
{5}	47
{6}	37
{7}	39
{8}	29
{9}	:

L1

Itemset	Support
{1}	68
{2}	72
{3}	60
{4}	43
{5}	47
{6}	37
{7}	39
{8}	29

(b) 우수 고객

그림 9. 1. 빈발 항목집합 생성
Fig. 9. Generating 1-Itemsets

2단계에서는 1-빈발 항목집합 테이블의 항목들의 조합에 의해 모든 경우의 수로 2-후보 항목집합을 생성하고 데이터베이스를 스캔해서 지지도를 조사한 후 최소지지도 이상이 아닌 후보 항목집합을 제거함으로써 2-빈발 항목집합 테이블을 생성한다. 그림 10은 일반고객과 우량고객의 2-빈발 항목집합 생성 과정이다.

C2

Itemset	Itemset
{1 2}	{3 5}
{1 3}	{3 6}
{1 4}	{3 7}
{1 5}	{3 8}
{1 6}	{4 5}
{1 7}	{4 6}
{1 8}	{4 7}
{2 3}	{4 8}
{2 4}	{5 6}
{2 5}	{5 7}
{2 6}	{5 8}
{2 7}	{6 7}
{2 8}	{6 8}
{3 4}	{7 8}

Scan D →

C2

Itemset	Sup	Itemset	Sup
{1 2}	45	{3 5}	30
{1 3}	47	{3 6}	21
{1 4}	36	{3 7}	25
{1 5}	35	{3 8}	14
{1 6}	22	{4 5}	24
{1 7}	22	{4 6}	17
{1 8}	19	{4 7}	19
{2 3}	32	{4 8}	12
{2 4}	29	{5 6}	17
{2 5}	35	{5 7}	18
{2 6}	18	{5 8}	12
{2 7}	19	{6 7}	13
{2 8}	19	{6 8}	11
{3 4}	31	{7 8}	10

L2

Itemset	Sup	Itemset	Sup
{1 2}	45	{2 4}	29
{1 3}	47	{2 5}	35
{1 4}	36	{3 4}	31
{1 5}	35	{3 5}	30
{1 6}	22	{3 6}	21
{1 7}	22	{3 7}	25
{2 3}	32	{4 5}	24

(a) 일반 고객

C2

Itemset	Itemset
{1 2}	{3 5}
{1 3}	{3 6}
{1 4}	{3 7}
{1 5}	{3 8}
{1 6}	{4 5}
{1 7}	{4 6}
{1 8}	{4 7}
{2 3}	{4 8}
{2 4}	{5 6}
{2 5}	{5 7}
{2 6}	{5 8}
{2 7}	{6 7}
{2 8}	{6 8}
{3 4}	{7 8}

Sca n D →

C2

Itemset	Sup	Itemset	Sup
{1 2}	50	{3 5}	31
{1 3}	48	{3 6}	25
{1 4}	34	{3 7}	31
{1 5}	33	{3 8}	16
{1 6}	27	{4 5}	23
{1 7}	34	{4 6}	18
{1 8}	17	{4 7}	26
{2 3}	45	{4 8}	16
{2 4}	31	{5 6}	24
{2 5}	42	{5 7}	23
{2 6}	26	{5 8}	14
{2 7}	30	{6 7}	19
{2 8}	19	{6 8}	11
{3 4}	30	{7 8}	12

L2

Itemset	Sup	Itemset	Sup
{1 2}	50	{2 7}	30
{1 3}	48	{3 4}	30
{1 4}	34	{3 5}	31
{1 5}	33	{3 6}	25
{1 6}	27	{3 7}	31
{1 7}	34	{4 5}	23
{2 3}	45	{4 7}	26
{2 4}	31	{5 6}	24
{2 5}	42	{5 7}	23
{2 6}	26		

(b) 우수 고객

그림 10. 2-빈발 항목집합 생성
Fig. 10. Generating 2-Itemsets

3단계에서는 2-빈발 항목집합의 항목들의 조합에 의해 3 후보 항목집합을 생성한다. 이때 3-후보 항목집합의 부분집합들이 2-빈발 항목집합에 있는 지를 조사해서 없다면 후보 항목집합에서 제거하게 된다. 이것은 연관규칙의 특성 중에서 상위 항목집합이 빈발하다면 그 부분집합은 빈발하며 부분집합이 빈발하지 않을 경우 상위집합도 빈발하지 않다는 특성을 이용하는 것이다. 이렇게 함으로써 데이터베이스를 스캔할 때 비교 횟수를 줄일 수 있다. 3-후보 항목집합이 생성되면 데이터베이스를 스캔해서 최소 지지도 이상의 3-빈발 항목집합 테이블을 생성하게 된다. 그림 11은 일반고객과 우량고객의 3-빈발 항목집합 생성 과정이다.

C3

Itemset	Itemset
{1 2 3}	{1 3 7}
{1 2 4}	{1 4 5}
{1 2 5}	{2 3 4}
{1 3 4}	{2 4 5}
{1 3 5}	{2 3 5}
{1 3 6}	{3 4 5}

Scan D →

C3

Itemset	Sup	Itemset	Sup
{1 2 3}	30	{1 3 7}	22
{1 2 4}	25	{1 4 5}	21
{1 2 5}	28	{2 3 4}	19
{1 3 4}	29	{2 4 5}	19
{1 3 5}	28	{2 3 5}	22
{1 3 6}	20	{3 4 5}	18

L3

Itemset	Sup
{1 2 3}	30
{1 2 4}	25
{1 2 5}	28
{1 3 4}	29
{1 3 5}	28
{1 3 6}	20
{1 3 7}	22
{1 4 5}	21
{2 3 5}	22

(a) 일반 고객

C3

Itemset	Itemset
{1 2 3}	{2 3 5}
{1 2 4}	{2 3 6}
{1 2 5}	{2 3 7}
{1 2 6}	{2 4 5}
{1 2 7}	{2 4 7}
{1 3 4}	{2 5 6}
{1 3 5}	{2 5 7}
{1 3 6}	{3 4 5}
{1 3 7}	{3 5 7}
{1 4 5}	{3 5 6}
{2 3 4}	{3 5 7}
	{4 5 7}

Scan D →

C3

Itemset	Sup	Itemset	Sup
{1 2 3}	36	{2 3 5}	29
{1 2 4}	26	{2 3 6}	20
{1 2 5}	31	{2 3 7}	25
{1 2 6}	19	{2 4 5}	21
{1 2 7}	27	{2 4 7}	21
{1 3 4}	28	{2 5 6}	20
{1 3 5}	29	{2 5 7}	21
{1 3 6}	21	{3 4 5}	19
{1 3 7}	30	{3 5 7}	16
{1 4 5}	19	{3 5 6}	15
{2 3 4}	23	{3 5 7}	13
		{4 5 7}	10

L3

Itemset	Sup	Itemset	Sup
{1 2 3}	36	{2 3 4}	23
{1 2 4}	26	{2 3 5}	29
{1 2 5}	31	{2 3 6}	20
{1 2 7}	27	{2 3 7}	25
{1 3 4}	28	{2 4 5}	21
{1 3 5}	29	{2 4 7}	21
{1 3 6}	21	{2 5 6}	20
{1 3 7}	30	{2 5 7}	21

(b) 우수 고객

그림 11. 3-빈발 항목집합 생성
Fig. 11. Generating 3-Itemsets

마지막으로 빈발 항목집합 테이블을 통해 연관규칙을 도출한다. 흥미로운 연관규칙을 위해 사용되는 신뢰도는 다음과 같이 구해질 수 있다.

연관규칙 R : X → Y에서 지지도를 supp(X), supp(Y)로 나타내면 신뢰도 conf(R) = supp(X ∪ Y) / supp(X)의 조건부 확률이 된다. 이것은 X를 구입한 사람이 Y도 구입할 확률로써 결과적으로 얼마나 조건부에 대하여 결과부가 자주 적용되는가를 보여주는 것이다[12,15].

본 논문에서 적용한 최소 신뢰도는 80%이며 2, 3-빈발 항목집합 테이블로부터 항목집합의 신뢰도를 구하고 최소 신뢰도 80% 이상의 연관규칙을 생성하게 된다.

4. 실험 결과 및 평가

4.1 실험 결과

그림 12는 신경망 분류기의 학습 후 모 전자 대리점의 전자제품 200가지의 훈련 데이터를 통해 테스트해 본 결과이다. 일반 고객집단의 적중률은 68%이며 우량 고객집단의 적중률은 62%, 전체 적중률은 65%이다.

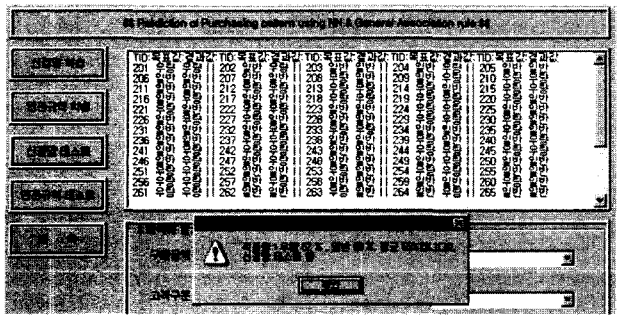

그림 12. 신경망 테스트 결과
Fig. 12. Testing Results of Neural Network

그림 13은 일반/우량 고객집단의 2-빈발 항목집합들로부터 연관규칙을 추출하기 위해 신뢰도를 구한 결과를 보여준다. 이 그림들은 빈발 항목 테이블로부터 엑셀을 이용하여 모든 신뢰도를 그래프로 출력하였다. 2-빈발 항목집합에서 항목 X, Y가 있을 때 연관규칙의 정의에 따라 X → Y, Y → X로의 연관규칙을 위한 2가지 신뢰도를 생성한다. 그림 13에서 일반1과 우량1은 항목집합에서 위쪽 방향으로의 규칙이며, 일반 2와 우량2는 아래쪽 방향으로의 규칙이다.

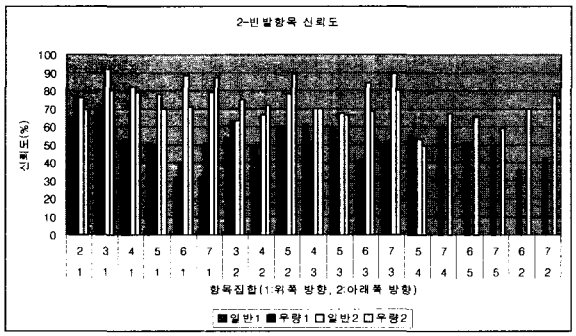

그림 13. 2-빈발항목 신뢰도
Fig. 13. Confidence of 2-Large Itemset

그림 14는 3-빈발 항목집합의 신뢰도를 보여준다. 3-빈발 항목집합에서 항목 X, Y, Z가 있을 때 연관규칙의 정의에 따라 {XY} → {Z}, {XZ} → {Y}, {YZ} → {X}로의 연관규칙을 위한 3가지 신뢰도를 생성한다. 그림 11에서 일반1과 우량1은 {XY} → {Z}로의 규칙, 일반2와 우량2는 {XZ} → {Y}로의 규칙이며, 마지막으로 일반3과 우량3은 {YZ} → {X}로의 규칙을 나타낸다.

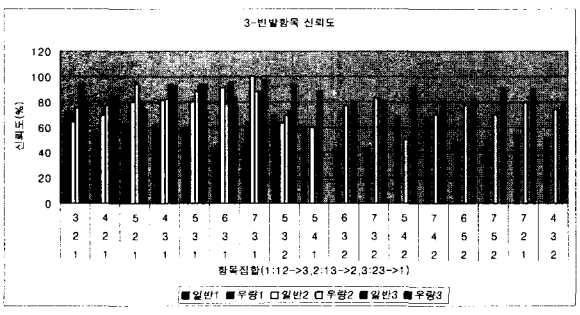

그림 14. 3-빈발항목 신뢰도
Fig. 14. Confidence of 3-Large Itemset

그림 14의 3-빈발 항목집합을 2-빈발 항목집합과 비교하면 일반 고객집단의 우량 고객집단에 대한 신뢰도 발생비율이 74%에서 47%로 떨어짐을 알 수 있다.

4.2 평가

학습 데이터 200건의 데이터를 이용하여 고객집단을 분류하고 일반 고객집단 100건과 우량 고객집단 100건의 데이터를 이용하여 그림 15, 16, 17과 같은 2, 3-항목집합의 연관규칙을 발견하였다. 연관규칙을 위한 최소 신뢰도는 80%이다.

일반 고객집단

{4(보온밥솥)}	→ {1(냉장고)}	: 82%
{6(전자레인지)}	→ {3(세탁기)}	: 84%
{6(전자레인지)}	→ {1(냉장고)}	: 88%
{7(가스렌지)}	→ {3(세탁기)}	: 89%
{3(세탁기)}	→ {1(냉장고)}	: 92%

우량 고객집단

{3(세탁기)}	→ {1(냉장고)}	: 80%
{7(가스렌지)}	→ {3(세탁기)}	: 80%
{7(가스렌지)}	→ {1(냉장고)}	: 87%
{5(VCR)}	→ {2(TV)}	: 89%

그림 15. 2-항목 연관규칙
Fig. 15. Association Rules of 2-Item

일반 고객집단

{1(냉장고) 5(VCR)}	→ {2(TV)}	: 80%
{1(냉장고) 5(VCR)}	→ {3(세탁기)}	: 80%
{2(TV) 5(VCR)}	→ {1(냉장고)}	: 80%
{1(냉장고) 4(보온밥솥)}	→ {3(세탁기)}	: 81%
{2(TV) 4(보온밥솥)}	→ {1(냉장고)}	: 86%
{3(세탁기) 7(가스렌지)}	→ {1(냉장고)}	: 88%
{4(보온밥솥) 5(VCR)}	→ {1(냉장고)}	: 88%

우량 고객집단

{2(TV) 3(세탁기)}	→ {1(냉장고)}	: 80%
{3(세탁기) 6(전자레인지)}	→ {2(TV)}	: 80%
{3(세탁기) 7(가스렌지)}	→ {2(TV)}	: 81%
{4(보온밥솥) 7(가스렌지)}	→ {2(TV)}	: 81%
{1(냉장고) 4(보온밥솥)}	→ {3(세탁기)}	: 82%
{1(냉장고) 7(가스렌지)}	→ {3(세탁기)}	: 83%
{2(TV) 7(가스렌지)}	→ {3(세탁기)}	: 83%
{5(VCR) 6(전자레인지)}	→ {2(TV)}	: 83%
{2(TV) 4(보온밥솥)}	→ {1(냉장고)}	: 84%
{3(세탁기) 6(전자레인지)}	→ {1(냉장고)}	: 84%
{1(냉장고) 5(VCR)}	→ {3(세탁기)}	: 88%

그림 16. 3-항목 연관규칙(80%)
Fig. 16. Association Rules of 3-Item(80%)

일반 고객집단

{1(냉장고) 6(전자레인지)}	→ {3(세탁기)}	: 91%
{3(세탁기) 5(VCR)}	→ {1(냉장고)}	: 93%
{2(TV) 3(세탁기)}	→ {1(냉장고)}	: 94%
{3(세탁기) 4(보온밥솥)}	→ {1(냉장고)}	: 94%
{3(세탁기) 6(전자레인지)}	→ {1(냉장고)}	: 95%
{1(냉장고) 7(가스렌지)}	→ {3(세탁기)}	: 100%

우량 고객집단		
{2(TV) 7(가스렌지)}	→ {1(냉장고)}	: 90%
{4(보온밥솥) 5(VCR)}	→ {2(TV)}	: 91%
{5(VCR) 7(가스렌지)}	→ {2(TV)}	: 91%
{2(TV) 4(보온밥솥)}	→ {1(냉장고)}	: 93%
{3(세탁기) 5(VCR)}	→ {2(TV)}	: 94%
{3(세탁기) 5(VCR)}	→ {1(냉장고)}	: 94%
{3(세탁기) 5(VCR)}	→ {2(TV)}	: 94%
{1(냉장고) 7(가스렌지)}	→ {3(세탁기)}	: 97%
{3(세탁기) 7(가스렌지)}	→ {1(냉장고)}	: 97%

그림 17. 3-항목 연관규칙(90%)
Fig.17. Association Rules of 3-Item(90%)

그림 18과 같이 각각의 고객집단에 따른 연관규칙이 비슷함을 보여주고 있다. 2-항목집합에서 연관규칙 {세탁기 → 냉장고}는 일반과 우량 고객집단간에 12%의 신뢰도 차이를 보이고, 연관규칙 {가스레인지 → 세탁기}는 우량 고객집단에서만 나타나는 연관규칙이다. 3-항목집합에서는 일반/우량 고객집단간의 연관규칙이 더욱 편차를 보임을 알 수 있다. 따라서 차별적 마케팅을 위한 고객 집단의 분류가 필요하며, 분류 예측율을 높일수록 마케팅의 효율을 높일 수 있다.

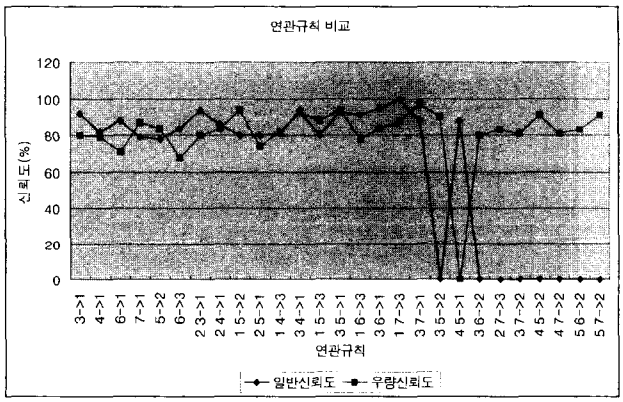

그림 18. 연관규칙 비교
Fig. 18. Comparison of Association Rules

최근 현대백화점에서 CRM(Customer Relationship-Management)의 일환으로 고객세분화작업(Selective Binding System)을 실시하여 2070의 법칙을 발견했다. 즉, 경기에 상관없이 상위 20%의 고객이 70%의 매출을 올린다는 것이다. 하지만 그렇다고 무조건 상위 20%의 고객을 대상으로 소위 귀족 마케팅을 실시하지는 않는다. 그 20%는 특별한 마케팅이 없어도 꾸준히 구매하는 사람들로 분석됐기 때문이다. 현재 가장 신경을 쓰는 고객집단은 최근에 구매를 한 고객들이며, 정밀분석을 해본 결과 이 고객집단이 구매 예상 시점에 있는 고객들보다 구매율이 더 높다는 사실이 드러났다. 자기가 산 물건에 대한 확인 때문에 광고에 가장 많이 관심을 갖고 또한 이것이 구매로 이어진다는 것이다. 또한 LG전자가 데이터베이스 마케팅을 도입, 고객정보를 전자 대리점 전산관리프로그램에 입력한 뒤 상품의 교체주기에 이른 고객에게만 집중적으로 광고물을 보내는 방식으로 마케팅을 한 결과 우편물 발송대상자 가운데 대리점을 찾는 사람이 40%에 달했고 이중 약 60%가 구매를 하는 놀라운 성과를 거두고 있다. 무작위로 발송할 때 잠재고객이 점포를 방문하는 비율이 2% 미만에 그친 것과 비교할 때 차이가 매우 크다는 것을 알 수 있었다[4,5].

본 논문의 시스템은 고객집단의 분류와 고객의 구매패턴을 분류하였다. 이에 위의 예와 같이 추가적인 데이터를 이용하여 분석한다면 보다 효율적인 마케팅이 될 수 있을 것이다.

5. 결론 및 향후 연구방향

본 연구에서 구현한 구매패턴 분류시스템은 상품 판매에 대한 의사 결정을 지원하기 위해 데이터의 분류에 있어 좋은 성능을 가진 신경망을 이용해 고객집단을 분류하고 상품 판매 트랜잭션의 상관관계를 찾는 일반화된 연관규칙을 연결하여 모 전자대리점의 고객의 상품 구매패턴을 예측해 보았다.

실제 어떤 두 집단간의 연관규칙을 조사한 결과 서로간의 비슷한 연관규칙이 있음을 알 수 있었고, 마케팅 의사결정을 위해 우량/일반 고객집단으로 분류해야 할 필요성이 있음을 밝혔다. 따라서 고객집단의 분류에 있어 예측율의 정확성을 높임으로써 차별적인 마케팅의 효율을 극대화 할 수 있음을 보였다.

본 시스템이 실제 업무에 잘 적용되기 위해서는 먼저 데이터베이스 구축에 있어서 고객과의 거래에 관련된 자료가 잘 정비되어 있어야 하며, 분류의 정확성을 높이기 위해 보다 많은 학습 데이터가 있어야 한다. 또한 신경망 시스템은 출력에 대한 분석이 어렵다는 단점이 있으므로 통계적 기법, 의사 결정 트리, 유전자 알고리즘 등의 데이터 마이닝 기법들을 통합하는 연구가 이루어져야 할 것이다. 그리고 일반화된 연관규칙을 탐사할 때 서로 다른 계층간의 연관관계와 순회패턴을 탐사할 수 있도록 연구되어져야 할 것이다.

앞으로 데이터베이스는 점점 대형화되어 질 것이고 이에 따라 데이터 마이닝에 의한 데이터베이스 마케팅은 그 필요성이 증가될 것이다. 따라서 향후에는 우선 어떤 항목이 필수적인 항목이고 어떤 항목이 불필요한 항목인지 선별하고 이를 효율적으로 구축하는 과정과 다음으로는 축적된 방대한 분량의 고객 데이터베이스를 정확하게 분석할 수 있는 적합한 데이터 마이닝 툴의 개발을 통해 마케팅 효율을 극대화해야 할 것이다.

참 고 문 헌

[1] 김대수, 신경망 이론과 응용(Ⅰ), 하이테크 정보, 1992.
[2] 박종수, 유원경, 홍기형, "연관규칙 탐사와 그 응용", 정보과학회지, pp37-43, 9월 1998.
[3] 박찬욱. 데이터베이스 마케팅, 연암사, 1996.
[4] 이동희, "Data mining을 이용한 리테일 뱅킹 전략에 관한 실증적 연구", 월간 금융('99.1), http://www.kfb.or.kr/
[5] 이상민, "DB마케팅", 금강기획 사보('98.3-4), http://www.diamond.co.kr/
[6] Rakesh Agrawal, Tomasz Imielnski, and Arun Swami. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93), pp. 207-216, May 1993.

[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB '94, pp.487-499, September 1994.
[8] D. Fudger and H. Hamilton, "A Heuristic for Evaluating Databases for Knowledge Discovery in Database: An Overview, in G. Piatesky-Shapiro and W. J. Frawly (eds.), Knowledge Discovery in Database, AAAI/MIT Press, 1-27, 1991.
[9] Joseph P. Bigus, Data Mining with Neural network, McGraw-Hill, 1996.
[10] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases", Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95), Zurich, Switzerland, pp.420-431, September 1995.
[11] J. Hertz, A. Korgh and Palmer, "Introduction to the Theory of Neural Computation", Addison-wesley, 1991.
[12] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules", 3st Conference on Information and knowledge Management(CIKM-94), pp. 401-407, November 1994.
[13] H. Mannila and K. J. Gaiha, "Dependency Inference", in Proc. of 3rd Intl. Conf. on VLDB, pp.155-158, 1987.
[14] A. Muller, "Fast sequential and parallel algorithms for association rule mining: a comparison", University of Maryland-College Park CS Technical Report, CS-TR-3515, 76 pages, August, 1995.
[15] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila, "Pruning and Grouping Discovered Association Rules", MLnet Workshop on Statistics, Machine Learning, and Discovery in Database, 47-52, April 1995.
[16] Ramakrishnan Srikant and Rakesh Agrawal, "Mining Generalized Association Rules", Proceedings of the 21st VLDB Conference, Zurich, Swizerland, pp. 407-419, 1995.

저 자 소 개

정 홍(Hong Chung)
1972년 : 한양대학교 원자력공학과(공학사)
1976년 : 고려대학교 경영대학원(경영학석사)
1996년 : 대구가톨릭대학교 전산통계학과 (이학석사)
1999년 : 대구가톨릭대학교 전산통계학과 (이학박사)
1972년~1981년 한국과학기술연구원 선임연구원
2000년~2001년 : 미국 Washington State University 연구교수
1981년~현재 계명대학교 컴퓨터전자공학부 부교수

관심분야 : 지능정보시스템, 소프트웨어공학
Phone : 053-580-5264
Fax : 053-580-5165
E-mail : jhong@kmu.ac.kr

김진상(Jin Sang Kim)
1978년 : 경북대 수학과(이학사)
1981년 : 한국과학기술원 전산과(이학석사)
1990년 : 임페리얼 칼리지 전산과(박사수료)
1982년~현재 : 계명대학교 컴퓨터공학부 교수

관심분야 : 기계학습, 데이터마이닝, 시멘틱웹
Phone : 053-580-5270
Fax : 053-582-7460
E-mail : jsk@kmu.ac.kr

이종민(Jong Min Lee)
1999년 : 계명대학교 컴퓨터공학과(공학사)
2001년 : 계명대학교 컴퓨터공학과(공학석사)
2001년~현재 : 보험개발원 정보통신본부

관심분야 : 지능정보시스템, 소프트웨어공학
Phone : 02-368-4084
Fax : 02-368-4047
E-mail : jmgate@kidi.or.kr