상용 데이타 마이닝 도구를 사용한 정량적 연관규칙 마이닝

(Mining Quantitative Association Rules using Commercial Data Mining Tools)

강 공 미 † 문 양 세 †† 최 훈 영 † 김 진 호 ††
(Gong-Mi Kang) (Yang-Sae Moon) (Hun-Young Choi) (Jin-Ho Kim)

요 약 상용 데이타 마이닝 도구에서는 기본적으로 이진 속성에 대한 연관규칙 마이닝만을 지원한다. 그러나, 일반적인 트랜잭션 데이타베이스는 이진 속성 뿐 아니라 정량적 속성을 포함한다. 이에 따라, 본 논문에서는 상용 데이타 마이닝 도구를 사용하여 정량적 연관규칙을 마이닝하는 체계적인 접근법을 제안한다. 이를 위해, 우선 상용 데이타 마이닝 도구를 사용하여 정량적 연관규칙을 찾아내기 위한 전체적인 프레임워크를 제안한다. 제안한 프레임워크는 정량적 속성을 이진 속성으로 변환하는 전처리 과정과 마이닝된 이진 연관규칙을 다시 정량적 연관규칙으로 변환하는 후처리 과정으로 구성된다. 다음으로, 전처리 과정을 위한 구간 분할의 개념을 제시하고, 기존의 평균 및 중앙치 기반 양분할 기법과 동일 너비 및 동일 깊이 기반 다분할 기법을 구간 분할의 개념으로 정형적으로 재정의한다. 그런데, 이들 기존 분할 기법은 속성 값의 분포를 고려하지 않은 문제점이 있다. 본 논문에서는 이를 해결하기 위하여 표준편차 최소화 기법을 제안한다. 표준편차 최소화 기법은 이웃한 속성 값의 표준편차 변화가 작다면 동일한 구간에 포함시키고, 표준편차 변화가 크다면 다른 구간으로 분할하는 매우 직관적인 분할 기법이다. 또한, 후처리 과정으로는 이진 연관규칙들을 통합하고 이를 다시 정량적 연관규칙으로 변환하는 방법을 제안한다. 마지막으로, 다양한 실험을 통하여 제안한 프레임워크가 바르게 동작함을 보이고, 표준편차 최소화 기법이 다른 기법에 비하여 우수함을 입증한다. 이 같은 결과를 볼 때, 제안한 프레임워크는 일반 사용자가 상용 데이타 마아닝 도구를 사용하여 정량적 연관규칙을 쉽게 마이닝 할 수 있는 매우 실용적인 접근법이라 생각한다.

키워드 : 연관규칙; 정량적 연관규칙; 데이타 마이닝; 상용 데이타 마이닝 도구

Abstract Commercial data mining tools basically support binary attributes only in mining association rules, that is, they can mine binary association rules only. In general, however, transaction databases contain not only binary attributes but also quantitative attributes. Thus, in this paper we propose a systematic approach to mine *quantitative association* rules---association rules which contain quantitative attributes---using commercial mining tools. To achieve this goal, we first propose an overall working framework that mines quantitative association rules based on commercial mining tools. The proposed framework consists of two steps: 1) a pre-processing step which converts quantitative attributes into binary attributes and 2) a post-processing step which reconverts binary association rules into quantitative association rules. As the pre-processing step, we present the concept of *domain partition*, and based on the domain partition, we formally redefine the previous bipartition and multi-partition techniques, which are mean-based or median-based techniques for bipartition, and are

· 본 연구는 첨단정보기술연구센터를 통하여 과학기술부/한국과학재단의 지원을 받았음

† 학생회원 : 강원대학교 컴퓨터과학전공
gmkang@hanmail.net
hychoi@kangwon.ac.kr
†† 종신회원 : 강원대학교 컴퓨터과학전공 교수
ysmoon@kangwon.ac.kr
(Corresponding author임)
jhkim@kangwon.ac.kr

논문접수 : 2007년 7월 16일
심사완료 : 2008년 1월 15일

정보과학회논문지: 데이타베이스 제35권 제2호(2008.4)

equi-width or equi-depth techniques for multi-partition. These previous partition techniques, however, have the problem of not considering distribution characteristics of attribute values. To solve this problem, in this paper we propose an intuitive partition technique, named *standard deviation minimization*. In our standard deviation minimization, adjacent attributes are included in the same partition if the change of their standard deviations is small, but they are divided into different partitions if the change is large. We also propose the post-processing step that integrates binary association rules and reconverts them into the corresponding quantitative rules. Through extensive experiments, we argue that our framework works correctly, and we show that our standard deviation minimization is superior to other partition techniques. According to these results, we believe that our framework is practically applicable for naive users to mine quantitative association rules using commercial data mining tools.

Key words : Association Rules; Quantitative Association Rules; Data Mining; Commercial Data Mining Tools

1. 서 론

대용량 데이타에서 잠재적 사용가치가 있는 패턴이나 추세 등의 규칙들을 발견하는 데이타 마이닝은 마케팅 전략 수립, 수요예측, 의료진단, 상품진열 등의 광범위한 분야에서 응용되고 있다. 이처럼 데이타에서 숨겨진 패턴을 탐색하는 데이타 마이닝에서 가장 많은 연구가 이루어진 분야가 연관규칙[1]이다. 연관규칙은 대용량의 데이타베이스에서 어떤 사건들이 함께 발생하거나, 또는 하나의 사건이 다른 사건을 암시하는 것과 같은, 사건간의 상호관계를 마이닝하는 것으로, 항목(들) X와 항목(들) Y 사이의 $X \rightarrow Y$ 형태의 규칙을 찾아낸다. 여기서 항목은 일반적으로 이진(boolean)속성을 가지며, 항목(들) X가 나타나면 항목(들) Y도 나타날 가능성이 높다는 연관관계를 표현한다.

본 논문에서는 상용 데이타 마이닝 도구에서 정량적 연관규칙을 마이닝하는 문제를 다룬다. 정량적 연관규칙이란 이진 속성이 아닌 정량적(수량적) 속성을 가진 항목들을 대상으로 하는 연관규칙을 의미한다. 그런데, 상용 데이타 마이닝 도구로는 이러한 정량적 연관규칙을 마이닝하기 어려운 문제가 있다. 그 이유는 기존의 상용 도구들은 이진 속성만을 다룰 뿐 정량적 속성은 지원하지 않기 때문이다. 또한, 연관규칙에서 정량적 속성을 다루기 위한 여러 연구가 진행되었는데, 이들 연구는 상용 데이타 마이닝 도구에는 적용이 어려운 문제가 있다. 왜냐하면, 이들 방법은 정량적 속성을 지원하기 위하여 대부분 연관규칙 마이닝 알고리즘 자체를 변경하였기 때문이다. 즉, 알고리즘 변경이 불가능한 기존 상용 도구에는 직접 적용하기 어렵다. 반면에, 본 연구에서는 상용 마이닝 도구는 변경이 어렵다는 제약 하에서 정량적 연관규칙 마이닝 문제를 다룬다. 이와 같이, 상용 마이닝 도구의 직접 활용으로 일반 사용자를 고려한다는 측면에서 본 연구의 결과는 실용성 및 효용성이 있다. 또한, 기존 연구들이 상용 마이닝 도구의 변경이 어려움을 간과한 반면, 본 연구는 이를 주된 제약사항으로 가정하고 연구를 전개했다는 측면에서 기존 연구와는 차별적이다. 이러한 이유에 의하여 본 논문에서는 상용 마이닝 도구에 적용이 가능한 정량적 연관규칙 마이닝 기법을 제시한다.

상용 데이타 마이닝 도구 기반의 정량적 연관규칙 마이닝을 위하여, 본 논문에서는 전처리 과정, 상용 데이타 마이닝 도구, 후처리 과정으로 구성되는 전체적인 동작 프레임워크를 제안한다. 먼저, 전처리 과정은 상용 마이닝 도구를 사용하여 연관규칙을 마이닝하기 이전의 과정으로서, 정량적 속성을 갖는 원 입력 데이타를 구간으로 분할된 이진 데이타로 변환하는 작업을 수행한다. 다음으로, 후처리 과정은 상용 마이닝 도구를 사용하여 연관규칙을 마이닝한 이후의 과정으로서, 마이닝된 규칙들을 통합하고 이진 속성을 다시 정량적 속성으로 변환하는 작업을 수행한다. 본 논문에서는 전처리 과정을 위해 다양한 구간 분할 방법을 제안하며, 후처리 과정을 위해 정량적 연관규칙을 통합하는 방법을 제안한다.

전처리 과정의 구간 분할 기법으로는 정량적 속성을 두 개의 구간으로 나누는 양분할(bipartition) 기법과 두 개 이상의 여러 구간으로 나누는 다분할(multi-partition) 기법을 각각 제안한다. 이를 위해 우선 구간 분할의 개념과 이를 이용하여 정량적 속성을 이진 속성으로 변환하는 과정을 정형적으로 정의한다. 그런 다음, 기존 연구에서 제안된 평균(mean)에 의한 양분할 기법과 중앙치(median)에 의한 양분할 기법을 제시한 구간 분할 개념으로 재정의한다. 다음으로, 다분할 기법으로 동일 너비(equi-width) 및 동일 깊이(equi-depth)를 사용하는 기존의 방법을 정형적으로 재정의한다. 그런데, 이러한 기존의 분할 기법은 정량적 속성 값에 대한 분포를 고려하지 않은 단순한 방법으로, 다양한 연관규칙을 찾을 수 없다는 단점이 있다. 따라서, 본 논문에서는 구간

별 표준편차를 최소화하는, 즉 구간별 표준편차의 합을 최소화하는 새로운 기법인 표준편차 최소화 기법을 제안한다. 표준편차 최소화 기법은 이웃한 속성 값의 변화가 심하지 않다면(즉, 표준편차가 작다면) 해당 속성 값들은 유사한 특성을 가지므로 동일한 구간에 포함시키고, 그 변화가 크다면(즉, 표준편차가 크다면) 해당 속성 값들을 다른 구간으로 분할한다는 직관에 기반한다. 본 논문에서 제안한 표준편차 최소화 기법은 양분할 기법과 다분할 기법 모두에 적용이 가능하다.

제안한 프레임워크의 후처리 과정은 매우 간단하게 수행된다. 먼저, 마이닝된 연관규칙들을 검사하여 이웃한 정량적 속성을 갖는 규칙들을 통합한다. 다음으로, 마이닝된 연관규칙들의 이진 속성을 원래의 정량적 속성으로 변환한다. 이와 같이 전처리 과정을 거쳐서 상용 마아닝 도구에 입력된 이진 속성 데이타는 후처리 과정을 통하여 결과적으로 다시 정량적 속성 데이타로 표현된다. 즉, 전처리 과정 이전에는 수량 값(혹은 범위)이 "트랜잭션 데이타"에 포함되어 있었으나, 후처리 과정까지 거친 후에는 수량 값(혹은 범위)이 "연관규칙"에 포함되어 나타나는 것이다.

본 논문에서는 양분할과 다분할을 구분하여 다양한 실험을 수행하였다. 이때 분할 기법의 우수성을 평가하기 위하여, 마이닝된 연관규칙의 개수와 해당 규칙들에 포함된 정량적 속성 값 범위를 사용하였다. 이는 연관규칙이 많을수록, 그리고 그 규칙에 포함된 정량적 속성 값의 범위가 클수록, 보다 정확하게 구간 분할을 수행한다고 볼 수 있기 때문이다. 실험 결과 제안한 표준편차 최소화 기법은 양분할 및 다분할 모두에 있어서 기존의 분할 방법보다 우수한 결과를 나타내었다. 먼저, 양분할에 있어서 표준편차 최소화 기법은 평균 기법에 비해 규칙 개수를 평균 57%, 속성 범위를 평균 15% 향상시켰으며, 중앙치 기법에 비해서는 평균 158%와 17%를 향상시킨 것으로 나타났다. 그리고 다분할에 있어서도 표준편차 최소화를 사용하는 경우는 그렇지 않은 경우에 비하여 최소 3%에서 최대 2배까지 우수한 결과를 나타내었다. 이는 제안한 표준편차 기반의 분할 기법이 기존 기법들에 비해 보다 정확하게 구간 분할을 수행함을 의미한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 논문의 이론적 배경이 되는 관련 연구를 소개하고, 제3장에서는 전체적인 프레임워크로서, 전처리 과정과 후처리 과정을 제시한다. 제4장은 전처리 과정으로, 기존의 양분할과 다분할 기법을 정형적으로 정의하고 표준편차 최소화 기법을 제안한다. 제5장은 후처리 과정으로, 상용 마이닝 도구에서 출력된 이진 연관규칙을 다시 정량적 연관규칙으로 표현하는 과정을 설명한다. 제6장에서는 제안한 분할 방법에 의한 실험 결과를 설명한다. 마지막으로 제7장에서는 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

이진 속성 기반의 연관규칙: 이진 속성이란 어떤 항목이 트랜잭션에 포함되거나 포함되지 않거나 하는 두 가지 중 하나의 값을 가지는 속성이다. (이에 반해, 정량적 속성은 어떤 항목이 수량 값이나 범위를 갖는 속성이다.) 이러한 이진 속성 기반의 연관규칙을 마이닝하는 가장 대표적인 방법으로서, Agrawal 등[2]의 APRIORI 알고리즘을 들 수 있다. APRIORI 알고리즘은 대용량 트랜잭션 데이타베이스에서 연관규칙을 마이닝 하기 위하여 "빈발항목 집합의 부분집합은 항시 빈발항목이다"는 개념을 사용하였다. APRIORI 알고리즘 이후에 DHP[3], PARTITION[4], DIC[5] 등 많은 알고리즘에 제안되었는데, 이들 알고리즘은 대부분 데이타베이스 스캔 횟수를 줄이거나 후보 집합의 개수를 줄여 성능을 향상시키는데 연구의 목적이 있었다. 예를 들어, DHP의 경우 해시 테이블을 사용하여 2-항목집합들을 효율적으로 관리하는 방법을 제시하였다. 다음으로, PARTITION은 트랜잭션 데이타를 분할하여 후보 항목집합들을 찾아서 입/출력 횟수를 줄이는 방법을 제안하였다. 또한 DIC(Dynamic Itemset Counting)는 모든 부분집합이 빈발할 것으로 예상되면 새로운 후보 항목 집합을 추가하여, 데이타베이스의 스캔 횟수를 줄이는 방법을 제안하였다. 이외에도, 참고문헌 [6]에서는 속성의 계층(taxonomy)을 사용하여 일반화된 연관규칙(generalized association rules)을 마이닝 하였으며, 참고문헌 [7]에서는 사용자 제약(constraint)에 기반하여 연관규칙을 마이닝 하였다. 그리고, 참고문헌 [8,9]에서는 샘플링이나 히스토그램을 사용하여 연관규칙 마이닝의 성능을 향상시켰으며, 참고문헌 [10]에서는 연관성이 없는 항목들을 찾아내어 부정 연관규칙(negative association rules)을 마이닝하는 기법을 제시하였다. 그러나, 지금까지 소개한 연관규칙 마이닝 알고리즘들은 모두 이진 속성을 대상으로 하고 있어, 본 논문에서 다루고자 하는 정량적 속성을 처리할 수 없다.

정량적 속성 기반의 연관규칙: 정량적 속성을 포함하는 트랜잭션 데이타베이스에서 정량적 연관규칙을 마이닝하기 위한 방법으로 여러 연구가 수행되었다[11-15]. 먼저 참고문헌 [11]에서는 정량적 속성을 일정한 범위의 소구간으로 분할한 후, 이웃한 소구간을 병합하면서 정량적 연관규칙을 마이닝하는 방법을 제안하였다. 참고문헌 [12]에서는 밀도(density)가 높은 범위를 판별하여 정량적 연관규칙을 마이닝하는 방법을 제안하였다. 그리고, 참고문헌 [13]에서는 PNSC(Principal Non-negative

Sparse Coding) 알고리즘으로 행렬에서의 고유 벡터를 찾아내는 방법으로, 참고문헌 [14]에서는 최빈수를 포함한 구간을 중심으로 주변 데이타의 발생 빈도를 사용하는 방법으로 정량적 연관규칙을 마이닝 하였다. 그런데 이들 방법은 모두 이진 기반의 연관규칙 마이닝 알고리즘 자체를 변경하여 새로운 알고리즘으로 정량적 연관규칙을 마이닝하는 전략을 취하였다. 결국, 이들 연구는 알고리즘 변경이 어려운 상용 데이타 마이닝 도구에는 적용할 수 없는 문제점이 있다.

이와 같이 알고리즘 자체를 변경하는 연구 이외에 평균(혹은 중앙치)를 사용하여 정량적 속성을 이진 속성으로 변환하는 연구[15]와 동일 깊이 혹은 동일 너비에 기반하여 변환을 수행하는 연구[11]가 있었다. 그리고, 이들 방법은 이진 속성으로의 변환 이후에 기존 연관규칙 마이닝 알고리즘을 그대로 사용할 수 있으므로, 결국 본 논문에서 목적하는 상용 마이닝 도구에의 적용이 가능하다. 따라서, 본 논문에서는 이들 평균 및 중앙치 기반 방법, 동일 너비 및 동일 깊이를 사용하는 방법을 제안하는 프레임워크의 전처리 과정에 있어서의 구간 분할 방법들로 제시한다.

상용 데이타 마이닝 도구: 상용 데이타 마이닝 도구로는 IBM의 Intelligent Miner (I-Miner)[16], SAS의 Enterprise Miner (E-Miner)[17], Silicon Graphics사의 MineSet[18], SPSS의 Clemetine[19] 등의 많은 제품이 사용되고 있다. 그리고 이들 대부분의 상용 도구들은 대표적인 마이닝 기법으로서 연관규칙 마이닝 기능을 제공한다. 본 논문에서는 이들 상용 도구 중에서 SAS의 E-Miner 제품을 실험에 사용한다. SAS의 E-Miner는 연관규칙은 물론 분류(classification), 군집 분석(clustering analysis), 신경망(neural network) 등의 기능을 제공하며, 세계적으로 가장 널리 사용되고 있는 마이닝 제품 중의 하나이다.

3. 제안하는 전체 프레임워크

상용 마이닝 도구를 사용해서 정량적 속성에 대한 연관규칙을 찾기 위해서는 전처리와 후처리의 두 가지 과정이 필요하다. 먼저, 전처리 과정은 상용 마이닝 도구를 사용하여 연관규칙을 마이닝하기 이전의 과정으로서, 수량적 속성을 갖는 원 입력 데이타를 구간으로 분할된 이진 데이타로 변환하는 작업을 수행한다. 다음으로, 후처리 과정은 상용 마이닝 도구를 사용하여 연관규칙을 마이닝한 이후의 과정으로서, 마이닝된 규칙들을 통합하고 이진 속성을 다시 정량적 속성으로 변환하는 작업을 수행한다.

그림 1은 상용 마이닝 도구를 사용해서 정량적 연관규칙을 찾기 위해 본 논문에서 제안하는 전체 프레임워크를 나타낸다. 그림을 보면, 상용 데이타 마이닝 도구를 중심으로 전처리 과정과 후처리 과정을 수행함을 알 수 있다. 전처리 과정은 수량적 속성의 원 입력 데이타를 상용 마이닝 도구가 처리할 수 있는 이진 속성의 데이타로 변환한다. 상용 마이닝 도구는 이렇게 변환된 이진 속성의 데이타를 입력으로 받아 이진 연관규칙들을 생성한다. 다음으로, 그림의 오른편을 보면, 후처리 과정은 상용 마이닝 도구에서 마이닝한 이진 연관규칙을 입력으로 받아 정량적 연관규칙들을 출력한다. 논문에서는 그림 1의 전체 프레임워크하에서 전처리 과정과 후처리 과정을 각각 제안한다. 먼저, 제4장에서는 구간 분할을 정의하고 구간을 분할하는 여러 가지 방법을 제시한다. 다음으로, 제5장에서는 분할된 구간을 다시 통합하는 후처리 과정을 논의한다.

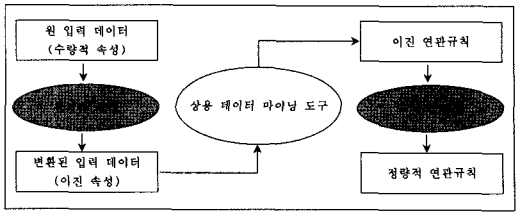


그림 1 상용 마이닝 도구 기반의 정량적 연관규칙 프레임워크

4. 전처리 과정

본 장에서는 정량적 속성의 데이타를 이진 속성 데이타로 변환하는 전처리 과정을 설명한다. 먼저 제4.1절에서는 원 입력 데이타의 정량적 속성을 이진속성으로 표현하기 위한 구간 분할을 정의한다. 다음으로 제4.2절과 제4.3절에서는 기존의 구간 분할 방법을 정형적으로 재정의한다. 마지막으로, 제4.4절에서는 본 논문에서 새롭게 제안하는 표준편차 최소화 기법을 설명한다.

4.1 구간 분할의 정의

정량적 속성의 도메인을 구간들의 집합으로 나눌 수 있다면, 정량적 속성은 구간을 값으로 하는 이진 속성으로 변환할 수 있다. 예를 들어, 도메인이 [0,100]인 속성이 있을 때, 이를 $\{I_1=(0,60),\ I_2=(61,100)\}$의 두 구간으로 나누었다 하자. 이때, 어떤 트랜잭션에서 해당 속성의 값이 30이면, 이 트랜잭션은 구간 I_1을 포함하는 새로운 트랜잭션으로 변환할 수 있다. 이러한 구간 결정의 개념을 정형적으로 표현하기 위해, 구간 분할의 대상이 되는 속성 A_i에 대해 다음 표 1의 표기법을 사용한다.

정의 1. 정량적 속성 A_i의 도메인 $D_i[b_i..d_i]$가 주어졌을 때, D_i를 제시된 방법에 따라 분할하여 구간의 집합 $\mathbb{I}_i$를 생성하는 작업을 구간 분할(domain partition)이라

표 1 분할 대상인 정량적 속성 A_i 에 대한 표기법.

표 기	내 용
$D_i[b_i..d_i]$	속성 A_i의 도메인을 나타내며 최소값으로 b_i를 최대값으로 d_i를 갖는다.
n_i	도메인 D_i가 분할된 구간의 개수
$I_{i,j}$	도메인 D_i가 분할된 구간 중에서 j-번째 구간
$\mathbb{I}_i$	도메인 D_i가 분할한 구간의 집합 = $\{I_{i,j} \mid 1 \le j \le n_i\}$

정의한다. □

다음 예제 1은 정량적 속성의 도메인을 구간들의 집합으로 구간 분할하는 과정을 설명한다.

예제 1. 표 2는 맥주(A_1)와 온도(A_2)의 두 정량적 속성에 대해 구간 분할을 수행한 예이다. 표에서 보듯이, 속성 맥주(A_1)에 대해서는 도메인 $D_1[1..54]$를 다섯 개의 구간 $I_{1,1} \sim I_{1,5}$로 나누었음을 알 수 있다. 즉, 맥주에 대해서 구간들의 집합 $\mathbb{I}_1 = I_{1,1}, I_{1,2}, .., I_{1,5}$가 생성되었다. 다음으로, 속성이 온도($A_2$)인 경우는 도메인 $D_2[3..34]$가 세 개의 구간 $I_{2,1}$, $I_{2,2}$, $I_{2,3}$로 나누어졌음을 알 수 있다. □

정량적 속성에 대한 구간 분할이 이루어진 후에는 이 속성을 포함하는 각 트랜잭션을 정량적 속성 값 대신 분할된 구간을 포함하는 새로운 트랜잭션으로 변환한다. 이를 위해서는 트랜잭션에 포함된 해당 속성 값이 분할된 구간 집합의 어떤 구간으로 매핑되는지 결정해야 한다. 예를 들어, 속성 A_i가 분할된 구간 집합 $\mathbb{I}_i = I_{1,1} = (0,60)$, $I_{1,2} = (61,100)$일 때, 트랜잭션 T에 포함된 속성 A_i의 값이 50이라면, 변환된 트랜잭션 T' 은 구간 $I_{1,1}$을 갖는 것으로 나타난다. 그런데, T의 속성 A_i의 값이, 단일 값 50이 아닌 범위 값, 예를 들어 (30,65)를 가질 경우에도 적절한 변환이 필요하다. 이와 같이 트랜잭션의 속성 값이 단일 값 혹은 범위 값을 갖는 경우 모두를 처리하기 위해 이진 변환 규칙을 다음과 같이 정의한다.

정의 2. 트랜적션 T의 속성 A_i가 단일 값 x를 갖는다면, T가 이진 변환된 트랜잭션 T' 은 $x \in I_{i,j}$인 이진 속성 $I_{i,j}$를 포함한다. 그리고, 트랜잭션 T의 속성 A_i가 범위 값 (s,e)를 갖는다면 T가 이진 변환된 트랜잭션 T' 은 $(s+e)/2 \in I_{i,j}$ 인 이진 속성 $I_{i,j}$를 포함한다. □

다음 예제 2는 정의 2의 이진 변환에 의해 트랜잭션이 변환되는 과정을 설명한다.

예제 2. 예제 1에서 사용한 표 2와 같이 맥주와 온도에 대한 구간 분할이 수행되었다 하자. 그리고 어떤 트랜잭션 T가 {맥주=18, 온도=[15,18]}과 같이 맥주와 온도에 대한 속성 값을 가진다 하자. 그러면, T가 이진 변환된 트랜잭션 T'은 $I_{1,2}, I_{2,2}$로 나타난다. 그 이유는 맥주 속성의 경우 단일 값인 18이 구간 $I_{1,2}$에 각각 포함되고, 온도 속성의 경우 범위 값인 [15,18]의 평균인 $(15+18)/2 = 16.5$가 $I_{2,2}$에 포함되기 때문이다. 궁극적으로, 이진 변환에 의해 정량적 속성을 갖는 트랜잭션 T ={맥주=18, 온도=[15,18]}는 이진 속성의 트랜잭션 $T' = I_{1,2}, I_{2,2}$로 변환됨을 알 수 있다. □

정량적 속성의 구간분할 결과는 분할 기법에 따라 달라질 수 있다. 이에 따라, 다음의 제4.2절, 제4.3절, 그리고 제4.4절에서는 전체 구간을 두 구간 혹은 여러 구간으로 분할하는 방법을 제시한다.

4.2 양분할 기법

본 절에서는 구간을 두 부분으로 나누는 양분할 기법으로 평균과 중앙치에 의한 분할 기법을 설명한다. 본 절의 양분할 기법은 참고문헌 [15]에서 제안된 것으로 제4.1절의 구간 분할 개념을 사용하여 정형적으로 재정의한다.

평균에 의한 구간 양분할: 이 방법은 주어진 속성에 대해 모든 트랜잭션을 대상으로 전체 평균을 구하여, 이를 중심으로 평균 미만과 평균 이상의 두 구간으로 분할하는 방법이다. 평균에 의한 구간 양분할 방법을 정형적으로 정의하면 다음과 같다.

정의 3. 정량적 속성 A_i를 포함하는 트랜잭션 집합을 $\mathbb{T} = T_1, T_2, .., T_t$라 하고, 각 트랜잭션 T_j에서 속성 A_i가 단일 값인 경우 $x_{i,j}$, 범위 값인 경우 $(s_{i,j}, e_{i,j})$를 갖는다 하자. 그러면, 평균에 의한 구간 양분할은 속성 A_i의 도메인 $D_i[b_i..d_i]$을 평균 m을 중심으로 $[b_i..m)$과 $[m..d_i]$의 두 구간으로 나눈다. 여기에서, 평균 m은 $\Sigma v_{i,j}/t$이고,

표 2 구간 분할 과정에 따른 이진 변환 규칙의 예

표기	A_1(맥주)	A_2(온도)
$D_i[b_i..d_i]$	$D_1 = [1..54]$	$D_2 = [3..34]$
n_i	$n_1 = 5$	n_2=3
$\mathbb{I}_i$	$\{I_{1,1}$=(1,11), $I_{1,2}$=(12,24), $I_{1,3}$=(25,35), $I_{1,4}$=(36,50), $I_{1,5}$ =(51,54)}	$\{I_{2,1}$=(3,12), $I_{2,2}$=(13,27), $I_{2,3}$= (28,34)}

$v_{i,j}$는 $x_{i,j}$ 혹은 $(s_{i,j}+e_{i,j})/2$ 이다. □

다음 예제 3은 평균에 의한 구간 분할의 예를 나타낸다.

예제 3. 표 2와 같이 속성 맥주(A_1)의 도메인이 $D_1[1..54]$이고, 맥주를 포함하는 세 개의 트랜잭션 $T_1(A_1)=18$, $T_2(A_1)=(20,30)$, $T_3(A_1)=(5,15)$이 있다 하자. 그러면, 정의 3에 의해 평균 m은 $(18+(20+30)/2+(5+15)/2)/3=17.7$로 계산된다. 따라서, 도메인 $D_1[1..54]$는 평균에 의한 분할에 의해 [1..17.7)와 [17.7..54]의 두 구간으로 나누어진다. □

중앙치에 의한 구간 양분할: 중앙치란 주어진 n개의 값들을 정렬했을 때 중앙, 즉 $n/2$번째에 해당하는 값이다[20]. 예를 들어, {2,3,5,7,11}이 주어졌을 때의 중앙치는 중앙에 있는 5가 되며, {2,3,5,7,11,13}의 경우는 중앙의 두 값인 5와 7의 중간 값인 $(5+7)/2=6$이 된다. 중앙치 기반의 구간 양분할 방법은 주어진 속성에 대해 모든 트랜잭션을 대상으로 전체 중앙치를 구하여, 이를 중심으로 중앙치 미만과 중앙치 이상의 두 구간으로 분할하는 방법이다. 중앙치에 의한 구간 양분할 방법을 정형적으로 정의하면 다음과 같다.

정의 4. 정량적 속성 A_i를 포함하는 트랜잭션 집합을 $\mathbb{T}=T_1,T_2,..,T_t$이라 하고, 각 트랜잭션 T_j에서 속성 A_i가 단일 값인 경우 $x_{i,j}$, 범위 값인 경우 $(s_{i,j},e_{i,j})$를 갖는다 하자. 이때, 중앙치 m은 단일 값인 경우 $x_{i,j}$, 구간 값인 경우 $(s_{i,j}+e_{i,j})/2$의 값을 정렬한 후 다음 공식 (1)과 같이 계산한다.

$$m=\begin{cases} \left\lceil \frac{t}{2} \right\rceil \text{번째 값} & , t=\text{홀수} \\ \dfrac{\left(\frac{t}{2}\right)\text{번째 값} + \left(\frac{t}{2}+1\right)\text{번째 값}}{2} & , t=\text{짝수} \end{cases} \quad (1)$$

그러면, 중앙치에 의한 구간 양분할은 속성 A_i의 도메인 $D_i[b_i..d_i]$을 $[b_i..m)$과 $[m..d_i]$의 두 구간으로 분할한다. □

다음 예제 4는 중앙치에 의한 구간 분할의 예를 나타낸다.

예제 4. 표 2에서와 같이 속성 맥주(A_1)의 도메인이 $D_1[1..54]$이고, 맥주를 포함하는 네 개의 트랜잭션이 표 3과 같다 하자. 트랜잭션 집합 $\mathbb{T}$의 중앙치를 구하기 위해 먼저 속성의 평균 값들을 정렬하면 {10, 18, 25, 42.5}가 된다. 따라서, 중앙치 m은 $(18+25)/2=21.5$로 계산된다. 그리고, 도메인 $D_1[1..54]$는 중앙치에 의한 분할에 의해 [1..21.5)와 [21.5..54]의 두 구간으로 나누어진다. □

4.3 구간의 다분할 방법

표 3 트랜잭션 T_j의 A_i 속성 값과 평균

트랜잭션(T_j)	맥주 속성 값($T_j(A_1)$)	평 균
T_1	18	18
T_2	(20,30)	25(=(20+30)/2)
T_3	(5,15)	10(=(5+15)/2)
T_4	(35,50)	42.5(=(35+50/2))

본 절에서는 전체 구간을 여러 구간으로 나누는 다분할 방법을 제시한다. 이때, 구간의 수는 사용자에 의해 주어지거나, 참고문헌 [11]의 결과 등에 의해 계산된다고 가정한다. 즉 구간의 수가 주어졌을 때, 전체 구간을 분할하는 방법을 제시한다. 본 절에서 제시하는 동일 너비와 동일 깊이에 의한 구간 다분할 방법은 참고문헌 [11]에서 이미 소개된 것으로서, 본 논문에서는 제4.1절의 구간 분할 개념을 사용하여 정형적으로 재정의한다. 우선, 동일 너비에 의한 구간 분할 기법을 정의하면 다음과 같다.

정의 5. 구간의 수가 n이라 할 때, 정량적 속성 A_i의 도메인 $D_i[b_i..d_i]$는 동일 너비에 의한 구간 분할에 의해 다음 공식 (2)와 같이 n개의 구간으로 나누어진다.

$$[b_i..(b_i+\delta)),\ [(b_i+\delta)..(b_i+2\delta)),...,[(b_i+(n-1)\delta)..d_i],$$
$$\text{where } \delta=\frac{d_i-b_i}{n}. \quad (2)$$

□

다음으로, 정의 6은 동일 깊이에 의한 구간 분할을 나타낸다.

정의 6. 구간의 수가 n이라 하고, 정량적 속성 A_i의 값 y에 대한 트랜잭션의 빈도수가 $f(y)$[1]로 주어졌을 때, 속성 A_i의 도메인 $D_i[b_i..d_i]$는 동일 깊이에 의한 구간 다분할에 의해 다음 공식 (3)과 같이 n개의 구간으로 나누어진다.

$$[y_0..y_1),[y_1..y_2),...,[y_{n-1}..y_n], \quad (3)$$
$$\text{where } \int_{y_i}^{y_{i+1}} f(y)dy \simeq \frac{1}{n}\int_{b_i}^{d_i} f(y)dy,$$
$$0 \le i \le n-1,\ y_0=b_i,\ \text{and } y_n=d_i$$ □

다음 예제 5는 동일 너비와 동일 깊이로 구간 다분할을 수행한 예이다.

예제 5. 맥주 속성(A_1)을 포함하는 트랜잭션들(T_j)이 있고, A_1의 도메인 $D_1[1..30]$을 다섯 개의 구간으로 나누려 한다. 우선, 동일 너비로 구간 분할을 할 경우, 정의 5에 의해 구간의 간격 δ를 5.8((30-1)/5)로 하여 $D_{1,1}[1..6.8)$, $D_{1,2}[6.8..12.6)$, $D_{1,3}[12.6..18.4)$, $D_{1,4}[18.4..$

1) 빈도수 $f(y)$는 전체 데이타베이스를 한 번 스캔하거나, 히스토그램[21] 등을 유지하여 구할 수 있다.

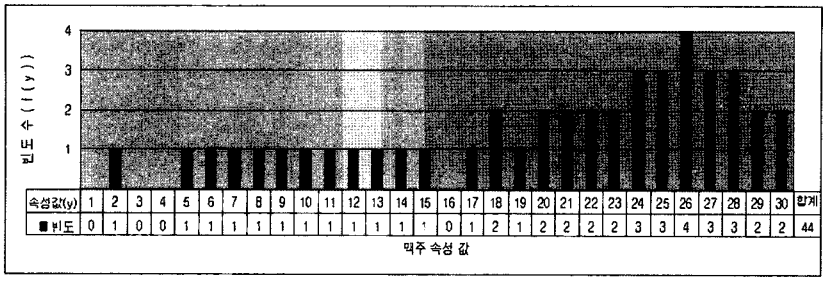


그림 2 맥주 속성(A_1)값의 트랜잭션 빈도 수($f(y)$)

24.2), $D_{1,5}[24.2..30]$의 다섯 구간으로 나눈다. 다음으로, 동일 깊이 구간 분할을 위한 빈도 수 $f(y)$가 그림 2로 주어졌다고 하자. 그러면 구간별 평균 트랜잭션의 수는 $\frac{1}{n}\int_{1}^{30} f(y)dy = 8.8$로 구할 수 있다. 따라서, 각 구간이 대략 8~9개의 트랜잭션을 갖도록 구간 분할을 수행한다. 그러면, 그 결과는 $D_{1,1}[1,..,11)$, $D_{1,2}[11,..,20)$, $D_{1,3}[20,..24)$, $D_{1,4}[24,..,27)$, $D_{1,5}[27,..,30]$이 된다. □

4.4 표준편차 최소화에 의한 구간 분할

본 절에서는 속성 값의 표준편차(혹은 분산)를 구간 분할에 사용하는 방법을 제안한다. 제4.2절과 제4.3절에서 제시한 구간 분할 방법들은 속성 값의 변화 분포를 고려하지 않아, 구간 분할이 바르게 이루어지지 않을 가능성이 높다. 이 문제를 해결하기 위하여 본 절에서는 연속된 속성 값의 변화가 심하지 않으면 해당 속성 값들은 유사한 특징을 가지므로 동일한 구간에 포함시키고, 그 차이가 크다면 다른 구간으로 분할한다는 직관에 기반한 새로운 구간 분할 방법을 제안한다. 속성 값의 표준편차를 사용하여 구간을 분할하는 방법을 다음과 같이 정형적으로 정의한다.

정의 7. 주어진 속성 A_i의 도메인이 n개의 구간 $\{I_{i,1}, I_{i,2},.., I_{i,n}\}$으로 분할되었다 하자. 그리고, 구간 $I_{i,j}(1 \le j \le n)$에서 속성 A_i의 값들을 대상으로 계산한 표준편차가 $\sigma(I_{i,j})$라 하자. 그러면, 분할된 모든 구간 $I_{i,j}$에 대한 표준편차의 합은 $\sum_{j=1}^{n}\sigma(I_{i,j})$이다. 표준편차 최소화 기법은 이러한 표준편차의 합 $\sum_{j=1}^{n}\sigma(I_{i,j})$가 최소화 되도록 구간 집합 $\{I_{i,j}\}$를 결정하는 구간 분할이다. □

다음 예제 6은 표준편차 기반의 구간 분할을 수행하는 예이다.

예제 6. 맥주 속성 A_1의 도메인 $D_1[1..30]$을 세 개의 구간으로 나누려한다. 이때, 두 방법 A와 B에 의해 표 4와 같이 구간 분할이 이루어졌다 하자. 그러면, 방법 A의 표준편차의 합은 15.0인 반면에 방법 B의 표준편차의 합은 11.0이 된다. 따라서, 표준편차 최소화 기법에서는 합이 최소가 되는 방법 B를 구간 분할에 이용한다. □

그런데, 표준편차의 합을 최소화하는 구간 집합을 찾는 문제는 지수 함수의 계산 복잡도를 갖는다. 즉, 속성(도메인)의 이산 값 개수가 p이고, 분할한 구간의 수가 $n(\ll p)$이라면, 대략 $\binom{p}{n}$개의 구간 분할 방법이 가능하다. 그리고, 이들 $\binom{p}{n}$개 모든 경우의 수에 대해서 $\Sigma\sigma(I_j)$을 최소화하는 구간 집합을 선택해야 한다. 그런데, 일반적으로 속성의 이산 개수 p는 구간의 수 n에 비해 훨씬 크므로, $\binom{p}{n}$인 모든 경우의 수를 고려하는 방법은 실용적으로는 사용할 수 없다. 따라서, 본 논문에서는 휴리스틱 알고리즘을 사용하여 이 문제를 해결한다. 제안하는 휴리스틱 알고리즘은 $\Sigma\sigma(I_j)$를 최소화하는 최적의 해답 대신에, 국부적 최적 값을 찾는 방법으로 가능한 경우의 수를 크게 줄여 근사적 해답을 찾는 방법이다.

그림 3은 표준편차 최소화 기법을 사용하여 구간 분할을 수행하는 알고리즘이다. 이 알고리즘은 속성 A_i의 도메인 $D_i[b_i..d_i]$에 대해 표준편차 기반의 구간 분할을 수행하는 근사 해답을 찾는다. 그림 3의 라인 (1)은 초기 분할을 수행하는 것으로, 우선 도메인 $D_i[b_i..d_i]$를 동일 너비로 나눈다. 라인 (1)의 초기 분할은 동일 너비 대신 동일 깊이를 사용할 수도 있다. 라인 (4)~(6)에서

표 4 표준편차 기반의 구간 분할 예제.

방법	구간 I_1	$\sigma(I_1)$	구간 I_2	$\sigma(I_2)$	구간 I_3	$\sigma(I_3)$	$\Sigma\sigma(I_i)$
A	[1..11)	7.0	[11..21)	1.0	[21..30]	7.0	15.0
B	[1.. 6)	4.0	[6..26)	3.0	[26..30]	4.0	11.0

는 각 구간의 경계 l_j를 제한된 범위인 (l_{j-1}, l_{j+1})에서 이동시키면서 표준편차의 합을 구하되, 그 합을 최소로 하는 새로운 l_j를 구한다. 이 과정은 (l_{j-1}, l_{j+1})에서 l_j의 국부적 최적값을 찾는 과정이다. 이러한 과정은 l_0와 l_n을 제외한 모든 l_j에 대해서 반복 수행한다. 라인 (4)~(6)의 과정은 모든 l_j에 변화가 없을 때까지 반복된다 (라인 (3)~(7)).

제안한 알고리즘의 계산 복잡도는 다음과 같다. 그림 3의 라인 (5)에서 l_j에 대하여 l_{j-1}과 l_{j+1}사이의 모든 속성 값을 대입하여 각 구간의 표준편차를 구한다. 여기서, l_j값들이 취할 수 있는 값은 구간$[l_0 .. l_n]$의 이산 값들로서 p개가 되고, 결국 라인 (5)의 계산 복잡도는 $O(p)$이다. 이때 계산 복잡도의 단위 연산은 표준편차를 한번 계산하는 연산으로 정한다. 만일, 표준편차 계산을 단위 연산으로 하지 않고, 덧셈 연산을 단위 연산으로 한다면, 이때의 계산 복잡도는 $O(p^2)$이 된다. 다음으로, 라인 (4)~(6)의 for 루프는 n번 실행되므로 계산 복잡도는 $O(pn)$이 된다. 만일 라인 (3)~(7)의 repeat 루프가 r번 수행된다 하면, 제안한 알고리즘의 전체의 계산 복잡도는 $O(rpn)$이 된다. 실험 결과 r은 p에 비하여 훨씬 작은 값으로 나타났으며, 따라서 $O(rpn)$은 $O\left(\binom{p}{n}\right)$에 비해 계산 복잡도를 크게 줄인 것이라 할 수 있다.

For a given domain $D_i[b_i .. d_i]$ and the number p of distinct attribute values.
(1) Do the equi-width partition as follows : //or equi-depth partition
$[l_0 .. l_1), [l_1 .. l_2), [l_2 .. l_3), ..., [l_{n-1} .. l_n]$,
where $l_0 = b_i$, $l_n = d_i$, and $l_j = b_i + j \cdot \delta$ $(1 \le j \le n-1)$.
(2) Compute $\sum_{j=1}^{n} \sigma([l_{j-1} .. l_j])$
(3) **repeat**
(4) **for** $j=1$ to $n-1$ **do**
(5) Set l_j as the value in the range (l_{j-1}, l_{j+1})
such that $\sum_{j=1}^{n} \sigma([l_{j-1} .. l_j])$ becomes minimum;
(6) **end for**
(7) **until** no change

그림 3 표준편차 최소화 기법의 구간 분할 알고리즘

제안한 표준편차 최소화 기법은 제4.2절의 양분할 및 제4.3절의 다분할 모두에 적용이 가능하다. 이에 대한 실험결과는 제6장에서 설명한다.

5. 후처리 과정

본 장에서는 상용 마이닝 도구에서 출력된 이진 연관규칙들을 입력으로 받아, 후처리를 거쳐 정량적 연관규칙들을 생성하는 과정을 설명한다. 후처리 과정으로는, 먼저 제5.1절에서 상용 마이닝 도구에서 출력된 이진 연관규칙들을 대상으로 연속된 구간을 통합하는 과정을, 제5.2절에서 이진 연관규칙을 정량적 연관규칙으로 표현하는 과정을 각각 설명한다.

5.1 연속 구간의 통합

본 절에서는 상용 마이닝 도구에서 출력된 이진 연관규칙들을 대상으로 연속된 구간을 통합하는 방법을 설명한다. 구간 통합이 필요한 이유는 정량적 속성을 이진 속성으로 강제로 변환하였기 때문이다. 예를 들어, 상용 데이타 마이닝 도구에서 표 5와 같은 이진 연관규칙들을 출력했다고 하자. 이때, 이진 연관규칙 1의 분할 구간 $I_{2,1}$, $I_{2,2}$는 연속된 구간이므로 구간 통합이 가능하다. 또한, 규칙 2의 $I_{5,3}, I_{5,4}, I_{5,5}$는 연속된 구간이므로 하나의 구간으로 통합할 수 있다.

표 5 마이닝된 이진 속성의 연관규칙

구분	이진 연관규칙
연관규칙 1	$I_{1,3}, I_{2,1}, I_{2,2} \Rightarrow A, I_{3,3}$
연관규칙 2	$B, I_{4,3} \Rightarrow I_{5,3}, I_{5,4}, I_{5,5}, I_{6,3}$

마이닝된 이진 연관규칙에서 구간을 통합하는 방법은 다음과 같다.

규칙 1. 이진 연관규칙 $X \Rightarrow Y$ 에서, X 혹은 Y에 속성 A_i가 분할된 연속 구간들 $I_{i,j}, I_{i,(j+1)}, .., I_{i,k}$가 있다면, 이들 연속된 구간은 속성 A_i에 대한 하나의 구간으로 통합하고, 이를 $I_{i,(j,k)}$로 표기한다. □

다음 예제 7은 이진 연관규칙에서의 연속된 분할 구간을 규칙 1에 기반하여 통합하는 과정을 설명한다.

예제 7. 표 5의 연관규칙 1과 2가 있다 하자. 규칙 1에 의해 속성 A_2에서 분할된 $I_{2,1}, I_{2,2}$는 연속적인 구간을 나타내므로, 이들을 통합하여 새로운 구간 $I_{2,(1,2)}$로 나타낸다. 즉, 이진 연관규칙 $I_{1,3}, I_{2,1}, I_{2,2} \Rightarrow B, I_{3,3}$는 $I_{1,3}, I_{2,(1,2)} \Rightarrow B, I_{3,3}$로 변환된다. 마찬가지로, 규칙 2는 $B, I_{4,3} \Rightarrow I_{5,(3,5)}, I_{6,3}$으로 변환된다. □

5.2 이진 연관규칙을 정량적 연관규칙으로 변환

본 절에서는 연속된 구간의 통합을 거쳐 최종 선택된 이진 연관규칙들을 정량적 연관규칙으로 다시 변환하는 과정을 설명한다. 이 과정은 매우 단순하게(straight-forward) 진행되므로, 간략한 예를 들어 설명한다. 속성 A_i와 도메인 D_i가 있으며 이들의 구간이 표 6과 같이 구간 $I_{i,j}$로 분할되었다 하자. 이때, 이진 연관규칙 $I_{1,1}, I_{2,1} \Rightarrow I_{3,1}$이 생성되었다면, 이는 표 6의 구간 정의에 의해 $D_1[1,10)$, $D_2[1,6) \Rightarrow D_3[1,3)$의 정량적 연관규칙으로 다시 표현할 수 있다.

6. 성능 평가

본 장에서는 제안한 방법의 구현 내용과 실험 결과를 설명한다. 제6.1절에서는 실험을 수행한 데이타 및 환경

표 6 이진 속성과 정량적 속성의 매핑 예.

속성A_i	도메인 $D_i[b_i..d_i]$	구간 $I_{i,j}$			
A_1	$D_1[1..20]$	$I_{1,1}$=[1,10)	$I_{1,2}$=[10,20]		
A_2	$D_2[1..30]$	$I_{2,1}$=[1, 6)	$I_{2,2}$=[7,14)	$I_{2,3}$=[15,20)	$I_{2,4}$=[21,30]
A_3	$D_3[1..12]$	$I_{3,1}$=[1, 3)	$I_{3,2}$=[3, 8)	$I_{3,3}$=[8,12]	

을 설명하고, 제6.2절에서는 실험 결과를 설명한다.

6.1 실험 환경 및 데이타

실험 데이타로는 참고문헌 [2]에서 처음 제시되어, 연관규칙 마이닝의 트랜잭션 데이타로 가장 널리 사용되는 합성 데이타(synthetic data)[2,11,22]를 이용하였다. 그런데, 이 합성 데이타는 이진 속성만을 지원하므로, 본 논문의 실험을 위해서는 정량적 속성을 가지도록 변환하여야 한다. 즉, 원래의 합성 데이타는 모든 속성이 이진 데이타이므로, 일부 속성에 대해서는 정량적 특성을 부여하여야 한다. 본 논문에서는 정량적 속성이 가지는 값의 범위가 0, 1, 2, ..., 100 중의 이산 값 하나를 갖도록 설정하였다. 이때, 각 이산 값 ℓ ($0 \le \ell \le 100$)은 다음 공식 (4)의 중요도 $w(\ell)$에 따라 트랜잭션 개수를 배정하였다.

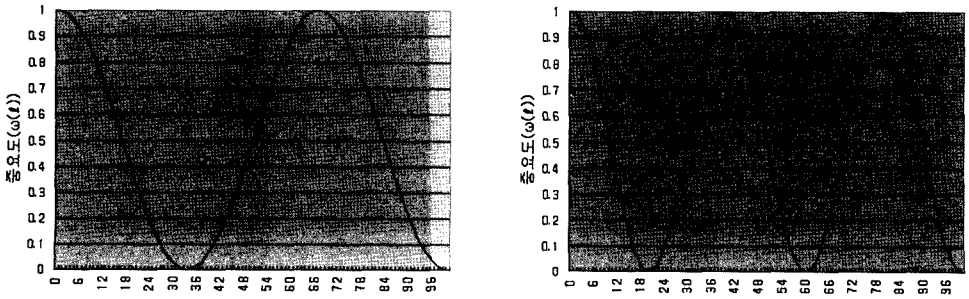
$$w(\ell) = \frac{1}{2}\cos(f \times \frac{\ell}{100} \times \pi) + \frac{1}{2} \qquad (4)$$

공식 (4)에서, f는 극값이 나타나는 주기를 나타낸다.

그림 4는 f를 달리한 경우의 중요도 $w(\ell)$의 그래프를 나타낸다. 그림에서 가로축은 0~100 사이의 이산 속성 값을 나타내고, 세로축은 각 속성 값에 대한 중요도를 나타낸다. 그림 5에서와 같이, 속성 값 ℓ의 중요도는 $w(\ell)$로 구해지므로, 전체 도메인 (0, 1, ..., 100)에서의 ℓ의 상대적 빈도는 $w(\ell)/\sum_{i=0}^{100} w(i)$이 된다. 이때, 해당 정량적 속성이 나타나는 트랜잭션 수를 N이라 하면, 속성 값 ℓ의 실제 빈도 수는 $(w(\ell)/\sum_{i=0}^{100} w(i)) \cdot N$ 으로 구할 수 있고, 이렇게 구해진 빈도수만큼 속성 값 ℓ이 트랜잭션에 나타나도록 설정하였다. 예를 들어, 그림 4(a)에서 속성 값 0의 중요도는 1로서, 속성 0은 실제로 $\frac{1}{\sum_{i=0}^{100} w(i)} = \frac{1}{50.5}$ 의 상대적 빈도를 가지게 된다. 이때, 해당 속성이 5000번 나타난다면, 속성 값 0은 5000× $\frac{1}{50.5}$, 즉 100개의 트랜잭션에 나타나도록 설정하였다.

실험은 양분할 기법과 다분할 기법으로 나누어 수행하였다. 먼저 양분할의 경우 1) 평균, 2) 중앙치, 3) 표준편차 최소화의 세 가지에 대해 실험을 수행하였다. 다음으로, 다분할의 경우 동일 너비와 동일 깊이에 대해 표준편차 최소화 기법을 사용하는 경우와 그렇지 않은 경우를 비교하였다. 이와 같이 각 방법으로 구간 분할을 수행한 후에는 상용 마이닝 도구인 SAS Enterprise Miner[23,24]를 사용하여 이진 연관규칙들을 찾아내었다. 그런 이후에 제5장에서 제시한 후처리를 통하여 각 방법에 대한 최종적인 정량적 연관규칙을 생성하였다.

제4장에서 제시한 분할 방법을 평가하기 위한 척도로는 1) 마이닝된 연관규칙의 수, 2) 정량 속성 값의 범위 크기를 비교하였다. 정량적 연관규칙 마이닝의 기존 연구들에서는 마이닝된 규칙을 평가하기 위하여 각 연구마다 고유의 흥미도(interestingness)를 주관적으로 정의하고, 이 흥미도를 평가 척도로서 사용하였다. 이와 같이 각 연구마다 제시된 주관적인 흥미도는 객관적으로 적용되기 어려운 문제점이 있으므로, 본 논문에서는 앞서의 두 가지 사항을 평가 척도를 사용한다. 두 가지 척도를 사용한 이유와 구하는 방법은 다음과 같다.



(a) f = 3 (b) f = 5

그림 4 극 값의 주기 f에 따른 중요도 $w(l)$의 그래프

- 연관규칙의 개수: 많은 연관규칙을 찾아낸 방법이 우수하다고 판단한다. 이는 동일한 조건인 경우 보다 많은 규칙을 찾아낼수록 구간 분할이 정확히 수행되었다고 생각할 수 있기 때문이다. 각 방법에 대한 연관규칙의 개수는 최종 결과 규칙을 카운트하여 구할 수 있다.
- 정량 속성 값의 범위 크기: 마이닝된 연관규칙에 포함된 정량적 속성 값이 나타나는 범위의 평균 크기를 비교하여, 크기가 클수록 더 우수하다고 판단한다. 이는 규칙 내에 포함된 정량적 속성 값의 범위가 클수록 보다 일반화된(넓은 범위를 고려한) 규칙을 찾아낸다고 볼 수 있기 때문이다. 각 방법에 대한 연관규칙의 범위는 각 규칙에 포함된 속성 값의 범위를 모두 더한 후 평균을 내어 구할 수 있다.

마지막으로, 표 7은 실험을 수행한 하드웨어 및 소프트웨어 환경을 나타낸다. 그리고, 실험에서는 (전체 속성 중에서) 정량적 속성의 비율을 5%, 10%, 20%로 달리하면서, 극값의 주기 f를 3, 5, 7로 달리하면서, 트랜잭션 개수(데이타베이스 크기)를 2만, 4만, 8만개로 달리하면서 연관규칙 개수와 속성 값 범위를 각각 측정하였다. 이때 이들 파라메터에 대한 디폴트(default) 값으로는 각각 10%, f=5, 4만을 사용하였다.

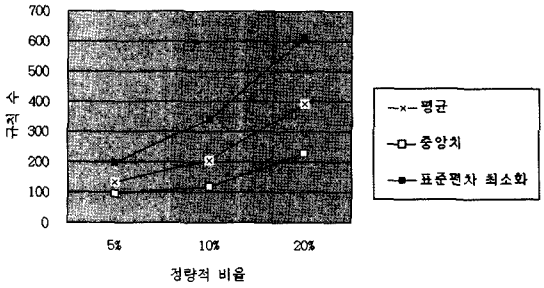
6.2 실험 결과

실험 1) 양분할 실험 결과

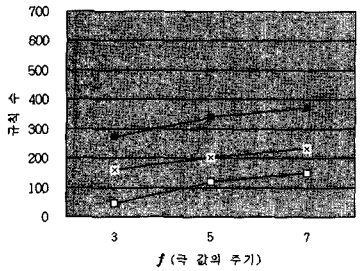
표 7 구현 및 실험 환경

구분		사양
소프트웨어	운영체제	Window XP
	프로그램 언어	Borland Delphi 5, C
	마이닝 도구	SAS E-Miner 9.2
하드웨어	CPU	Intel Pentium Ⅳ 3.0Gz
	RAM	1 GB
	HDD	30 GB

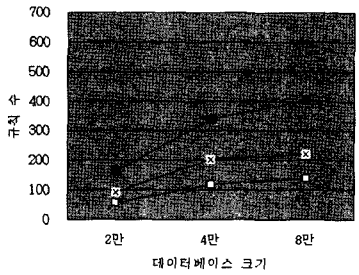
그림 5는 양분할에 의한 실험 결과로서, 평균, 중앙치, 표준편차 최소화의 세 가지 기법에 대해 측정한 연관규칙의 개수를 나타낸다. 그림 5(a)는 정량적 속성 비율은, 그림 5(b)는 극 값의 주기 f를, 그림 5(c)는 트랜잭션 개수를 각각 달리하면서 실험한 결과이다. 그림 5의 실험 결과를 보면, 표준편차 최소화 기법이 평균이나 중앙치 기법보다 결과가 우수함을 알 수 있다. 즉, 표준편차 최소화 기법이 다른 두 기법에 비해 보다 많은 연관규칙을 찾아냄을 알 수 있다. 이는 제안한 표준편차 최소화 기법이 속성 값의 밀집 지역이나 희소 지역을 평균 및 중앙치 기법보다 명확히 구분하기 때문이다. 그리고 평균 기법이 중앙치 기법에 비해 우수한 결과를 보이는데, 이는 중앙치 기법의 경우 하나의 구간을 트랜잭션 개수가 동일한 두 개의 구간으로 나누는 반면에, 평균 기법의 경우 나누어진 두 구간의 트랜잭션 개수가 제각기 다르기 때문이다. 즉, 중앙치 기법에 의해 분할된 두



(a) 정량적 속성 비율에 따른 규칙 수



(b) f에 따른 규칙 수



(c) 트랜잭션 수에 따른 규칙 수

그림 5 양분할에 의해 마이닝된 연관규칙의 개수

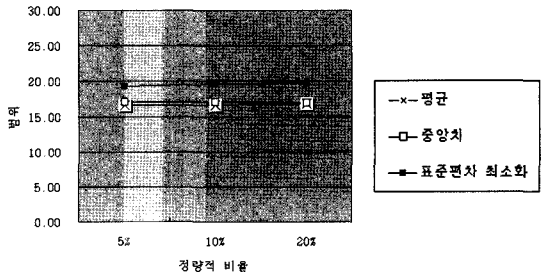
구간은 모두가 빈발하지 않을 가능성이 높은 반면에, 평균 기법에 의한 두 구간 중 하나는 빈발할 가능성이 비교적 높기 때문이다. 그림 5의 실험 결과를 요약하면, 표준편차 최소화 기법은 평균 기법에 비해 평균 57%, 중앙치 기법에 비해 평균 158% 많은 연관규칙을 찾는 것으로 나타났다.

그림 5를 보면, 세 가지 실험 결과 모두 가로축의 값이 증가함에 따라 연관규칙 개수가 증가함을 알 수 있다. 우선, 그림 5(a)에서 정량적 속성의 비율이 증가할수록 마이닝된 규칙의 수가 증가한다. 이는 정량적 속성 비율이 높을수록 생성되는 구간의 개수가 많아지고, 이는 결국 연관규칙 마이닝에 있어서 항목 수가 증가하는 결과를 가져오기 때문이다. 다음으로, 그림 5(b)에서 극값의 주기 f가 증가할수록 연관규칙 개수가 증가한다. 이는 f가 클수록 하나의 구간에 포함되는 트랜잭션 개수가 평균적으로 많아지고, 이에 따라 여러 구간이 연관규칙 생성에 사용되기 때문이다. 마지막으로, 그림 5(c)에서는 트랜잭션 개수가 많을수록 더 많은 연관규칙이 마이닝되고 있다. 이는 마이닝에 사용되는 트랜잭션의 개수가 많아질수록 보다 정확한 마이닝이 이루어지고(여러 항목이 규칙 생성에 참여하고), 이에 따라 연관규칙 개수가 증가하기 때문이다.

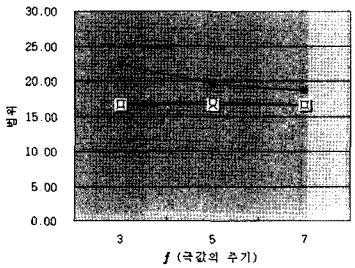
다음으로, 그림 6은 평균, 중앙치, 표준편차 최소화의 세 가지 기법 각각에 의한 정량적 속성 범위를 나타낸다. 앞서 설명한 바와 같이, 각 연관규칙에 포함된 정량적 속성의 범위 평균을 실험 결과로 나타낸 것이다. 그림 6을 보면, 표준편차 최소화 기법의 범위가 평균 및 중앙치 기법에 비해 크게 나타남을 알 수 있다. 이는 표준편차를 최소화하는 과정이 결국 구간을 넓게(밀도가 높은 값들은 가능한 이어서 넓게, 밀도가 낮은 값들도 마찬가지로 가능한 이어서 넓게) 분할하는 효과를 나타내기 때문이다. 그런데, 그 차이는 그림 5의 연관규칙 개수보다는 크게 줄었음을 알 수 있다. 또한 평균과 중앙치의 경우 정량적 속성의 범위 차이가 거의 없음을 알 수 있다. 이는 범위의 합이 아닌 평균을 사용했기 때문이다. 즉, 각 규칙에 포함된 정량적 속성의 범위를 구하여 합한 후, 이들을 규칙의 개수로 나누어 평균을 취하였기 때문이다. 그림 6에서 정량적 속성 비율, 극 값의 주기, 트랜잭션 개수의 변화에 따른 정량적 범위의 변화가 적은 이유도 이와 같이 평균을 취하였기 때문이다. 그림 6의 실험 결과를 종합하면, 제안한 표준편차 최소화 기법은 평균 기법에 비해 평균 17%, 중앙치 기법에 비해 평균 15% 정량적 속성 범위를 늘린 것으로 나타났다.

실험 2) 다분할 실험 결과

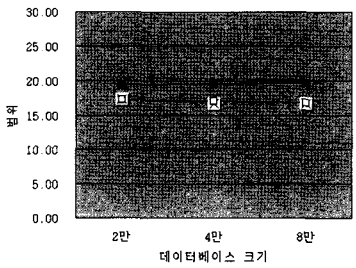
그림 7은 다분할을 수행하는 여러 기법에 대해 연관규칙 개수를 측정한 실험 결과이다. 실험에서 분할 구간의 개수(n)는 고정 값을 사용하지 않고, 극 값의 주기



(a) 정량적 속성 비율에 따른 범위

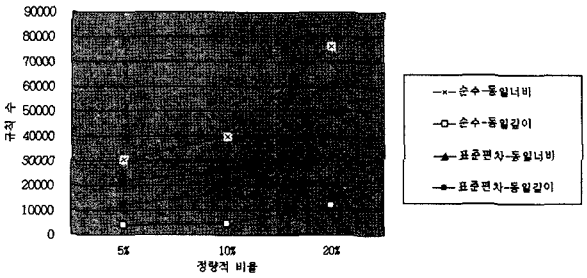


(b) f에 따른 범위

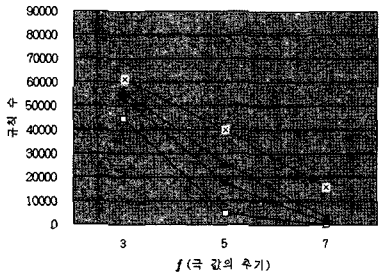


(c) 트랜잭션 수에 따른 범위

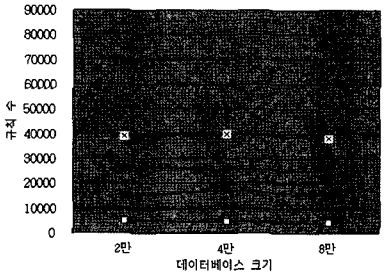
그림 6 양분할에 의해 마이닝된 연관규칙들의 정량적 속성 범위



(a) 정량적 속성 비율에 따른 규칙 수



(b) f에 따른 규칙 수



(c) 트랜잭션 수에 따른 규칙 수

그림 7 다분할에 의해 마이닝된 연관 규칙의 개수

(f)에 변동되어 변하도록 $n=f+1$의 관계로 설정하였다. 예를 들어, f=3인 경우는 분할 구간을 4로, f=5인 경우는 분할 구간을 6으로 설정하였다[2]. 실험 방법으로는 동일 너비와 동일 깊이 각각에 대해서 표준편차 최소화 기법을 사용하는 경우와 그렇지 않은 경우를 구분하여 실험하였다. (각 방법을 표준편차-동일너비, 순수-동일너비, 표준편차-동일깊이, 순수-동일깊이라 부르기로 한다.) 그림 7의 결과를 보면, 순수-동일너비(표준편차 최소화를 사용하지 않는 동일 너비 기법)에서 가장 많은 수의 연관규칙이 생성됨을 알 수 있다. 마이닝된 연관규칙들을 분석한 결과, 이는 특정 구간에 트랜잭션이 편중되는 현상에 의한 것이다. 즉, 동일 너비에 따라 특정 구간에 트랜잭션이 매우 높게 편중되고(반면에, 나머지 구간에는 트랜잭션이 거의 배정되지 않고), 이들 편중된 구간은 거의 모든 다른 항목들을 대상으로 규칙들을 생성하기 때문이다. 그런데, 이와 같이 특정 분할 구간에 의해 생성된 연관규칙들은 대부분이 의미 없는 규칙이라 할 수 있다. (예를 들어, 슈퍼마켓에서 90%의 사람이 쇼핑 봉투를 구매한다면, 쇼핑 봉투를 포함하는 연관규칙은 실제로 의미가 없다고 할 수 있는 것과 같은 이치이다.) 따라서, 본 실험 결과에서는 이상 현상이 발생하는 순수-동일너비는 다른 기법과의 비교에서 제외한다.

그림 7에서 순수-동일 너비를 제외한 세 가지 기법을 비교하면 표준편차-동일너비, 표준편차-동일깊이, 순수-동일깊이 순으로 많은 연관규칙을 생성한다. 특히 동일 깊이에 대해서 표준편차 최소화를 사용하는 경우(표준편차-동일깊이)가 그렇지 않은 경우(순수-동일깊이)에 비하여 많은 연관규칙을 생성함을 알 수 있다. 이는 양분할의 실험 결과와 마찬가지로, 표준편차 최소화 기법이 속성 값의 밀집 지역이나 희소 지역을 다른 기법보다 명확히 구분하기 때문이다. 그리고 이러한 경향은 그림 7(a)~7(c)의 세 가지 실험 결과에서 모두 동일하게 나타났다. 반면에, 표준편차-동일너비는 순수-동일너비에 비해서 규칙 개수가 오히려 줄어들었는데, 이는 표준편차 최소화 전략이 동일너비에서 나타난 이상 현상을 완화시키기 때문으로 해석된다. 그림 7의 결과를 종합하면, 표준편차-동일너비와 표준편차-동일깊이는 순수-동일깊이에 비해서 각각 평균 394%, 평균 209% 규칙개수를 늘린 것으로 나타났다.

그림 7에서 가로축 값 변화에 따른 연관규칙의 변화 경향을 해석하면 다음과 같다. 우선, 그림 7(a)에서 정량적 속성의 비율이 증가할수록 마이닝된 규칙의 수가 증가하는데, 이는 그림 5(a)의 양분할과 마찬가지로 정

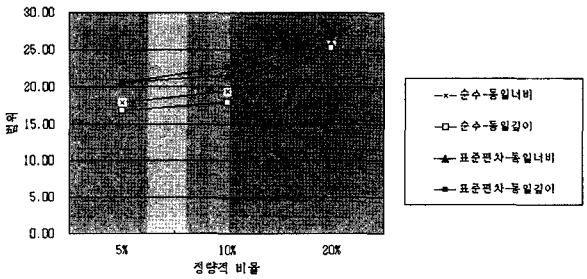
2) 본 논문에서 $n=f+1$의 관계를 사용한 이유는, 그림 4의 중요도 그래프를 보면 주어진 f에 대해 $n+1$개의 영역이 생성되기 때문이다. 예를 들어, f=3인 경우 중요도 그래프가 네 개의 영역으로 구분되고, f=5인 경우 여섯 개의 영역으로 구분됨을 볼 수 있다. 이 방법 이외에, 분할 구간의 수를 결정하는 정형적 방법으로 참고문헌 [11]의 결과를 사용할 수도 있다.

량적 속성 비율이 높을수록 생성되는 구간의 개수가 많아지기 때문이다. 반면에, 그림 7(b)는 그림 5(b)와는 반대로 극 값의 주기 f가 증가할수록 연관규칙 개수가 감소함을 볼 수 있다. 이는 분할 구간 개수가 둘로 고정된 양분할과는 달리 다분할에서는 분할 구간 개수를 $f+1$로 동적으로 조정하였기 때문이다. 즉, 분할 구간의 수가 많아짐에 따라 구간에 속하는 트랜잭션 빈도수가 줄어들고, 이에 따라 연관규칙의 수가 줄어드는 것이다. 마지막으로, 그림 7(c)을 보면, 트랜잭션 개수 변화에 따른 연관규칙 개수 변화는 크지 않음을 알 수 있다. 이는 다분할에서는 분할 구간의 수가 많아, 트랜잭션 개수를 증가하여도 각 구간에 속하는 트랜잭션 개수의 비율이 양분할에 비해 적기 때문으로 해석된다.

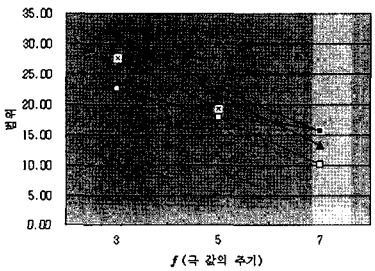
다음으로, 그림 8은 다분할 환경에서 순수-동일너비, 표준편차-동일너비, 순수-동일깊이, 표준편차-동일깊이의 네 가지 기법에 대한 정량적 속성 범위를 측정한 실험 결과이다. 그림을 보면, 대다수의 경우에 있어서 표준편차 최소화를 사용하는 두 가지 기법이 이를 사용하지 않는 나머지 두 가지 기법에 비해 보다 넓은 정량적 범위를 형성함을 알 수 있다. 특히, 동일 너비에 있어서 많은 규칙을 생성했던 순수-동일너비보다 표준편차-동일너비가 (평균적인) 정량적 범위가 크게 나타남을 볼 수 있다. 앞서 설명한 바와 같이 이는 순수-동일너비가 이상 현상에 의해 의미 없는 많은 규칙을 생성한다는 사실을 뒷받침하는 결과이기도 하다. 그림 8에서 가로축 값의 변화에 따른 범위의 변화는 그림 7의 연관규칙 개수의 변화와 유사한 경향을 보이는 것으로 나타났는데, 이는 양분할에 대한 실험 결과인 그림 6과 그림 5의 관계와 유사한하므로, 자세한 분석은 생략한다. 그림 8의 실험 결과를 요약하면, 표준편차 최소화 기법은 다른 기법에 비해 평균적으로 3%~18% 정량적 범위를 증가시킨 것을 나타났다.

실험 3) 전처리 및 후처리 과정이 전체 성능에 미치는 영향 분석

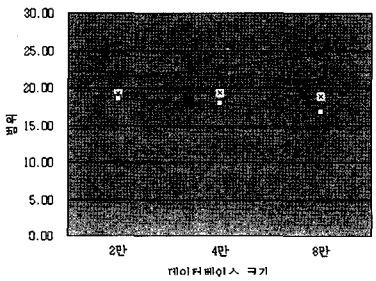
표 8은 전체 마이닝 시간과 이중에서 전처리 및 후처리 과정이 차지하는 시간을 나타낸 것이다. 표의 실험은 50개의 속성 중에서 10%에 정량적 속성을 부여한 경우이다. 극 값의 주기 f를 5로 하여 한 여섯 구간으로 나누는 다분할 기법을 실험하였으며, 데이타베이스 크기는 4만개로 하였다. 표에서 보듯이, 전처리 및 후처리 과정은 전체 마이닝 시간의 3.6%~5.4%로 그리 크지 않은 것으로 나타났다. 이는 상용 데이타 마이닝 도구를 사용하는데 있어서 전처리 및 후처리 과정의 성능 오버헤드가 그다지 크지 않음을 의미한다. 또한, 표준편차를 사용한 방법이 그렇지 않은 방법에 비해서 전처리 및 후처리 과정의 오버헤드가 크게 심하지 않음도 확인할 수 있다. 따라서 제안한 방법은 성능적인 측면에서 충분히 활용 가능한 방법이라 사료된다.



(a) 정량적 속성 비율에 따른 범위



(b) f에 따른 범위



(c) 트랜잭션 수에 따른 범위

그림 8 다분할에 의해 마이닝된 연관규칙들의 정량적 속성 범위

표 8 전체 마이닝 시간과 전처리 및 후처리 시간 비교

구분	전체 시간		마이닝 시간		전처리/후처리 시간	
	초	%	초	%	초	%
순수-동일너비	102.1	100.0	98.4	96.4	3.7	3.6
순수-동일깊이	92.4	100.0	87.4	94.6	5.0	5.4
표준편차-동일너비	100.3	100.0	94.9	94.6	5.4	5.4
표준편차-동일깊이	96.6	100.0	91.6	94.8	5.0	5.2

7. 결 론

본 논문에서는 상용 데이타 마이닝 도구로 정량적 연관규칙을 마이닝하기 위한 체계적인 접근법을 제안하였다. 정량적 연관규칙을 마이닝하는 기존 알고리즘은 상용 데이타 마이닝 도구에는 직접 적용하기 어려운 문제점이 있다. 이러한 문제점을 해결하기 위하여, 본 논문에서는 상용 데이타 마이닝 도구를 중심으로 전처리 과정과 후처리 과정을 두어 상용 도구의 연관규칙 알고리즘을 수정하지 않고도 정량적 연관규칙을 마이닝하는 방법을 제안하였다.

본 논문의 공헌을 요약하면 다음과 같다. 첫째, 상용 데이타 마이닝 도구를 사용하여 정량적 연관규칙을 찾아내기 위한 전체적인 프레임워크를 제안하였다. 제안한 프레임워크는 상용 데이타 마이닝 도구를 중심으로 정량적 속성을 이진 속성으로 변환하는 전처리 과정과 이진 연관규칙을 정량적 연관규칙으로 변환하는 후처리 과정으로 구성된다. 둘째, 구간분할의 개념을 정형적으로 정의하고, 기존의 양분할 및 다분할을 구간 분할 개념을 사용하여 재정의하였다. 본 논문에서는 양분할에 대해서 평균 및 중앙치 기법을, 다분할에 대해서 동일 너비 및 동일 깊이 기법을 구간 분할 개념을 사용하여 각각 재정의하였다. 셋째, 기존 구간분할 방법들이 트랜잭션의 분포를 고려하지 않는 단점을 해결하기 위하여 표준편차 최소화 기법을 제안하였다. 표준편차 최소화 기법은 이웃한 속성 값의 표준편차가 작다면 해당 값들은 동일한 구간에 포함시키고, 표준편차가 크다면 이들을 다른 구간으로 분할하는 방법이다. 넷째, 상용 데이타 마이닝 도구를 사용하여 마이닝된 이진 연관규칙들을 원래의 정량적 연관규칙으로 변환하는 후처리 과정을 제안하였다. 다섯째, 양분할과 다분할에 대한 다양한 실험을 통해, 제안한 프레임워크가 바르게 동작함을 보이고, 표준편차 최소화를 사용하는 방법이 그렇지 않은 경우에 비해 우수한 결과를 나타냄을 입증하였다.

이 같은 결과를 볼 때, 제안한 방법론은 상용 데이타 마이닝 도구를 활용하는 일반 사용자가 정량적 연관규칙을 쉽게 마이닝 할 수 있는 일반적이고 합리적인 프레임워크라 생각한다. 향후 연구로는 1) 마이닝된 정량적 연관규칙의 중요성을 판별하는 작업, 2) 표준편차 이외에 상관관계(correlation) 등의 다른 통계적 특성을 구간 분할에 적용하는 작업, 3) k-평균 알고리즘(k-means algorithm)과 같은 클러스터링 기법을 사용하여 구간을 분할하는 방법과의 비교 작업 등을 수행할 예정이다.

참 고 문 헌

[1] Argrawal, R., Imielinski, T. and Swami, A., "Mining Association Rules in Large Databases," In *Proc. Int'l Conf. on Management of Data, ACM SIGMOD*, Washington D.C, pp. 207-216, May. 1993.

[2] Agrawal, R. and Srikant, R., "Fast Algorithms for Mining Association Rules in Large Databases," In *Proc. the 20th Int'l Conf. on Very Large Data Bases*, Santiago, Chile, pp. 487-499, Sept. 1994.

[3] Park, J.-S., Chen, M.-S. and Philip S. Y., "An Effective Hash-based Algorithm for Mining Association Rules," In *Proc. Int'l Conf. on Management of Data, ACM SIGMOD*, San Jose, California, pp. 175-186, May, 1995.

[4] Savasere, A., Omiecinski, E. and Navathe, S., "An Efficient Algorithm for Mining Association Rules in Large Databases," In *Proc. the 21st Int'l Conf. on Very Large Databases*, Zurich, Switzerland, pp. 432-443, Sept. 1995.

[5] Brin, S., Motwani, R., Ullman, J. D. and Tsur, S., "Dynamic Itemset Counting and Implication Rules for Market Basket Data," In *Proc. Int'l Conf. on Management of Data, ACM SIGMOD*, Tucson, Arizona, pp. 255-264, 1997.

[6] Srikant, R. and Agrawal, R., "Mining Genralized Association Rules," In *Proc. the 21st Int'l Conf. on Very Large Databases*, pp. 407-419, Sept, 1995.

[7] Srikant, R., Vu, Q. and Agrawal, R., "Mining Association Rules with Items Constraints," In *Proc. the 3rd Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 67-73, Aug. 1997.

[8] Toivonen, H., "Sampling Large Databases for Association Rules," In *Proc. the 22th Int'l Conf. on Very Large Data Bases*, Mumbai(Bombay), India, pp. 134-145, Sept. 1996.

[9] Park, J.-S., Yu, P.-S. and Chen, M.-S. "Mining Association Rules with Adjustable Accuracy," In *Proc. the ACM Sixth Int'l Conf. on Information and Knowledge Management*, Las Vagas, Nevada,

pp. 151-160, Nov. 1997.
[10] Savasere, A., Omiencinski, E. and Navathe, S., "Mining for Strong Negative Associations in a Large Database of Customer Transactions," In *Proc. the 14th Int'l Conf. on Data Engineering*, Olrando, Florida, pp. 494-502, Feb, 1998.
[11] Srikant, R. and Agrawal, R., "Mining Quantitative Association Rules in Large Relational Tables," In *Proc. Int'l Conf. on Management of Data, ACM SIGMOD*, Montreal Canada, pp. 1-12, June. 1996.
[12] Wang L., David W. C. and Yiu, S. M., "An Efficient Algorithm for Finding Dense Regions for Mining Quantitative Association Rules," *Computers & Mathematics with Applications*, Vol.50, No.3-4, pp. 471-490, Aug. 2005.
[13] Hu, C., et al., "Mining Quantitative Associations in Large Database," In *Proc. the 7th Asia-Pacific Conf. on Web Technologies Research and Development, APWeb2005, Shanghai China*, pp. 405-416, Mar. 2005.
[14] 이혜정, "병렬 처리를 이용한 효과적인 수량 연관규칙에 관한 연구", 순천향대학교 대학원, 전산학과, 박사학위 논문, 2007. 02.
[15] Imberman, S. and Domanski, B., "Finding Association Rules From Quantitative Data Using Data Booleanization," In *Proc. the 7th Americas Conf. on Information Systems*, City University of New York, 2001.
[16] IBM. http://www-07.ibm.com/software/kr/data/db2/product/intelligent_miner_data.html,
[17] SAS Enterprise Miner. http://www.sas.com/technologies/analytics/datamining/miner/
[18] Silicon Graphics MineSet. http://www.sgi.com/
[19] SPSS Clemetine. http://www.spss.com/clementine/
[20] Mendenhall, W. and Beaver, R. J., Introduction to Probability and Statistics, Eighth Edition, Thomson Information, pp. 23-56, 2005.
[21] Gibbons, P., Matias, Y. and Poosala, V., "Fast Incremental Maintenance of Approximate Histograms," In *Proc. the 23th Int'l Conf. on Very Large Data Bases*, Athens, Greece, pp. 466-475, Aug. 1997.
[22] Grahne, G. and Zhu, J., "Fast Algorithms for Frequent Itemset Mining Using FP-Trees," *IEEE Trans. Knowl. on Data Engineering*, Vol.17, No.3, pp. 1347-1362, Oct. 2005.
[23] 강현철, 한상태, 최종후, 김차용, 김은성, 김미경, "SAS Enterprise Miner를 이용한 데이타 마이닝(방법론 및 활용)", 자유아카데미, 1999.
[24] 최종후, 한상태, 강현철, 김차용, 김은성, 김미경, "SAS Enterprise Miner를 이용한 데이타 마이닝(기능과 사용법)", 자유아카데미, 1999.

강 공 미

1986년 2월 한양대학교 가정학과 학사 1999년 2월 한국방송통신대학교 전자계산학과 학사. 2002년 2월 강원대학교 교육대학원 컴퓨터교육과 석사. 2006년 2월~현재 강원대학교 IT특성화대학 컴퓨터학부 박사과정. 2006년 2월~현재 남원주 중학교 교사. 관심분야는 Data Mining, Quantitative Association Rules, Data Warehouse, OLAP, Web Database Technology

문 양 세

1991년 2월 한국과학기술원 과학기술대학 전산학과 학사. 1993년 2월 한국과학기술원 전산학과 석사. 2001년 8월 한국과학기술원 전자전산학과 전산학전공 박사. 1993년 2월~1997년 2월 현대전자산업(주) 통신사업본부 주임연구원. 2001년 9월~2002년 2월 (주)현대시스콤 호처리개발실 선임연구원 2002년 2월~2005년 2월 (주)인프라밸리 기술연구소 기술위원(이사). 2005년 3월~현재 한국과학기술원 첨단정보기술연구센터 연구원. 2005년 3월~현재 강원대학교 컴퓨터과학과 조교수. 관심분야는 Data Mining, Knowledge Discovery, Stream Data, Storage System, Database Applications, Mobile/Wireless Communication Services & Systems

최 훈 영

2003년 2월 강원대학교 컴퓨터공학과 학사. 2006년 2월 강원대학교 교육대학원 컴퓨터교육전공 석사. 2005년 3월~2006년 2월 한국과학기술원 첨단정보기술연구센터 연구보조원. 2007년 3월~현재 강원대학교 컴퓨터과학전공 박사과정. 관심분야는 Computer Education, Data Mining, Knowledge Discovery, Embedded Systems & Algorithms, GPU-based Algorithms

김 진 호

1982년 2월 경북대학교 전자공학과 학사 1985년 2월 한국과학기술원 전산학과 석사. 1990년 2월 한국과학기술원 전산학과 박사. 1995년 8월~1996년 7월 미국 미시간 대학교 객원 교수. 2003년 2월~2004년 2월 미국 Drexel University 객원 교수. 1999년 3월~현재 한국과학기술원 첨단정보기술연구센터 연구원. 1990년 8월~현재 강원대학교 컴퓨터과학과 교수. 2006년 9월~현재 강원대학교 중앙교육연구전산원 원장. 관심분야는 Data warehouse, OLAP, Data Mining, Real-time/Embedded Database, Main-memory database, Data Modeling, Web Database Technology