클러스터링과 특성분석을 이용한 구간 데이터에서 다차원 연관 규칙 마이닝

(Mining of Multi-dimensional Association Rules over Interval Data using Clustering and Characterization)

임 승 환 † 권 용 석 ††
(Seung-Hwan Lim) (Yong-Suk Kwon)

김 상 욱 †††
(Sang-Wook Kim)

요 약 비 트랜잭션 데이터를 대상으로 연관 규칙을 도출하기 위해서, 데이터의 속성들을 구간화하는 기법들이 활발하게 연구되었다. 이러한 기존의 연구들은 구간화 단계에서 구간 범위의 변화에 따른 연관 규칙의 신뢰도 변화를 반영하지 않고, 구간화 단계와 연관 규칙을 도출하는 단계들을 독립적으로 수행하였다. 이로 인해 속성들의 구간이 부적절하게 설정되고, 이 결과 높은 신뢰도를 갖는 연관 규칙들이 최종 결과에서 누락된다. 따라서 본 논문에서는 속성들을 구간화하는 단계와 연관 규칙들을 도출하는 단계를 병합하여 동시에 수행함으로써, 가장 신뢰도가 높은 연관 규칙들을 도출할 수 있는 구간을 설정하는 방안을 제안한다. 이를 위해서 연관 규칙의 우변의 속성들을 대상으로 계층적 클러스터링을 수행하고, 각 클러스터들에 대해서 특성 분석을 수행한다. 실험 결과, 제안하는 기법은 기존의 기법들에 비해서 높은 신뢰도를 갖는 연관 규칙들을 발견하는 것으로 나타났다.

키워드 : 연관 규칙, 데이터마이닝, 클러스터링, 특성 분석

Abstract To discover association rules from non-transactional data, there have been many studies on discretization of attribute values. These studies do not reflect the change of discovered rules' confidence according to the change of the ranges of the discretized attributes, and perform the discretization stage and the rule discovery stage independently. This causes the ranges of attributes not properly discretized, thereby making the rules having high confidence excluded in the result set. To solve this problem, we propose a novel method that performs the discretization and rule discovery stages simultaneously in order to discretize ranges of attributes in such a way that the rules having high confidence are discovered well. To the end, we perform hierarchical clustering on the attributes in the right hand side of rules, then do characterization on every cluster thus obtained. The experimental result demonstrates that our method discovers the rules having high confidence better than existing methods.

Key words : Association Rules, Data Mining, Clustering, Characterization

1. 서 론

비 트랜잭션 데이터들을 대상으로 신뢰도가 높은 연관 규칙들을 도출하기 위해서는 유사한 연관 규칙을 보이는 데이터들에 동일한 항목을 부여하는 것이 중요하다. 각 데이터들이 부여받는 항목은 각 속성의 구간화 단계에서 결정되므로, 결과로 도출된 연관 규칙들의 신뢰도는 각 속성의 구간 설정 방법에 절대적인 영향을 받는다고 할 수 있다. 따라서 속성의 구간 설정 방법에 대한 연구는 활발하게 논의되어 온 주제이다[1-4].

참고문헌 [2]는 연관 규칙 마이닝의 후처리 단계에서 규칙들을 병합함으로써, 규칙의 속성들이 구간 값을 갖도록 가공하는 방법을 제안하였다. 또한, 참고문헌 [1]은 연관 규칙 마이닝의 전처리 단계에서 데이터를 구간화하는 방법을 제안하였다. 속성들의 구간이 어떻게 설정되었는지에 따라서 도출되는 연관 규칙들과 이들의 신뢰도도 변화한다. 그렇지만, 참고문헌 [1,2]의 방법들은 연관 규칙을 도출하는 과정과는 상관없이 전처리 단계, 후처리 단계에서 속성들의 구간을 결정하므로, 구간 결정에 따른 연관 규칙들의 신뢰도 변화를 반영할 수 없다는 단점을 갖고 있다. 이는 이러한 기법들이 연관규칙 도출 시에 신뢰도가 높은 연관 규칙을 누락시킬 수 있는 가능성을 갖고 있음을 의미한다.

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2008-000-20872-0)
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '클러스터링과 특성분석을 이용한 구간 데이터에서 다차원 연관 규칙 마이닝'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한양대학교 전자컴퓨터통신공학과
shlim@hanyang.ac.kr
†† 정 회 원 : 삼성전자 무선연구소
yongsuk.kwon@samsung.com
††† 종신회원 : 한양대학교 전자컴퓨터통신공학과 교수
wook@hanyang.ac.kr

논문접수 : 2009년 8월 14일
심사완료 : 2009년 10월 20일

따라서 본 논문에서는 연관 규칙 도출 작업과 속성들의 값을 구간화 하는 작업을 동시에 수행하면서, 상호작용을 통해서 높은 신뢰도를 갖는 연관 규칙들을 도출하는 방법을 제안한다. 또한, 제안하는 기법의 구간 설정 방법을 통해서 기존의 기법들에서는 발견할 수 없었던 의미 있는 연관 규칙들을 도출할 수 있음을 정량적으로 검증한다.

2. 관련연구

참고문헌 [1]은 연관 규칙 마이닝의 후처리 단계에서 연관 규칙들을 병합함으로써 규칙이 포함하는 속성들을 구간화하는 방법을 제안하였다. [1]에서는 연관 규칙의 우변(RHS)은 단일 속성으로 표현되고, 좌변(LHS)은 n개의 속성들로 표현되는 형태 즉, $A_1 \wedge A_2 \wedge ... \wedge A_n \Rightarrow B$와 같은 규칙들을 도출하는 방법을 제안하였다. 이러한 기법의 기본적인 아이디어는 다음과 같다. 우선, 좌변에 포함되는 n개의 속성들을 각각 일정한 크기를 갖는 작은 구간들로 분할한다. 이러한 방법으로 좌변에 포함되는 속성들을 이용하여 n차원의 격자를 생성한다. 생성된 격자에서 좌변과 우변의 조건을 동시에 만족하는 칸들을 표시한다. 이렇게 표시된 칸들은 각각 하나의 작은 규칙을 의미하게 된다. 이후, 표시된 칸들 중에서 인접하고 있는 칸들을 병합하여 하나의 규칙으로 만듦으로써 도출된 규칙의 수를 줄인다.

연관 규칙에서 우변은 분석가가 관찰을 통하여 이미 알고 있는 현상이고, 좌변은 분석가가 알고자하는 우변의 원인을 의미한다. [1]의 기법은 우변이 단일 속성으로 이루어져 있어야만 한다는 제약사항을 갖고 있다. 그러나 실제 세계에서 관찰되는 현상들 중에서 단일 속성으로 표현할 수 있는 경우는 매우 드물다. 우변은 단일 속성 보다는 다수의 속성으로 표현될 때에 현상을 보다 정확하게 반영할 수 있다.

따라서 우변이 단일 속성으로만 표현된다는 것은 도출된 연관 규칙의 유용성을 저하시키는 원인이 된다. 이에 본 연구에서는 다수의 속성으로 표현되는 우변을 갖는 연관규칙을 도출하는 방안을 제안한다.

참고문헌 [2]는 연관 규칙 마이닝의 전처리 단계에서 데이터를 구간화하는 방법을 제안하였다. 그 과정은 다음과 같다. 우선, 데이터들이 갖고 있는 속성들에 대하여 각 속성들을 기준으로 클러스터링을 수행한다. 이 단계에서 각 속성마다 데이터들의 클러스터들이 생성된다. 이후 각 속성별로 클러스터에 포함되는 데이터들이 갖는 값의 범위를 해당 클러스터의 구간으로 설정한다.

[2]에서는 이렇게 설정된 구간들을 이용하여 연관 규칙을 도출한다. 서로 다른 속성들을 기준으로 생성된 클러스터들에 공통적으로 포함되는 데이터들이 기준치 이상 존재한다면, 이들 클러스터들은 상호 연관이 있다고 할 수 있고, 이를 통해서 연관 규칙을 도출할 수 있다.

연관 규칙에 포함되는 속성들의 구간이 어떻게 설정되었는지에 따라서 도출되는 연관 규칙들과 이들의 신뢰도도 변화한다. 그러나 앞서 살펴본 바와 같이 [1,2]는 연관 규칙을 도출하는 과정과 별개로 속성들의 구간화 작업을 전처리, 후처리 단계에서 수행한다. 이는 속성들의 구간화에 따르는 연관 규칙들의 신뢰도 변화를 반영하고 있지 못하므로, 부적절한 구간을 설정하는 원인으로 작용한다. 따라서 이러한 기법들은 신뢰도가 높은 연관 규칙들을 결과에서 누락할 수 있는 가능성을 갖고 있다. 이에 본 논문에서는 속성들의 구간화 단계와 연관 규칙 도출 단계를 동시에 수행하여, 신뢰도가 높은 연관 규칙들이 도출되도록 속성들의 구간을 설정하는 방법을 제안한다.

3. 제안하는 기법

본 논문에서는 분석 대상이 특정 속성의 변화에 따라서 다른 속성들이 어떠한 변화를 빈번하게 보이는지를 연관 규칙의 형태로 도출하고자 한다. 이를 위해서 본 논문에서 도출하는 연관 규칙은 좌변은 단일 속성, 우변은 다수의 속성들로 표현된다. 우변에 다수의 속성들을 포함함으로써 특정 속성의 변화에 따른 분석 대상의 변화를 세밀하게 표현할 수 있다. 예를 들어, 일조량의 변화에 따른 기온의 변화만을 분석하는 것보다, 일조량의 변화에 따른 기온, 습도, 풍량 등의 변화를 함께 고려하여 분석하는 것이 현실 세계를 보다 자세하게 표현하고 있다고 볼 수 있다.

또한, 본 논문에서는 속성들의 구간을 결정하는 과정과, 연관 규칙을 도출하는 과정을 동시에 수행하고자 한다. 이러한 상호작용을 통하여 높은 신뢰도를 갖는 연관 규칙들을 도출할 수 있으며, 기존의 연구에서 부적절하게 설정된 속성들의 구간으로 인해 발견하지 못했던 연관 규칙들을 도출할 수 있다.

3.1 용어 정리

본 절에서는 표 1을 통하여 앞으로의 논의 전개를 위해서 필요한 용어 및 기호들을 정리한다. 본 논문에서 도출하는 연관 규칙은 좌변과 우변에 포함되는 속성들의 구간으로 나타낸다. 이러한 규칙은 $A_i \Rightarrow B_{1,x} \wedge B_{2,y} \wedge ... \wedge B_{n,z}$의 형태를 가진다. 여기서 A_i는 좌변에 해당하는 속성 A의 i번째 구간을 의미한다. 우변은 n개의 속성들의 구간들로 구성되고, $B_{m,w}$는 우변에 해당하는 속성 B_m의 w번째 구간을 의미한다. 본 논문에서는 계층적 클러스터링 기법을 이용하여 연관 규칙을 도출하는데, 이 과정에서 다수의 클러스터들이 생성된다. 이러한 클러스터들 중에서 식별자가 j인 클러스터를 C_j로 나타낸다.

표 1 용어 정리

$A_i \Rightarrow B_{1,x} \wedge B_{2,y} \wedge \ldots \wedge B_{n,z}$ [지지도, 신뢰도]: 규칙의 형태
A_i: 좌변에 해당하는 속성 A의 i번째 구간
$B_{m,w}$: 우변에 해당하는 속성 B_m의 w번째 구간
C_j: 식별자가 j인 클러스터

3.2 클러스터링

본 논문에서 도출하고자 하는 규칙은 우변이 n개의 속성들로 구성된 형태를 갖고 있다. 따라서 우변의 속성들을 구간화하기 위해서는 단일 속성으로 표현되는 좌변을 구간화 하는 방법과는 달리, n개의 속성들을 대상으로 분석을 수행하는 작업이 필요하다. 우변의 속성들을 가장 의미 있는 구간들로 분할하기 위해서는 유사한 속성 값을 갖는 데이터 집합들의 식별이 선행되어야 한다. 본 논문에서는 이를 위해서 계층적 클러스터링 기법을 이용한다.

클러스터링의 수행을 위해서 데이터들을 우변의 속성들의 값을 토대로 n차원의 공간상의 한 점으로 나타낸다. 각 차원은 우변의 속성들을 나타내고, 각 점들은 계층적 클러스터링의 수행을 위해서 필요한 초기 클러스터들이 된다. 또한, 점들 간의 유클리드 거리는 각각의 점들이 나타내는 데이터들 간의 유사도의 척도로서 이용된다.

이들 초기 클러스터들을 대상으로 클러스터의 개수가 1이 될 때까지 가장 가까운 거리에 있는 두개의 클러스터들을 식별하여 이들 클러스터들의 병합을 진행한다. 클러스터 간의 거리를 측정하는 방법에는 최소, 최대, 평균, 중심 거리를 이용하는 방법들이 있으며, 본 논문에서는 그 중에서 가장 널리 사용되는 최소 거리 기법을 이용하였다. 최소 거리 기법은 두 클러스터에 각각 속해 있는 데이터들의 쌍의 거리 값들 중에서 최소값을 두 클러스터 간의 거리로 부여하는 방법이다[5,6].

이러한 방법으로 도출된 클러스터들은 각각 우변의 속성들의 구간 값을 나타낸다. 즉, 클러스터 C_j에 포함되는 데이터들이 갖는 속성 B_m의 값들의 범위가 C_j가 나타내는 B_m의 구간인 것이다.

이는 각각의 클러스터들마다 좌변 속성의 구간들과 결합하여 연관 규칙으로 나타낼 수 있음을 의미한다. 즉, C_j가 나타내는 구간들 $B_{1,x}$, $B_{2,y}$, ..., $B_{n,z}$는 좌변 속성의 임의의 구간 A_i와 결합하여 연관 규칙 $A_i \Rightarrow B_{1,x} \wedge B_{2,y} \wedge \ldots \wedge B_{n,z}$를 나타낼 수 있다. 각 클러스터들은 최대 좌변 속성의 구간들의 개수만큼의 연관 규칙들을 나타낼 수 있는데, 본 논문에서는 분석가가 설정한 임계값 이상의 신뢰도를 갖는 연관 규칙들만을 인정하는 것으로 한다.

3.3 특성 분석

계층적 클러스터링에서 병합 과정이 진행됨에 따라, 대체로 클러스터가 나타내는 연관 규칙들의 우변 속성들의 구간들의 범위가 커지게 된다. 또한, 이로 인해 임계값 이상의 신뢰도를 만족하는 연관 규칙들의 개수도 줄어드는 경향을 보인다.

두개의 연관 규칙이 동일한 좌변과, 동일한 신뢰도를 갖는 경우, 우변 속성들의 범위가 클수록 보다 유용한 규칙이라고 할 수 있다. 예를 들어서, 연관 규칙(40<=age<45) ⇒ ($50,000<=salary<$55,000) [지지도: 0.3]과 (40<=age<45) ⇒ ($50,000<=salary<$70,000) [지지도: 0.3]의 경우, 후자가 보다 많은 경우를 포함하고 있으므로, 전자에 비해서 유용한 연관 규칙임을 알 수 있다. 따라서 본 논문에서는 계층적 클러스터링의 병합 과정에 따른 클러스터들의 특성 변화를 살펴보면서, 신뢰도의 손실이 없는 경우, 가급적 넓은 범위를 갖도록 속성들의 구간들을 설정하는 기법을 제안한다.

본 논문에서 클러스터의 특성은 연관 규칙의 좌변 속성의 항목들과 해당 클러스터에 속한 데이터들과의 연관성의 정도로 표현된다. 즉, 클러스터 C_j의 특성은 좌변 속성의 구간들 A_1, A_2, A_3, A_4, A_5 등과 C_j에 포함된 데이터들 간의 신뢰도로 표현된다. 따라서 C_j가 A_i에 대하여 갖는 특성 score(C_j,A_i)는 식 (1)과 같이 정의된다. 여기서, P(C_j|A_i)는 A_i를 만족하는 데이터들 중에서 C_j에 포함되는 데이터들의 비율 즉, $A_i \Rightarrow C_j$의 신뢰도를 의미한다.

$$score(C_j,A_i) = P(C_j|A_i) \quad (1)$$

신뢰도의 손실 없이 속성들의 구간들이 넓은 범위를 갖도록 설정하는 방법은 다음과 같다. 클러스터 C_j, C_k, 이들의 병합으로 인해 생성된 클러스터 C_l에 대해서 특성 분석을 수행하여 좌변 속성의 임의의 구간 A_i에 대한 각 클러스터들의 특성 score(C_j,A_i), score(C_k,A_i), score(C_l,A_i)를 계산한다. score(C_l,A_i)의 값이 score(C_j,A_i)나 score(C_k,A_i) 보다 작은 경우에, 이들 클러스터에 대한 병합을 중단하고, score(C_j,A_i), score(C_k,A_i) 중에서 임계값 이상인 연관 규칙을 최종 결과 집합에 포함 시킨다. 그 외의 경우에는 계층적 클러스터링의 병합 과정을 진행한다.

4. 성능 평가

4.1 실험 환경

본 논문에서는 성능 분석을 위하여 보스턴시의 집 가격과 관련된 데이터를 사용하였다. 이는 1997년 7월에 수집된 것으로서 506개의 레코드, 14가지의 속성들로 구성되어 있다[7].

본 연구에서 성능 평가의 대상으로 선정한 속성의 구간화 기법들은 본 논문에서 제안하는 기법인 RMUC

(Rule Mining Using Clustering), 클러스터링을 이용하는 기법인 Clustering, 포함하는 데이터의 개수가 동일하도록 구간을 설정하는 기법인 Equi-depth로 총 세 가지이다. 기법 Clustering과 기법 Equi-depth는 참고문헌 [2]에서 성능 분석을 위해서 사용된 기법들이다.

위의 세 가지 기법들을 통해서 우변 속성들의 구간을 설정하고, 이를 이용하여 연관 규칙을 도출한다. 본 실험에서는 속성 RM, AGE, MEDV의 조합들을 우변으로 갖고, 그 외의 속성들을 각각 좌변으로 갖는 연관 규칙을 도출 한다. 이는 어떤 속성이 방의 개수, 집의 나이, 집의 가격에 영향을 미치는지를 분석하는 것이다. 본 논문에서는 우변의 속성 RM, AGE, MEDV를 각각 B1, B2, B3로 나타낸다. 따라서 본 실험에서 도출하는 연관 규칙들의 우변은 B1, B2, B3의 가능한 조합들인 B1, B2, B3, B1&B2, B1&B3, B2&B3, B1&B2&B3의 총 7가지의 경우로 구성된다.

4.2 실험 결과

본 논문에서는 두 가지 종류의 실험을 수행하였다. 실험 1에서는 세 가지 기법들을 이용해서 도출한 연관 규칙의 개수, 평균 신뢰도를 비교하였고, 실험 2에서는 규칙들이 포괄하는 평균 면적, 평균 신뢰도/면적을 비교하였다.

실험 1에서는 기법 RMUC, 기법 Clustering, 기법 Equi-depth를 통하여 B1, B2, B3의 구간들을 설정하고, 이를 통해서 도출된 연관 규칙들의 개수, 규칙들이 갖는 신뢰도의 평균값을 비교하였다. 그림 1과 그림 2는 각각 도출된 연관 규칙의 개수, 평균 신뢰도를 그래프의 형태로 보인 것이다.

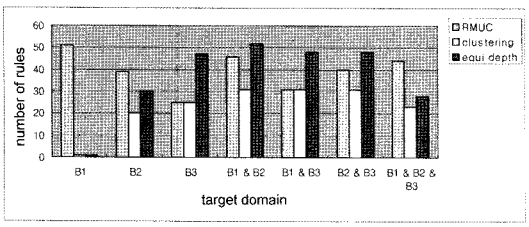

그림 1 도출된 연관 규칙들의 개수 비교

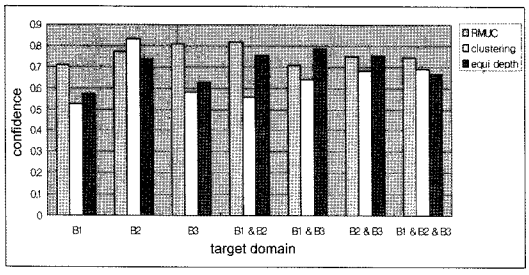

그림 2 도출된 연관 규칙들의 평균 신뢰도 비교

실험 1의 결과, 제안하는 기법인 RMUC가 기법 Clustering, 기법 Equi-depth에 비하여 전체적으로 많은 연관 규칙들을 도출하고, 이들의 신뢰도도 높은 것으로 나타났다.

기법 RMUC는 총 276개, 평균 39.43개, 평균 신뢰도 0.757554를 갖는 규칙들을 도출한 반면에, 기법 Clustering은 총 162개, 평균 23.14개, 평균 신뢰도 0.756102를 갖는 규칙들을 도출하였다. 또한, 기법 Equi-depth는 총 254개, 평균 36.29개, 평균 신뢰도 0.723875를 갖는 규칙들을 도출하였다.

기법 RMUC는 기법 Clustering에 비해서 평균 신뢰도는 유사한 반면에, 도출된 규칙의 수는 1.7배 가량 증가하였는데, 이는 기법 RMUC가 기법 Clustering이 도출하지 못한 높은 신뢰도를 갖는 연관 규칙들을 도출하였음을 의미한다. 또한, 기법 RMUC는 기법 Equi-depth에 비해서 도출되는 규칙의 수는 유사한 반면에, 평균 신뢰도는 0.033 가량 높은 값을 보이는데, 이는 기법 RMUC가 기법 Equi-depth에 비해서 우변 속성들에 보다 의미 있는 구간 범위가 설정되었음을 의미한다.

기법 RMUC는 연관 규칙의 우변에 다수의 속성들이 포함되는 경우인 B1&B2&B3에 기법 Clustering, 기법 Equi-depth에 비해서 우수한 성능을 보였다. 이를 통해서, 기법 RMUC가 다수의 속성들을 우변으로 갖는 연관 규칙을 도출하는 데에 적당한 기법임을 알 수 있다.

실험 2에서는 기법 RMUC, 기법 Clustering, 기법 Equi-depth를 통하여 B1, B2, B3의 구간들을 설정하고, 이를 통해서 도출된 연관 규칙들이 포괄하는 면적의 평균, 평균 신뢰도/면적의 값을 비교하였다.

연관 규칙의 우변의 속성들은 다차원 공간으로 나타낼 수 있는데, 본 논문에서는 이 공간상에서 특정 연관 규칙을 지지하는 지역의 넓이를 해당 연관 규칙의 우변의 면적이라고 부른다. 일반적으로 이러한 면적이 증가할수록 우변의 속성 구간 안에 해당 연관 규칙을 만족하지 않는 데이터들이 포함되는 비율이 증가하므로, 해당 연관 규칙의 신뢰도는 떨어지게 된다. 따라서 연관 규칙의 신뢰도를 해당 연관 규칙의 우변의 면적으로 나눈 값은 연관 규칙의 유용성에 대한 척도로 활용될 수 있다.

그림 3은 실험 2의 결과를 보인 것이다. 실험 결과, 제안하는 기법인 RMUC가 기법 Clustering, 기법 Equi-depth에 비하여 우변에 다수의 속성들이 포함된 경우에, 우수한 성능을 보이는 것으로 나타났다. B1&B2의 경우, 기법 RMUC가 기법 Clustering과 기법 Equi-depth에 비하여 평균 신뢰도/면적의 값으로서 각각 6.32배, 4.78배 높은 값을 보였고, B1&B3의 경우, 각각 5.57배, 4.59배, B2&B3의 경우, 각각 6.36배, 2.88배, B1&B2&B3의 경우, 각각 1.72배, 1.85배 높은 값을 보였다.

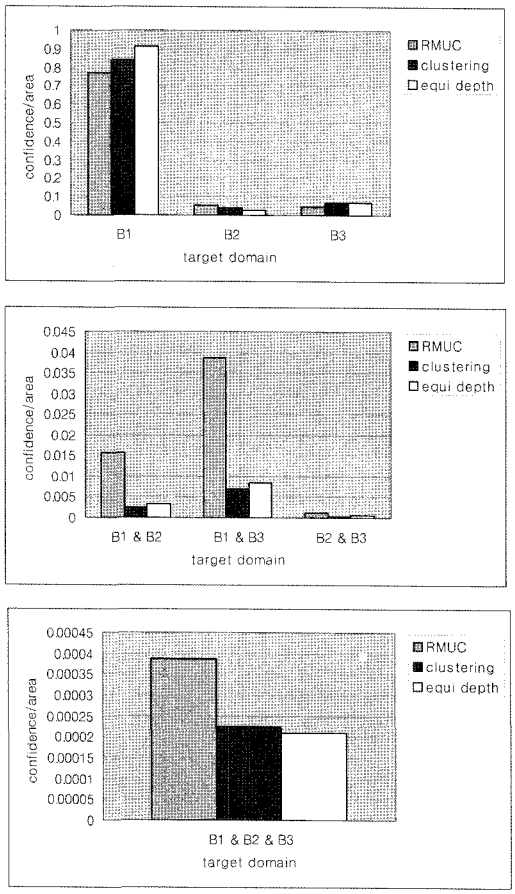

그림 3 평균 신뢰도/면적의 비교

실험 1과 실험 2의 결과를 통해서, 제안하는 기법이 기존의 기법들에 비해서, 다수의 속성들을 우변으로 갖는 연관 규칙을 도출하는 데에 유용한 기법임을 알 수 있다.

5. 결 론

본 논문에서는 비 트랜젝션 데이터를 대상으로 속성들을 구간화하여, 연관 규칙을 도출하는 방안에 대하여 논의하였다. 기존의 연구들은 속성들을 구간화하는 단계와 연관 규칙을 도출하는 단계를 분리하여 독립적으로 수행하였다. 이러한 제약으로 인해, 기존의 기법들은 속성들의 구간 범위를 결정하는 데에, 속성들의 구간화에 따른 연관 규칙의 신뢰도의 변화를 반영하지 못한다. 따라서 부적절한 구간 범위 설정으로 인해, 신뢰도가 높은 연관 규칙들이 누락될 수 있는 가능성을 갖고 있다.

이에 본 논문에서는 속성들의 구간의 범위를 결정하는 과정과 연관 규칙을 도출하는 과정을 병합하여 수행함으로써, 가장 높은 신뢰도를 갖는 연관 규칙들을 도출할 수 있는 구간 범위를 설정하는 방법을 제안하였다. 이를 위해서 연관 규칙의 우변의 속성들을 대상으로 계층적 클러스터링을 수행하여, 각 클러스터마다 특성 분석을 수행하였다. 클러스터의 특성은 해당 클러스터가 만족하는 연관 규칙의 신뢰도를 의미한다. 계층적 클러스터링의 병합 과정을 관찰하면서, 병합에 참여한 클러스터들과 병합되어 새롭게 생성된 클러스터의 특성 값을 비교하여, 병합 이후에 클러스터의 특성 값이 감소하는 경우에 병합 이전의 클러스터를 결과 집합에 포함시킨다.

실험 결과, 제안하는 기법은 기존의 클러스터링을 이용한 기법에 비해서 1.7배 가량 많은 연관 규칙들을 도출하는 것으로 나타났는데, 이는 제안하는 기법이 클러스터링 기법에서 누락되었던 연관 규칙들을 도출할 수 있음을 의미한다. 제안하는 기법은 데이터의 개수가 동일하도록 속성의 구간을 설정한 기법에 비해서, 연관 규칙의 신뢰도가 0.033 가량 향상된 것으로 나타났다. 또한, 제안하는 기법은 기존의 기법들에 비해서, 다수의 속성들을 우변으로 갖는 연관 규칙을 도출하는 데에 우수한 성능을 보이는 것으로 나타났다.

참 고 문 헌

[1] B. Lent, A. Swami, and J. Widom, "Clustering Association Rules," In *Proc. IEEE Int'l. Conf. on Data Engineering, IEEE ICDE*, pp.220-231, 1997.

[2] R. J. Miller and Y. Yang, "Association Rules Over Interval Data," In *Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD*, pp.452-461, 1997.

[3] R. Povinelli, *Identifying Temporal Patterns for Characterization and Prediction of Financial Time Series Events*, Springer Berlin, 2001.

[4] M. Kamber, J. Han, and J. Chiang, "Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, pp.207-210, 1997.

[5] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," In *Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD*, pp. 103-114, 1996.

[6] T. Zhang, R. Ramakrishnan, and M. Livny, "Data Clustering System BIRCH and Its Applications," *Data Mining and Knowledge Discovery*, vol.1, no.2, pp.141-182, 1997.

[7] D. Harrison and D. L. Rubinfeld, "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, vol.5, pp.81-102, 1978.