전자상거래에 적용 가능한 고객분류기†

°김선철, 이준욱, 이용준*, 류근호
충북대학교 컴퓨터과학과
*한국전자통신연구원 우정정보화팀

A Customer Classifier for EC Mall

Sun Cheul Kim, Jun Wook Lee, Yong Jun Lee* and Keun Ho Ryu
Dept. of Computer Science, Chungbuk National University / *ETRI
E-mail Address :{sckim,junux,khryu}@dblab.chungbuk.ac.kr

요 약

분류기법은 과거데이터를 분석하여 새로운 데이터에 대한 예측에 사용되며, 결정트리 알고리즘을 많이 사용한다. 따라서, 이 기법은 전자상거래에서 DB 마케팅을 위해 데이터베이스에 저장되어 있는 고객데이터를 분석하여 암시적인 고객들의 행위규칙을 찾고, 예측하기 위하여 사용할 수 있다. 기존의 분류알고리즘들은 전자상거래에서 일반적인 연속형 고객데이터를 처리하는 데는 많은 문제점를 가지고 있다. 이러한 문제를 해결하기 위하여 연속형 데이터를 범주형 데이터로 변환하는 알고리즘을 구현하였다. 이 논문은 전자상거래에 적용하기 위한 고객분류기로서 ID3알고리즘에 1차원 클러스터링알고리즘을 결합하여 사용한다.

1 서 론

최근 들어 전자상거래 시장이 활성화되면서 많은 쇼핑몰들은 새로운 유형의 차별화 된 서비스를 위한 방법들을 적용하고 있다. 기업들은 마케팅전략으로서 가격우위전략, 특화된 제품판매, 이벤트행사 등을 통해 고객을 확보하려는 고객중심의 마케팅전략을 전개하고 있다. 이를 위해서 수년간 누적되어 온 데이터베이스를 이용하여 마케터가 생각하지 못했던 고객 데이터들간에 존재하는 암시적인 정보를 찾아 효율적인 의사결정을 지원하기 위하여 여러 종류의 데이터마이닝기술이 적용되고 있다.

전자상거래의 기반이 되는 데이터베이스상에서 고객데이터는 각 속성들이 연속형 데이터와 범주형 데이터로 혼합되어 있으며, 그 규모가 크고, 일정한 형태의 특징들을 내포하고 있다.

분류기법은 기존의 데이터셋를 이용하여 레코드들 간에 존재하는 일정한 규칙을 찾아내며, 고객데이터베이스에 적용하여 고객들간의 행위규칙을 유도해 낼 수 있는 적합한 방법이다.

이 논문은 분류기법으로 가장 많이 사용되고 있는 ID3알고리즘[Quin89]에 기반을 두고 있다. 그러나 고객 데이터셋의 특성으로 인하여 ID3알고리즘을 바로 사용하는데는 어려움이 있다. 이러한 문제점을 해결하기 위하여 연속형 자료를 범주형 자료로 변환하는 클러스터링 알고리즘을 구현하였으며, ID3알고리즘과 클러스터링알고리즘을 결합함으로서 전자상거래에 적용 가능한 고객 분류 기법을 제시한다.

이 논문의 구성은 다음과 같다. 먼저 제2장에서는 관련연구를 통해 기본적인 분류기에 적용 가능한 데이터마이닝 알고리즘에 대하여 기술하고, 제3장에서는 고객분류기 모형을 제시한다. 제4장에서는 기존 분류기법을 확장하는 클러스터링알고리즘을 제시한다. 마지막으로 제5장에서는 결론 및 향후 연구사항을 기술한다.

2 관련연구

데이터마이닝에서 분류기법의 목적은 입력데이터를 분석하여 각 클래스에 대한 정확한 표현이나 모델을 개발하는데 있으며, 여러 분야에서 분류알고리즘들이 적용되고 있다. 기계학습분야에서 ID3[Quin89], C4.5[Quin93], CART[Mitc97] 알고리즘과 대용량 데이터를 처리할 수 있는 SLIQ[Meht96], SPRINT[Shaf96], Rainforest[Rost98]와 같은 알고리즘들이 제시되어 있으며, 통계학에서는 DISCRIM, CANDISC, STEPDIC절차에 의한 판별함수를 사용하는 판별분석이 제시되었다.

ID3(Inducted Decision Tree)는 여러 개의 입력변수 값과 한 목적변수를 갖는 레코드가 주어져 있는 경우에 적용하는 알고리즘이다 그러나 연속형 변수가 존재하는 경우에는 적용하기 어렵고, 정확도가 매우 낮다는 문제점을 가지고 있다. 이 기법은 먼저 트레이닝 셋으로부터 규칙을 생성하고, 생성된 규칙에 테스트 데이터를 적용하여 정확도를 구하며, 생성된 규칙을 새로운 레코드셋에 적용하여 어떤 그룹에 속하는가를 예측할 수 있다[Quin86].

C4.5는 ID3의 알고리즘을 확장하여 미지의 변수를 포함하는 레코드를 처리할 수 있고, 연속형 자료를 일정한 범위로 분할하는 기능을 가지고 있으며, 생성 가능한 규칙에 대한 확률을 적용하는 특징을 가지고 있다. 그러나 연속형 데이터를 일정

† 이 논문은 한국전자통신 연구원의 "통합 우정 물류 실시간 관제 시스템 개발" 사업의 일부 연구비 지원에 의해 수행 되었음

한 범위로 분할하기 때문에 데이터가 갖는 특성을 잃어버릴 수 있고, 확률을 적용하기 때문에 너무 많은 규칙을 생성하는 문제점을 가지고 있다[Quin86].

3 전자상거래에 적용하기 위한 고객분류기

관련연구에서 기술한 것처럼 분류의 목적은 과거의 데이터를 통해 특정한 그룹에 대한 규칙을 찾아내고, 새로운 객체에 대하여 찾아진 규칙을 적용함으로서 어떤 그룹에 속하는지를 예측하는 것이다. 예를 들어 "기업이 수년동안에 걸쳐 데이터베이스에 저장된 고객데이터를 이용하여 우수고객과 일반고객, 이탈고객으로 그룹핑하고, 또한 각 그룹에 대한 규칙을 찾고, 잠재성 고객들에 대한 예측을 하고자 한다." 이러한 예측을 기법으로 분류기법이 많이 사용되는데, 실제 데이터베이스상에서 고객들간에 존재하는 규칙을 찾기 위해서는 많은 고려사항들이 존재한다. 특히 데이터베이스 내에 저장되어 있는 고객데이터는 이질적 성질을 지니고 있다. 데이터마이닝에서 사용되는 분류알고리즘들은 연속형 자료에 대한 처리가 어렵다. 따라서 연속형 자료를 범주형으로 변환하는 방법이 필요하며[Agra92], 이에 연속형 자료에 대한 분류방법으로 1차원 클러스터링 알고리즘을 제시한다. 그림 1은 전자상거래에 적용 가능한 고객 분류기의 모형이다.

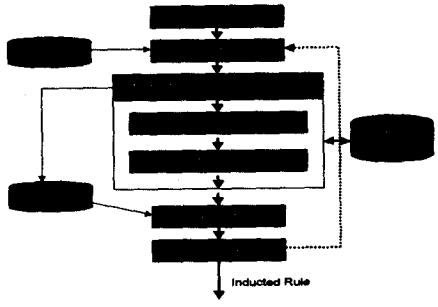

그림 1. 고객분류기시스템모형도

위 모형에서 개념계층은 시간단위나 지역코드등을 일반화하기 위한 배경지식을 표현하는 모듈로서 데이터베이스 내에 저장되고, 일반화 기법을 적용할 수 있는 기능을 제공하다. 또한, 1차원 분류기를 통하여 연속형 데이터를 분할하여 범주형으로 변환시키는 기능을 제공하는데, 이 알고리즘은 데이터의 분포에 따라 분할하기 때문에 그 특성을 잃지 않는다.

4 1차원 클러스터링알고리즘의 구현

클러스터링 기법은 모집단 또는 범주에 대한 사전정보가 없는 경우에 관측 값들 사이의 거리(또는 유사성)을 이용하여 전체를 몇 개의 그룹 또는 군집으로 나누는 분석법이다. 클러스터링에서 "클러스터간의 거리"에 대한 정의에 따라 여러 가지 클러스터링방법으로 나뉘어 진다. 이 논문에서는 가까운 객체끼리 차례로 묶는 계층적 방법을 사용한다. 거리를 나타내는 방법으로는 유클리드거리, 마할라노비스거리, 민코우스키거리 등이 있으며, 유클리드 거리를 사용한다.

[정의1] 유클리드거리는 두점사이의 최단거리로 정의된다.

$$d(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+\ldots+(x_{ip}-x_{jp})^2}$$

[정의2] N개의 관측값 $x_1 \cdots, x_N$에 대하여 x_i , x_j 사이의 거리 d_{ij}를 다음과 같이 정의한다.

$$d_{ij}=d(x_i,x_j)$$

[정의3] 두 군집간의 거리는 최단거리로 정의한다. 즉, 두 군집 C_1, C_2 사이의 거리는 다음과 같이 정의한다.

$$d(C_1,C_2)=\min(d(x,y):x\in C_1, y\in C_2)$$

위의 정의로부터 1차원 공간상에 위치하는 각 객체들의 클러스터를 생성하는 알고리즘을 그림2에서 보여주고 있다.

```
T ← user defined value /* 임계치 */
 initialize Object obj, Vector vector
 while(vector 크기 > 임계치 T) {
    result_vector ← Clustering(vector)
     return result_vector
 }
 clustering(_vector) {
    for(int i=0 ; i<vector 크기-1 ; i++) {
        interval ← cal_distance(obj(i), obj(i+1)) 메소드호출
        if(min_dist>interval) {
           min_dist ← interval;
           left_index←i; right_index←i+1;
        }
        combine(Obj(left_index),Obj(right_index))
 }
 /* 두 객체간의 거리를 계산하는 메소드 */
 cal_distance(Obj1, Obj2){
    if (Obj2 > Obj1)
       return Obj2.leftpoint - Obj1.rightpoint;
    else
       return Obj1.leftpoint - Obj2.rightpoint;
 }
 /* 두 객체를 병합하는 메소드 */
 combine(Obj1, Obj2, Obj1_idx, Obj2_idx) {
    if (Obj2>Obj1)
       new_Obj.leftpoint = Obj1.leftpoint
       new_Obj.righpoint = Obj2.rightpoint
    else
       new_obj.leftpoint=Obj2.leftpoint
       new_obj.rightpoint=Obj1.rightpoint
    remove(Obj1_idx) from vector
    remove(Obj2_idx) from vector
    add elementAt(_left_index) to vector
 }
```

그림 2. 1차원 클러스터링 알고리즘

분류알고리즘에 고객데이터를 적용하기 위해서는 연속형 자료를 범주형 자료로 변환해야 한다. 표 1은 고객데이터베이

스에 입력되어 있는 여러 개의 속성들 중에서 나이를 나타내는 30개의 연속형 자료의 구성을 보여주고 있으며, 그림 3에서 보는 바와 같이 5개의 집단으로 구분될 수 있는 형태를 지니고 있다. 따라서, 위 알고리즘을 적용하기 위해서는 임계치를 결정해야 하며, 아래의 데이터에 대해 클러스터링을 수행하기 위해서는 데이터의 구조상 값을 5로 부여한다. 임계치는 클러스터의 생성 개수를 나타내며, 데이터의 분포형태에 따라 알맞게 부여해야 한다. 그 값에 따라 분류기로부터 유도된 규칙의 정확도가 좌우될 수 있기 때문이다.

표 1. 연속형 고객데이터(나이)

OID	Age	OID	Age	OID	Age
1	24	11	54	21	55
2	27	12	56	22	42
3	28	13	19	23	39
4	33	14	24	24	34
5	34	15	27	25	34
6	25	16	28	26	25
7	29	17	35	27	33
8	40	18	40	28	56
9	41	19	53	29	23
10	53	20	34	30	25

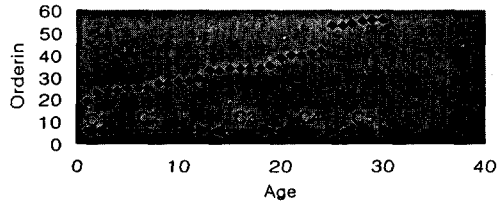

그림 3. 나이 데이터 분포도

앞에서 기술한 연속형 데이터에 1차원 클러스터링 알고리즘을 적용하면 25번의 병합과정을 거쳐 표 2와 같은 5개의 객체가 생성되고, 그 각각의 객체가 클러스터가 되는 것이다. 병합하는 방법은 두 객체간에 최소거리를 계산하여 가장 가까운 거리에 있는 객체를 병합하는 방법을 사용한다. 병합되는 객체는 시작점과 끝점을 가지고 있으며, 객체간의 거리 계산은 각 시작점과 끝점을 사용한다.

표 2. 생성된 클러스터

Object ID	Left Point	Center Point	Right Point	Cluster
13	19	19	19	C1
29,1,14,25,30,2,15,26,3,16	23	25.5	28	C2
4,27,5,20,24,6,7	33	34	35	C3
7,23,8,18,9,22	39	40.5	42	C4
10,19,11,21,12,28	53	54.5	56	C5

표 2에서 보여주는 바와 같이 각 클러스터는 범위 값을 갖게 되며 다음과 같은 특징을 갖는다. 클러스터는 시작점과 끝점, 그리고 중심점을 갖는 객체로서 저장된다. 따라서, 새로 삽입되는 객체가 존재하는 클러스터범위의 내부 값이라면 해당 클러스터에 포함되고, 밖의 값이면 가장 가까운 두 클러스터 에 대한 비교를 통해 삽입된다. 또한, 각 클러스터의 중심점을 갖고 있기 때문에 최단거리 기법이 아닌 중심 연결 방법을 사용할 수도 있으며, 가장 큰 특징은 데이터의 분포에 따라 분할하기 때문에 불필요한 분할을 하지 않고 분할 시에 데이터의 특성을 잃지 않는다

이 알고리즘은 고객의 행위 규칙을 찾는데 사용할 ID3알고리즘의 연속형 자료에 대한 처리문제를 해결하기 위하여 제시되었으며, 적절한 임계치가 부여되었을 경우에는 매우 정확한 클스터링을 할 수 있다.

5 결 론

이 논문에서는 전자상거래에 적용하기 위한 고객 분류기 시스템모형을 제시하였다. 분류기법으로 ID3알고리즘을 채택했으며, 이 알고리즘은 범주형 자료에 대해서는 매우 빠르고 정확하지만 연속형 데이터에 대해 분류 정확도가 현저히 떨어지는 문제점을 지니고 있다. 이러한 문제점을 해결하기 위해 1차원 클러스터링 알고리즘을 제시했다. 이 알고리즘은 데이터의 분포를 고려하여 객체들을 병합하기 때문에 유동적으로 객체들을 클러스터링하여 양질의 범주형 자료를 생성했다. 또한, 이 시스템은 개념계층을 배경지식으로 가지고 있기 때문에 데이터마이닝 기술의 한 지류인 일반화 기법을 적용 할 수 있다.

향후 연구과제로는 현재 구현한 ID3알고리즘과 1차원클러스터링 알고리즘의 가시화 기법에 대한 연구가 이루어져야 하며, 1차원 클러스터링 알고리즘은 2차원으로 확장하여 공간 데이터 처리에 대한 연구를 진행할 계획이다.

[참고문헌]

[Agra92] Agrawal. R. et.al., "An Internal Classifier for Database Mining Applications," VLDB , 1992.

[Meht96] Mehta. M., Agrawal. R., and Rissanen. J., "SLIQ : A fast Scalable Classifier for Data Mining", EDBT, 1996.

[Quin86] Quinlan. J. R. "Induction of Decision Trees", Machine Learning, 1, pp.81-106, 1986.

[Quin89] Quinlan. J. R. "Induction of decision trees using minimum description length principle, Information and Computation", 1989.

[Quin93] Quinlan. J. R. "C4.5: Programs for and Neural Networks," Cambridge University Press, Cambridge, 1996.

[Shaf96] Shafer. J. et.al., "SPRINT: A scalable parallel classifier for data mining," VLDB, 1996.

[Rast98] Rastogi. R. et.al., "PUBLIC: A decision tree classifier that integrates building and pruning," VLDB, 1998.

[Mitc97] Mitchell.T. M. "Machine Learning", The MecGraw-Hill Companies.Inc, 1997.

[이강태99] 이강태, "시간연관규칙 탐사 기법," 충북대학교 석사학위 논문, 1999.