신경망과 다단계 연관규칙을 이용한 구매 패턴 분류 시스템의 설계

Design of Purchasing Pattern Classification System Using Nural Network and Multiple-Level Association Rules

이종민, 정 홍
계명대학교 컴퓨터전자공학부
Jong Min Lee, Hong Chung
Keimyung University, Faculty of Computer and Electronic Engineering

요약

신경망을 이용해 고객집단을 분류하고 고객의 특성에 따라 세분화된 고객들에 대해 다단계 연관규칙을 적용해서 고객의 상품 구매패턴을 찾아 줌으로써 마케팅 전략 결정을 지원하는 구매패턴분류 시스템을 설계한다. 고객분류를 위한 신경망 시스템은 다층 퍼셉트론에 역전파 알고리즘을 이용한다. 주소, 구매금액, 구매횟수, 고객 구분, 상권 등과 같은 고객정보를 입력층에 입력변수로 지정하고, 이에 따른 우량/일반고객을 출력변수로 지정한 후 신경망을 학습시키면, 실제의 우량/일반의 값과 예측되는 우량/일반의 값의 차이를 최소화시키면서 모형을 형성시켜 나가게 된다. 구매패턴 분류 시스템은 다단계 연관규칙을 이용한다. 고객분류 서브시스템을 통해 고객집단이 세분화되면 각각의 고객집단에 대해 TID와 품목 트랜잭션을 입력으로 cumulate 알고리즘과 개념계층을 이용해 일반화 과정을 수행하면서 빈발 항목을 찾게 되고 이를 근거로 항목간의 연관규칙을 찾아내게 된다.

1. 서론

수요가 공급을 초과하는 시장상황에서 최근 공급이 수요를 초과하는 시장상황으로 반전되면서 시장의 주도권이 메이커에서 고객으로 옮겨지게 되었다. 그리고 컴퓨터와 정보통신 기술 등의 급격한 발전으로 고객 개개인의 자료를 수집하고 이를 데이터베이스화하는 것이 매우 용이하게 되었고, 시장이 성숙되고 고객의 요구가 다양화되면서 경쟁이 치열해 짐으로써 신규고객의 창출보다 기존고객의 유지가 보다 중요하게 되면서 DB마케팅이 등장하게 되었다[9]. DB마케팅이란 고객에 대한 여러 가지 정보를 컴퓨터를 이용하여 DB화하고 이를 바탕으로 고객 개개인과의 장기적인 관계 구축을 위한 마케팅 전략을 수립, 집행하는 모든 활동을 말한다[8].

본 논문은 신경망을 이용해 고객집단을 분류하고 고객의 특성에 따라 세분화된 고객들에 대해 다단계 연관규칙을 적용해서 고객의 상품 구매패턴을 찾아 줌으로써 마케팅 전략 결정을 지원하는 구매패턴분류 시스템을 설계한다.

2. 연관규칙과 Cumulate 알고리즘

일반적으로 연관규칙 탐사는 두 단계로 나누어진다[2,7]. 첫번째는 데이터베이스에서 추출될 수 있는 모든 항목집합 중 최소 지지도보다 높은 지지도를 갖는 모든 항목집합을 찾는 단계로서 추출된 항목집합을 빈발 항목집합(large itemset)이라고 한다. 두 번째 단계는 첫 단계에서 얻는 주요 항목 집합을 이용하고 데이터베이스를 참조하여 연관규칙을 찾는 단계이다.

항목들의 집합 $I = \{i_1, i_2, \ldots, i_m\}$라 하고 D를 트랜잭션의 집합(데이터베이스)이라고 하자. 각 트랜잭션 T는 $T \subseteq I$인 항목집합이다. 각 트랜잭션은 TID라는 유일한 식별자를 가진다. X가 I에 있는 몇 개의 항목들의 집합($X \subseteq I$)이라 할 때 $X \subseteq T$이면 트랜잭션 T는 X를 포함한다고 말하고 T가 집합 X를 지지한다고 한다. 연관규칙에서 $X \Rightarrow Y$는 $X \subseteq I$, $Y \subseteq I$, 그리고 $X \cap Y = \varnothing$임을 암시한다. 규칙 $X \Rightarrow Y$가 트랜잭션의 집합 D에서 신뢰도 c와 지지도 s를 가진다면 X를 포함하는 D에 있는 트랜잭션중에 c%가 Y를 포함하고 있고 D에 있는 트랜잭션중에 $X \cup Y$를 가진 트랜잭션이 s%임을 뜻한다. X의 지지도를 supp(X)로 정의하면 이것은 X를 지지하는 T에 있는 모든 트랜잭션 수를 의미한다. 만약 주어진 최소 지지도 s_{min}에 대하여 $supp(X) \geq s_{min}$이라면 집합 X는 빈발하다고 한다[1].

많은 응용 분야에서 데이터 항목 사이의 흥미로운 연관은 상대적으로 상위 수준의 개념에서 발생한다. 각 항목은 분류(taxonomy) 기준에 의하여 한 분류에 속하는데, 연관규칙 탐사에서는 이런 분류를 이용하여 보다 상위개념의 규칙을 찾아내게 된다. 일반화된 연관규칙의 탐사 문제에서는 각 항목의 분류를 포함하는 연관성을 찾게 된다[5].

일반적으로 각 항목을 일반화 시키기 위해 그림 1과 같은 개념 계층을 이용한다.

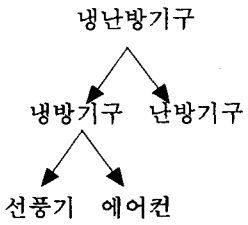

그림 1 개념 계층의 예

일반화된 연관규칙의 발견에 사용되는 다단계 연관규칙의 유도에는 Cumulate 알고리즘[5]이 사용된다. 이 알고리즘은 현 단계에서 계산된 후보 항목들의 상위 항목을 추가하고 포함되지 않은 항목은 트랜잭션에서 제거하며, 개념계층을 순회하여 각 항목의 상위항목을 찾지 않고 상위 항목을 미리 계산하면서 동시에 후보 항목들에 없는 상위 항목들은 제거한다. 그리고 항목과 상위항목을 모두 포함하는 항목집합을 전지하는 방법으로 일반화된 연관규칙을 유도한다. 이를 알고리즘으로 표시하면 다음과 같다.

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

3. 시스템 설계

그림 2는 전체 시스템의 구조도이다. 데이터베이스를 입력으로 하여 신경망을 사용한 분류 시스템을 통해 고객집단을 우량고객과 일반고객으로 분류한다. 그리고 각각의 고객집단에 대해 다단계 연관규칙과 개념계층을 이용해 고객의 구매 패턴에 대한 연관규칙을 발견하게 된다.

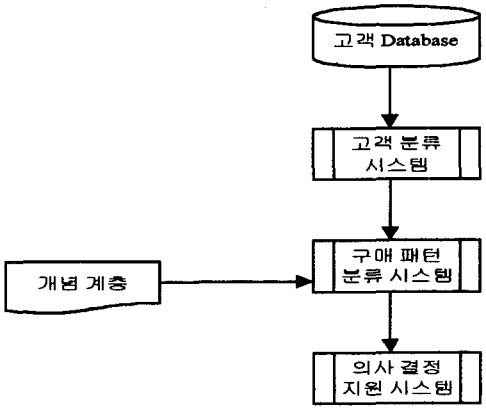

그림 2 시스템 구조도

3.1 고객분류 시스템

고객분류를 위한 신경망 시스템은 다층 퍼셉트론에 역전파 알고리즘을 이용한다. 주소, 구매금액, 구매횟수, 고객 구분, 상권 등과 같은 고객정보를 입력층에 입력변수로 지정하고, 이에 따른 우량/일반고객을 출력변수로 지정한 후 신경망을 학습시키면, 실제의 우량/일반의 값과 예측되는 (판별되는) 우량/일반의 값의 차이를 최소화시키면서 모형을 형성시켜 나가게 된다. 그림 3은 신경망을 이용한 고객분류 시스템의 구성이다.

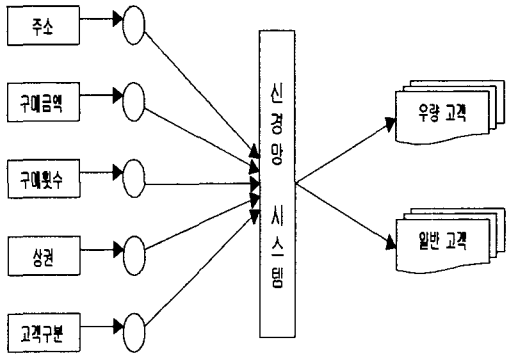

그림 3 고객 분류 시스템

역전파 알고리즘을 구현하고 학습하는 과정에서 지역 최소점에 빠지는 경우에 해결할 수 있는 획기적인 방안이 아직 고안되지 않은 상황이기 때문에 지역 최소점에 빠지거나 진동하는 경우에

는 학습계수나 초기값(weight, offset), 은닉 노드 수 등을 조절해 주어야한다.

3.2 구매패턴 분류 시스템

구매패턴 분류 시스템은 다단계 연관규칙을 이용한다. 그림 4는 구매패턴 분류 시스템의 구성이다. 고객분류 서브시스템을 통해 고객집단이 세분화되면 각각의 고객집단에 대해 TID와 품목 트랜잭션을 입력으로 cumulate 알고리즘과 개념 계층을 이용해 일반화 과정을 수행하면서 빈발 항목을 찾게 되고 이를 근거로 항목간의 연관규칙을 찾아내게 된다. 우량 고객은 개념 계층에서 서로 다른 수준의 연관규칙까지 유도할 수 있도록 하여 차별화 시킨다.

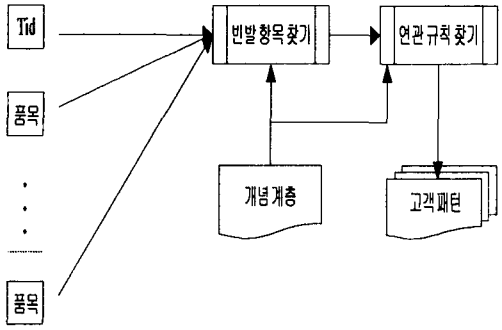

그림 4 구매패턴 분류 시스템

모든 트랜잭션의 항목들의 빈발항목을 계산하고 개념계층을 이용하여 상위 수준의 개념을 이끌어 내어 최소 지지도를 만족하는 빈발 항목집합을 찾고 빈발 항목집합으로부터 항목간의 연관규칙을 이끌어 내어 고객의 차후 상품 구매 성향을 알아낼 수 있도록 한다.

4. 시스템 실험

실험에 사용된 데이터는 전자대리점 데이터베이스를 이용하였다.

고객 분류 시스템에서 신경망 시스템은 데이터 추출시 최종 출력 요소의 값 중 개수가 적은 것이 있을 때 제대로 신경망에 학습되지 않을 수 있기 때문에 가능한 최종 출력 요소의 값이 비슷한 개수로 선택했다.

입력변수에 주소는 대구시를 동구, 서구, 남구, 북구, 수성구, 달서구와 기타지역 등 7개로 분류하고, 상권(주거지와 주거지에 가까운 대리점을 분류)은 A, B, C, D, G와 V 등 6개로 분류한다. 구매횟수는 1회, 2회, 3회 이상 등 3가지로 분류하고 고객구분은 임시와 고정 2가지로 분류한다. 구매금액은 200만원 이하, 200~300만원, 300만원 이상 등 3가지로 분류한다. 출력변수는 우량고객과 일반고객으로 나뉘고 분류 결과는 템플릿으로 출력한다[4].

고객 번호 50번에 대한 사항을 신경망 시스템을 통해 고객 분류를 한 결과는 그림 5와 같다.

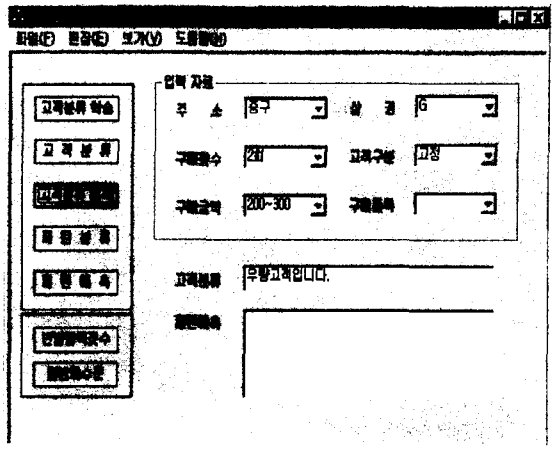

그림 5 고객 분류 결과 출력화면

고객 패턴 분류 시스템의 일반화된 연관규칙에 사용되는 트랜잭션 데이터의 품목들은 모두 정렬되어 있어야 한다. 입력변수는 트랜잭션 데이터베이스의 TID와 품목들이다. 개념계층은 코드화 되어있으며 미리 주어진다. 빈발항목은 빈발 4-항목까지 구하고 지지도와 신뢰도를 계산해 데이터베이스에 저장해서 재사용이 가능하도록 한다. 그리고 사용자가 어떤 수준에서 연관 규칙을 알고자 하는지 일반화 수준을 조절할 수 있도록 한다. 고객 번호 50번에 대한 구매패턴분류 결과는 그림 6과 같다.

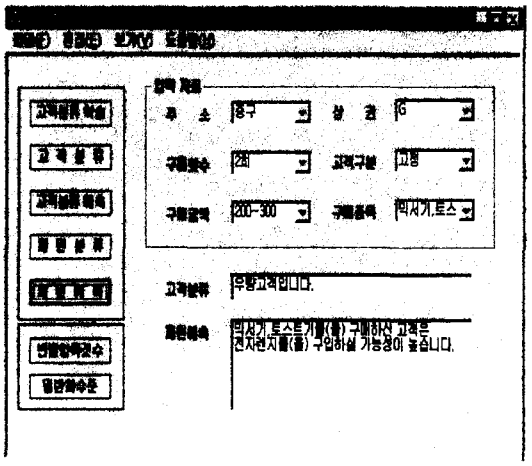

그림 6 전체 시스템의 출력화면

이와 같이 일반화 연관규칙에 의한 출력은 템플릿에 의해 사용자의 의사결정을 지원하는 예측을 보여준다.

5. 결론 및 평가

본 논문에서 설계한 구매패턴분류 시스템은 상품 판매에 대한 의사 결정을 지원하기 위해 데이

터의 분류에 있어 좋은 성능을 가진 신경망과 상품 판매 트랜잭션의 상관 관계를 찾는 연관규칙을 연결하여 고객의 상품 구매패턴을 유도하였다.

상품의 구매패턴을 분류할 때 사용자가 빈발 항목들의 개수를 지정할 수 있도록 하고 개념 계층의 수준을 지정할 수 있도록 하여 사용자 편이성을 도모하고 서로 다른 수준의 연관관계를 알 수 있도록 알고리즘을 개선하였다.

본 시스템이 실제 업무에 잘 적용되기 위해서는 먼저 데이터베이스 구축에 있어서 고객의 거래에 관련된 고객자료가 잘 정비되어져 있어야 한다. 또한 신경망 시스템은 출력에 대한 분석이 어렵다는 단점이 있으므로 통계적 기법, 의사 결정 트리, 유전자 알고리즘 등의 데이터마이닝 기법들을 통합하는 연구가 이루어져야 할 것이다.

앞으로 DB는 점점 대형화되어 질 것이고 이에 따라 데이터마이닝에 의한 DB마케팅은 앞으로 더 많은 필요성을 가지게 될 것이다. 이를 위해서는 모든 데이터베이스에 대해 같은 메커니즘을 사용할 것이 아니라 데이터베이스의 데이터 특성에 따라 적합한 데이터마이닝 기법을 적용하여야 할 것이다.

6. 참고 문헌

[1] Rakesh Agrawal, Tomasz Imielinski, 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.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", *VLDB '94*, pp. 487-499, September 1994.

[3] Joseph P. Bigus, *Data Mining with Neural network*, McGraw-Hill, 1996.

[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules", *Conference on Information and knowledge Management(CIKM-94)*, pp. 401-407, Nov. 1994.

[5] Ramakrishnan Srikant and Rakesh Agrawal, "Mining Generalized Association Rules", *Proceedings of the 21st VLDB Conference*, Zurich, Swizerland, pp. 407-419, 1995.

[6] 김대수, 신경망 이론과 응용(Ⅰ), 하이테크 정보, 1992.

[7] 박종수, 유원경, 홍기형, "연관규칙 탐사와 그 응용", 정보과학회지, pp. 37-43, 9월 1998.

[8] 박찬욱. 데이터베이스 마케팅, 연암사, 1996.

[9] 이상민, "DB마케팅", 금강기획 사보 ('98.3-4), http://www.diamond.co.kr/.