□ 기술해설 □

데이타 마이닝의 유용성

숙명여자대학교 윤종필*·김희숙**·최옥주***

1. 서 론

데이타 마이닝은 점점 중요한 연구 분야로 관심의 대상이 되고 있다[10]. 여러 분야에서 많은 프로토타입과 상용 툴들이 개발되어 이용되고 있다. 그러나, 대부분의 데이타 마이닝 툴은 너무 많은 필요 이상의 정보를 생성하거나, 또는 사용자의 의도와 관계없는 정보이어서 그 유용성이 떨어지는 양상을 보이고 있다. 최근에 데이타 마이닝의 유용성에 대한 문제가 제기되고 여러 논문에서도 큰 이슈화되고 있는 실정이다. 본 논문에서는 이러한 배경으로 데이타 마이닝의 유용성에 대해 논의하고자 한다. 데이타 마이닝의 유용성은 크게 두가지 연구 방향을 생각할 수 있다. 첫째, 마이닝 된 데이타 혹은 지식의 유용성의 정도를 검증하기 위한 연구 방향이고, 둘째로, 유용한 데이타 혹은 지식을 생성하는 마이닝 기법을 개선하기 위한 연구 방향이다. 전자에서는 여러 가지 검증의 지표가 제안되어 이용되고 있다. 마이닝된 정보의 유용성을 정량적으로 지표화 하려는 노력을 의미한다. 이에 반하여, 후자의 경우는 데이타 마이닝 작업 초기부터 사용자가 원하는, 또는 사용자에게 유용한 정보를 추출하기 위한 마이닝 툴(또는 방법)을 개발하는 노력을 의미한다. 유용한 정보를 추출하기 위하여 사용자와 상호 작용을 하거나 때로는 사용자 또는 데이타의 정보(혹은 스키마 정보)로 부터 마이닝 작업을 시작하기도 한다[24]. 본 논문에서는 주로 전자의 데이타 마이닝 유용성에 대하여 논의하고자 한다.

기존의 실험적인 표본 조사를 통한 사례의 적용에서, 실세계의 데이타를 통한 데이타 마이닝 기술로 보다 유용한 정보를 추출하고 또한 기대하지 않았던 지식까지도 추출할 수 있는 마이닝 기법을 개발하는 것이 목표이기도 하다. 이와 같은 목표 달성을 위해 시도하고 있는 핵심 이슈들을 여러 각도에서 정리하려고 한다. 본 논문의 구성은, 2장에서 데이타 마이닝에서 필요로 하는 입력 정보들을 정리하고, 3장에서 데이타 마이닝 방법을 분류하고, 4장에서 데이타 마이닝의 유용성의 검증 지표를 소개한다. 마지막 5장의 결론으로 본 논문을 마치려 한다.

2. 데이타 마이닝 입력정보

실세계의 데이타베이스는 여러 가지 상황에 따른 입력 데이타가 있을 수 있다.

분석의 대상이 되는 데이타의 물리적 형태는 텍스트 또는 데이타베이스의 테이블로 이루어져 있는데, 주어진 입력에 따라 여러 지식의 형태를 도출시킬 수 있고, 입력의 구체성에 따라 데이타 마이닝의 대상탐색공간이 줄어들 수도 있다. 데이타의 논리적 형태로는 데이타베이스의 스키마 정보 혹은 사용자의 질의문을 들 수 있다. 이와 같은 정보는 모두 대상 데이타와 직접 종속된 정보들이다. 데이타 마이닝의 입력 정보로는 대상 데이타에 독립적인 정보도 있을 수 있는데, 이는 여러 마이닝 연산들의 결과 임계치를 결정하기도 한다.

*종신회원
**학생회원
***정 회 원

한편, 데이타 마이닝의 입력정보는 입력 데이타가 불완전(incomplete)하여 주어지지 않거나, 잘못 주어질 수도 있다. 이에 따른 여러 가지 입력 사항에 대해 알아본다.

2.1 일반적으로 주어지는 입력정보

위에서도 언급되었듯이, 일반적으로 주어지는 데이타 마이닝 입력정보는 다음과 같이 크게 두 가지로 나뉘어 진다.

- 논리적 데이타 입력: 데이타 스키마 정보, 사용자 질의문, 결정(determination)규칙, 도메인 지식, 사용자의 원하는 목표(goal) 등
- 물리적 데이타 입력: Belief 지식, a priori 임계치 등

2.1.1 Belief

실세계에서 때때로 불확실성에 근거하여 결정을 내려야 할 때 적절한 단정을 내리기 위해 belief를 사용한다. 어떤 사건에 대한 사람이 가지고 있는 경험적 요소를 포함한 믿음의 정도를 나타낸다[16, 21]. 예를 들어, 증권 상담자가 증권에 대한 예측을 할 때 주가와 회사 정보 또는 환율 등의 통계적인 상관성(correlation), 정성적인 전문지식 기반 통계 방법에 근거하여 증권 예측을 결정한다.

Belief의 정도를 Bayesian 네트워크를 이용하여 표현하면, 다음과 같은 확률로 주어진다[12].

$P(e \mid \zeta)$: ζ가 주어질때, e의 확률(ζ: 확률을 제공하는 사람의 지식 상태, e: 사건)

2.1.2 A priori 지식

프레디키트(predicate) 규칙 "if x then y"에서 x는 y를 수반하는 경향이 있고 x는 y로의 가능성이 증가함을 나타낸다. 이와 같은 규칙에 대한 관련 임계치를 설정함으로써 데이타 독립적 입력정보를 제시할 수 있다. 예를 들면, 자주 발생되는 항목의 집합을 찾기 위해 최소 지지도와 최소 신뢰도를 만족하는 large itemset을 찾는다[1].

2.1.3 사용자 질의문

사용자는 데이타베이스 시스템에서 제공되는 질의문을 이용하여 사용자의 관심사를 표현할 수 있다. 대부분의 사용자의 질의는 그 질의가 사용자의 의도나 관심사를 내포한다고 믿는다. 즉 사용자의 전제사항(hypotheis)이나 관심사를 반영한다[24].

사용자의 질의문에 따라 대상 데이타베이스(데이타 집합)를 크게 세 부분집합으로 나눌 수 있다. 예를 들면, "SELECT attributes FROM table names WHERE a condition"의 결과로 query Q는 전체 데이타 집합, T_1, ···, T_l이 다음과 같이 나뉘어 진다[24].

i) Positively-related query view(PRQV) : 테이블에서 질의를 만족하는 결과

$$PRQV(Q)=Q(D)=\pi_{A1, \cdots, Ak}(\sigma_{condition}(T_1 \infty..\infty T_l))$$

ii) Negatively-related query view (NRQV) : 테이블에서 질의를 만족하지 않는 결과

$$NRQV(Q)=\pi_{A1, \cdots, Ak}(T_1 \infty..\infty T_l)-PRQV(Q)=\pi_{A1, \cdots, Ak}[(\sigma \neg_{condition}(T_1 \infty..\infty T_l))]$$

iii) Unrelated-related query view(URQV) : 데이타베이스에서 질의에 포함되지 않은 결과

$$URQV(Q) = D - [PRQV(Q) \cup NRQV(Q)]=D-[\pi_{A1, \cdots, Ak}(T_1 \infty..\infty T_l)]$$

2.1.4 결정 규칙(Determination rule)

구체적인 데이타로부터 일반적인 규칙을 생성하는 induction에서 두개의 규칙사이에 함수적 관계가 있다고 가정하고 규칙을 세우는 것이다. 예를 들면, $p(x, y) \Rightarrow q(x, z)$는 모든 문자, 즉, p, q, x, y, z가 변수인 경우, 결정규칙에 속하는데, 이와 같은 입력정보로부터 customer-service(x, y)⇒complain(x, z)라는 새로운 규칙이 도출될 수 있다[17].

2.1.5 데이타 스키마/도메인 지식

주로 대용량 데이타베이스, 특히, 분산 데이타베이스로부터 데이타 마이닝 또는 지식 발견을 위하여 입력되는 정보로서 이질 데이타베이스에 대한 스키마 정보와 도메인 정보를 뜻한

다[7]. 분산 데이타베이스로부터의 데이타 마이닝은 크게 두 가지 방향 즉, 先마이닝 後통합 방향과 先통합 後마이닝 방향이 있다. (1) 전자의 방향의 경우, 각 사이트의 데이타베이스로부터 도출된 정보를 전체 분산 데이타베이스의 대표성 있는 정보로 재구성하기 위하여 필요로 하는 정보이고, (2) 후자의 경우, 통합된 데이타베이스로부터 마이닝을 시도할 때, 이용할 수 있는 각 사이트에 관한 정보를 뜻한다.

2.2 불완전 입력(Incomplete data인 경우)

데이타베이스의 내용에는 값이 불완전하게 없는 것도 있을 수 있다. 결측치(missing value)는 다양한 방법으로 다루어질 수 있다. 단순하게 무시하거나, 알려진 값으로부터 유추하거나 평균값으로 대체시키기도 한다[12]. 실시간 데이타베이스로부터 마이닝을 시도할 경우, 나타날 수 있는 문제이기도 하다.

2.3 오류 입력(Outlier 혹은 Noise)

데이타베이스에 입력되는 데이타는 오류(noise)가 있을 수 있기 때문에 보통 DBMS의 일관성 기능을 이용하여 미리 제거하는 것이 일반화되어 있다. 그러나, DBMS를 이용하지 않는 기타 데이타베이스의 경우는 자료가 정확하다고 가정할 수가 없기 때문에 정제할 필요가 있다.

3. 데이타 마이닝 방법

데이타 마이닝, 즉 대용량 데이타베이스 또는 정보 저장소에서 추출된 지식은 데이타의 가치와 지식의 유용성을 캐내는 것으로 여러 분야에 광범위하게 활용된다.

데이타 마이닝 방법은 방대한 자료로부터 숨겨진 어떠한 정보 패턴을 추출하는데 관심이 있는 단계로 마이닝의 목적에 따라 여러 가지로 분류되어 질 수 있다. 추출된 패턴들은 어떤 사실(fact)들의 집합으로, 의사결정이나 업무의 예측이나 분류에 사용되거나 관찰된 현상의 설명등으로 해석되어진다.

3.1 분류(Classification) 방법

분류의 목적은 입력 데이타를 분석하여 각 클래스에 대해 정확한 표현(description)이나 모델을 개발하는 것이다. 여기서 입력 데이타는 속성이나 특성에 대한 레코드(예를 들어, training set)들로 구성된다. 클래스 분류는 테스트 데이타를 분류하기 위해 학습 데이타로부터 도출될 수 있다. 일반적으로 수백만 표본을 가진 큰 데이타 집합을 분류하기 위해서 의사결정 트리 분류자(Decision Tree Classifier)를 사용한다[19]. 의사결정 트리 분류자는 다른 분류 방법에 비해 상대적으로 빠르며 데이타베이스를 액세스할 수 있는 SQL 질의로 전환될 수 있다. 그림 1은 6개의 표본에 대한 의사결정 트리 분류의 예이다.

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

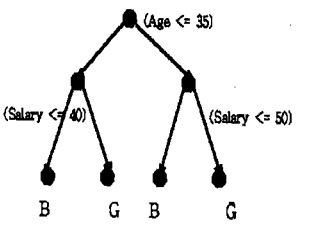

그림 1 의사결정 트리의 예

예를 들면, B는 Age〈=35인 경우는 Salary〈=40을 만족하고, Age〉35인 경우는 Salary〈=50을 만족하는 클래스이다.

3.2 클러스터링(Clustering) 방법

유사한 특성을 갖는 클래스를 함께 그룹화하고 분할하는 방법으로, 어떤 그룹이 사전에 정의되어 있지 않다는 점에서 분류 방법과 차이점이 있다. 이 방법은 유사한 속성을 가진 객체들끼리 군집화한다. 일반적으로, k개의 군집에 대한 클러스터링 방법은 다음과 같이 나뉘어질 수 있다[5].

• 메트릭 거리 기반 : 데이타간의 거리를 계산함으로 데이타들을 k개의 군집으로 나눌 수 있다. 즉, 하나의 데이타는 한 군집에 속함으로 그 군집내의 다른 데이타와의 거리가 다른 군집내의 데이타와의 거리보다 작다는 관계가 성립된다. 예를 들면, "k-평균값 기반"은 비슷한 성향을 가진 레코드들은 데이타의 공간상에서 서로 근접하게 분포할 것이므로 각 공간에서의

거리를 근거로 k개의 그룹으로 클러스터링하고 분할된 각 그룹에 대해 그룹내의 유사성을 중심으로 특징을 찾아내는 것이다.

• 모델 기반 : 각 클러스터를 위한 모델이 가설되고 가장 적합한 모델에 의거하여 클러스터를 찾는 것이다. 만일 D가 데이타이고 M_l 이 클러스터 l(l∈{1, …, k})을 위한 가설된 모델일 때 다음과 같은 조건부 확률을 가지게 된다.

$$\text{Prob}(M_l \mid D) = \text{Prob}(M_l \mid D)\frac{\textit{Prob}(M_l)}{\textit{Prob}(D)}$$

이 경우, Prob(M)은 모델이 한 클러스터에 속할 확률이고, Prob(D)는 데이타가 한 클러스터에 속할 확률을 뜻한다.

위에서, 두 클러스터 Xi, Xj 사이의 거리 척도를 소개하면 다음과 같다[25].

• Euclidean 거리 (D0)

$$D0 = [(\vec{X}\boldsymbol{i} - \vec{X}\boldsymbol{j})]^{\frac{1}{2}}$$

• 맨해튼 거리 (D1)

$$D1 = |(\vec{X}\boldsymbol{i} - \vec{X}\boldsymbol{j})| = \sum_{k=1}^{d} |(\vec{X}\boldsymbol{i}^{(k)} - \vec{X}\boldsymbol{j}^{(k)})|$$

• 클래스간 평균거리 (D2)

$$D2 = \left[\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}(\vec{X}\boldsymbol{i} - \vec{X}\boldsymbol{j})^2}{N_1 N_2}\right]^{\frac{1}{2}}$$

• 클래스내의 평균거리 (D3)

$$D3 = \left[\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{N_1+N_2}(\vec{X}\boldsymbol{i} - \vec{X}\boldsymbol{j})^2}{(N_1+N_2)(N_1+N_2-1)}\right]^{\frac{1}{2}}$$

그밖에, 두개의 확률분포 p와 q사이의 거리, 즉 relative entropy라고 부르는, 거리는 다음과 같이 정의된다[14]. 이 경우 p는 데이타의 분포를 나타내고 q는 참조 관련분포(예, 균등분포)를 나타낸다.

• Entropy 거리.

$$D(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{i=1}^{n} \boldsymbol{p}(x_i)\ \boldsymbol{log}_2\ \frac{\boldsymbol{p}(x_i)}{\boldsymbol{q}(x_i)}$$

3.3 요약(Summarization) 방법

데이타의 부분집합을 표현하기 위해 함축된 패턴을 추출하는 것이 요약의 목적이다. 요약방법에는 데이타의 조각을 수직으로 표현하느냐 수평으로 표현하느냐에 따라 두가지 방법이 있다. 데이타의 조각을 수평으로 표현하는 방법은 데이타의 부분집합을 적당한 통계나 논리적 조건에 의해 생산하는 것이다. 수직으로 표현하는 방법은 필드사이의 관계를 예측하는 것과 같다. 즉, 필드사이의 관계를 찾아내는 것이다[5]. 요약방법에 대해서는 현재 많은 연구가 이루어지고 있으며, 본 절에서는 연관규칙과 일반화 등에 대하여 알아보기로 한다.

3.3.1 연관 규칙(Association rule)

트랜잭션의 모임으로부터 한 항목집합의 존재와 다른 항목집합과의 연관 관계를 요약할 수 있다. 연관 관계는 항목들 사이에 존재하는 유사성 또는 패턴을 의미한다. 연관 규칙의 기본 개념을 설명하면, $I=\{i_1, i_2, \cdots, i_k\}$라는 항목(item)집합 즉 트랜잭션의 집합을 고려한다. 이 경우 $i_1, i_2, \cdots, i_k \subset I$이다. ϕ이 아닌 항목 집합 X, Y에 대해 X⊂I, Y⊂I 에 대한 연관규칙 X→Y는 $X \cap Y = \phi$의 특성을 갖는다. X는 규칙의 가정, Y는 규칙의 결과라고 한다. 항목집합 I의 부분집합 X에 대해, X⊆T이면 T는 X를 만족한다고 정의한다. 최소지지도(minimum support threshold)를 만족하는 X⊆I를 Large 항목집합이라 한다[1, 22].

3.3.2 일반화(Generalization)

데이타 집합을 요약하므로 한 레벨 상위 개념을 표현할 수 있는 방법이다. 예를 들어 판매 데이타베이스에서 상품번호, 상품명, 제조일, 가격 등의 애트리뷰트는 상품이라는 상위 애트리뷰트로 요약될 수 있다. 하위 레벨에서 관련 있는 상위 레벨로 관련 있는 데이타의 큰 집합으로 추상화시키는 과정이다[7, 9].

• 데이타 큐브(Data cube) 방법 : 방대한 데이타를 다차원으로 분류하여 표현하는 데이타 큐브로부터 통합된, 혹은 일반화된 데이타를 추론하는 방법이다[3, 25].

• 속성기반 일반화 방법 : 데이타베이스를 구성하는 속성들의 도메인에 대한 일반화된 관계(예, 개념계층)를 이용하여 데이타의 추상화단계를 높이는 방법이다[7].

3.4 변화(Change) 방법

순차정보, 시계열 정보, 어떤 다른 순서(예를 들어 유전자의 단백질 순서 등)를 설명하는 방

법이다, 이 방법에서는 관찰순서가 중요하다. 본 절에서는 시계열과 예측모델링에 대하여 알아본다[5].

3.4.1 시계열(Time-series)

시계열 데이타는 복잡한 데이타 객체에서의 중요한 클래스가 되며 금융이나 과학 응용프로그램에서 많이 볼 수 있다. 예를 들어, 주식 가격 지수, 상품 판매량, 원격통신 데이타, 1차원 의학 신호, 오디오 데이타, 환경측정 신호 등이 있다.

대부분의 경우 시계열 데이타베이스에서 주어진 질의순차와 유사한 시계열을 찾는 것이 필요하다. 순차식 X와 순차식 X′이 있고 이때, X′은 outlier값이나 다른 비례식과 기준선에 의해 수정되었다면, 순차식 X, X′은 유사성이 있다고 간주 할 수도 있다. 이때, 시계열 데이타의 유사성은 통계식으로 표현할 수 있다.

시계열은 정수의 유한한 순차식으로 표현된다: $X(x_1, x_2, \cdots, x_n)$. 두개의 순차식 $X=(x_1, x_2, \cdots, x_n)$, $Y=(y_1, y_2, \cdots, y_n)$가 F-similar하다는 것은 순차식 X의 부분 순차식 X′이 함수 f를 사용하여 순차식 Y의 부분 순차식 Y′과 거의 매핑되는 함수($f \in F$)가 존재한다는 것이다[8]. 예를 들면, 유사한 순서를 보이는 사건들 사이의 시간적인 관계를 다음과 같이 표현할 수 있다: $Y(t)=f(X_1(t), X_2(t), \cdots, X_k(t))$. 여기서 정수와 정수를 매핑하는 변형함수 $f \in F$는 모든 일차식($x \rightarrow ax+b$)이나 비례식($x \rightarrow ax$), 이차식, 항등식 등으로 구성된다[4, 8]. 유사성에 대해서는 다음 장에서 자세히 언급한다.

3.4.2 예측(Prediction) 모델링

예측 모델링은 데이타베이스에서 다른 필드를 기반으로 특정필드를 예측하는 것이다. 일반적으로 주어진 다른 필드(입력데이타)나, 트레이닝 데이타(학습데이타, 관찰에 의한 목표변수), 문제의 사전 지식을 표현한 가정집합들에 의해 예측하고자 하는 변수의 값들을 결정한다. 입력데이타에 대해 non-linear transformation와 조합된 linear regression은 널리 사용 사용되는 기법이다. 분류별 변수(클래스)의 상태를 예측하는 기법으로는 기본적으로 밀도평가(density estimation) 문제이다[5]. 클래스 C=c, 특정 벡터 x에 대해 주어진 필드 X=x에 대한 확률을 구한다면, C와 X의 공동 밀도로부터 확률을 구할 수 있다. 그러나 공동 밀도는 흔히 알려지지 않고 평가하기가 어렵다.

3.4.3 회귀분석 (Regression Analysis)

종속변수에 대한 독립변수 모형을 분석하는 것으로 다음과 같은 회귀함수를 이용한다.

$Y_i=\alpha+\beta X_i$

변수들간의 관련성을 규명할 수 있는 수학적 모형을 측정된 변수들의 자료로부터 추정하는 통계적 방법으로 이 추정 모형을 사용하여 필요한 예측(prediction)을 하거나 통계적 추론(inference)을 하게 된다[5].

4. 유용성 측정 지표

본 절에서는 마이닝된 정보를 측정하는 정량적 지표에 대하여 알아본다. 연관규칙에서 많이 이용되고 있는 지지도와 신뢰도로부터 설득력, 경이도 또는 흥미도, 분포도, 상관도, 유사도 등을 간략하게 다룬다.

4.1 지지도(Supportiveness)

규칙의 통계적 중요성으로 전체 트랜잭션 N에 대한 x와 y를 만족하는 트랜잭션의 비율이다.

$$S(x, y)=\frac{P(x, y)}{N}$$

자주 발생되는 패턴 혹은 규칙의 빈도를 나타내므로, 그 패턴이나 규칙의 유용성이 증대되려면 지지도가 증대되어야 한다[1, 22]. 예를 들면, 수퍼마켓에서 item x, y에 대하여 추출된 규칙 $x \rightarrow y$의 지지도가 85%일때 일반적으로 모든 구매자중 85%가 item x, y 모두를 구매한다는 것이다.

4.2 신뢰도(Confidence)

위 지지도와 같은 상황에서 x를 만족하는 트랜잭션에 대한 y를 만족하는 트랜잭션의 비율

이다. 이는 규칙이 실행되어 정확도를 기할 수 있는 강도를 나타낸다[1, 22].

$$C(x, y)=\frac{P(x, y)}{P(x)}$$

높은 지지도는 정확한 예측을 제공한다. 예를 들면, 수퍼마켓에서 item x, y에 대하여 추출된 규칙 x→y의 신뢰도가 85%일때 x를 구매할 때는 85%가 y도 구매한다는 것이다.

4.3 설득치(Conviction)

신뢰도와 비슷하며, 수퍼마켓에서 두 item의 상관도가 없을 경우(또는 x와 y가 독립적일 경우)에도 측정할 수 있도록 개선된 지표이다. 예를 들면, 두 item x와 y가 무관할 경우 규칙 x→y신뢰도는 P(y)가 되는 오류를 개선하기 위함이다[6].

$$N(x, y)=\frac{P(x)P(\neg y)}{P(x, \neg y)}$$

전체 데이타베이스에서 B를 구매하지 않을 확률 대비 A를 구매하고 B를 구매하지 않을 확률이다. 이로부터, 설득력이 증가할수록 A를 구매할 때 B를 구매할 확률이 증가하게 된다. 즉, 전체 데이타베이스에서 일반적으로 B를 구매하지 않을 경우는 빈번하나, 일단 A를 구매할 경우는 B를 구매한다는 것이다.

4.4 흥미도(Interestingness) 또는 경이도(Surprisingness)

일반적으로 생성되는 규칙의 흥미도 또는 경이도는, 수퍼마켓의 경우 item x와 y가 전혀 무관할 경우에 나타난다. 즉 아래의 흥미도에서 x와 y가 독립적일 경우 1을 나타낸다.

$$I(x, y)=\frac{P(x, y)}{P(x)P(y)}$$

그러나, x와 y가 종속적일 경우는 종속의 정도에 따라 1보다 작아지게 된다. 즉, x와 y 각각의 지지도는 커지고, x와 y에 대한 신뢰도가 작아질수록 흥미도 또는 경이도는 커지게 된다[6, 15, 20, 23]. 이는, 앞서 2.1.3에서 정의된 세가지 뷰 중에서 URQV의 흥미도가 가장 높게 나타나게 된다는 것이다.

4.5 분포도(Variance)

통계학에서 사용되는 변동(variability)의 측정치로 사용된다. n개의 튜플을 갖는 $R=\{x_1, C_1>, <x_2, C_2>, \cdots, <x_n, C_n>\}$이 주어지는데, 각 x_i는 일반화된 애트리뷰트로 되어 있는 유일한 튜플이고 각 C_i는 각 x_i의 수를 표기한다. R에서 발생하는 각 x_i의 확률이 다음과 같이 주어진다[14].

$$p(x_i)\frac{c_i}{(c_1+c_2+\ldots+c_n)}$$

균등하게 분포된 튜플에서 각 t_i의 확률은 다음과 같다.

$$q(x_i)=\frac{(c_1+c_2+\ldots+c_n)}{n}=\frac{1}{(c_1+c_2+\ldots+c_n)}=\frac{1}{n}$$

균등분포로부터 R에서 튜플의 분포도는 다음과 같이 정의된다.

$$V(x)=\frac{\sum_{i=1}^{n}(p(x_i)-q(x_i))^2}{n-1}$$

4.6 상관도(Correlation)

두 item x와 y간의 상호 관계성을 나타내는 지표로, 이전의 x와 y의 독립성에 근거한 흥미도 또는 경이도와는 다르다고 할 수 있다. 흔히 통계학에서 정의된 상관도는 다음의 상관계수에 근거한다고 할 수 있다.

$$R(x, y)=\frac{cov(x, y)}{SD(x)SD(y)}$$

여기서, cov(x, y)는 두 item x와 y에 관한 covariance를 의미하고, SD(x)는 x에 관한 표준편차(standard deviation)를 의미한다. 이는, 앞서 2.1.3에서 정의된 세가지 뷰 중에서 PRQV의 상관도가 가장 높게 나타나게 됨을 알 수 있다.

4.7 유사도(Similarity)

많은 데이타 탐색과 데이타 마이닝 애플리케이션에서 데이타 객체사이의 유사도는 중요한 문제이다. 복잡한 데이타 객체에서 유사도는 데이타의 Nontrieval문제를 정의하는 것이다.

두개의 시계열 데이타 $X=(x_1, x_2, \cdots, x_n)$, $Y=(y_1, y_2, \cdots, y_n)$와 숫자 $0<\gamma, \varepsilon<1$이 주어지고 순차식 X, Y가 (F, γ, ε)-similar하다는 것은 함수 $f\in F$와 부분 순차식 $X_f=(xi_1, xi_2, \cdots, xi_n)$, $Y_f=(y_1, y_2, \cdots, y_n)$가 존재한다는 것이다[8].

$y_{jk}/(1+\varepsilon) \le ax_{ik}+b \le y_{jk}/(1+\varepsilon)$, where $i_k \le i_{k+1}$, $y_k \le y_{k+1}$ for all $k=1, \cdots, \gamma n-1$, $\forall k, 1 \le k \le \gamma n$.

파라매타 $\gamma(0<\gamma\le1)$는 Y와 매핑되는 부분 순차식 X의 길이를 제어하기 위해 사용하고 ε은 순차식간의 밀접함을 제어하기 위해 사용된다. y_{jk}, x_{ik}가 상위의 조건을 만족한다면 ε-close하다고 말할 수 있다.

또한 X, Y, F, ε이 주어질 때, X와 Y의 유사도는 다음과 같이 표현할 수 있다.

SimF, $\varepsilon(X, Y) = \{\max\gamma \mid X, Y \text{ are } (F, \gamma, \varepsilon\text{-simllar}\}$

SimF, $\varepsilon(X, Y)$은 0과 1사이의 숫자이고 1과 가까운 숫자일수록 유사도를 더 만족한다고 할 수 있다[8].

5. 결 론

데이타베이스가 대용량화됨에 따라 저장되는 데이타의 양은 폭발적으로 증가하였으나, 이를 효과적으로 활용할 수 있는 정보와 지식이 부족하고, 관련 응용 도메인에 적절한 규칙성을 반영하지 못하고 있었다.

본 논문에서는 대용량 데이타베이스에 존재하는 여러 유용한 지식을 추출하는 방법으로서 분류화, 클러스터링, 요약규칙, 시간에 따른 분석 및 예측 등을 요약 제시하였고, 이렇게 추출된 패턴, 정보, 지식들의 유용성을 측정하는 지표를 정리하였다. 각각의 마이닝 방법에 따른 유용성의 정도가 달라지겠고, 경우에 따라서는 지표의 정량적 특성도 다르게 표현될 것이다.

더 나아가, 데이타 마이닝 기법뿐만 아니라 데이타의 샘플링과 성능향상[2]을 통하여 방대한 데이타로부터 다양한 지식 탐사가 가능해지고, 발견된 규칙 또는 지식의 유용성 측정을 통하여 업무 분야의 특성에 따라 효과적으로 반영되며 의사 결정 및 마케팅, 동향 분석 및 예측 등에 유용한 정보로 제공될 것이다.

참고문헌

[1] R. Agrawal, T. Imielinske, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", Proc. ACM SIGMOD, 1993.

[2] R. Agrawal, T. Imielinski, and, A. Swami, "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning and Discovery in Knowledge-Based Databases, Vol. 5, No. 6, December 1993.

[3] D. Barbara, and M. Sullivan, "Quasi-Cubes: A Space-Efficient Way to Support Approximate," submitted for publication, 1998.

[4] B. Bollobas, G. Das, D. Gunopulos, and H. Mannila, "Time-Series Similarity Problems and Well-Separated Geometric Sets", 13th Annual ACM Symposium on Computational Geometry Nice, 1997.

[5] P.S. Bradley, U. M. Fayyad, O.L. Mangasarian, "Data Mining: Overview and Optimization Opportunities", http://elib.stanford.edu, Technical Report MP-TR-98-01, 1998.

[6] S. Brin, R. Motwami, J. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", Proc. SIGMOD, pp. 255-264, 1997.

[7] M.S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Tran. Knowledge and Data Engineering, Vol 8, No 6, Dec 1996.

[8] G. Das, D. Gunopulos, and H. Mannila, "Finding Similar Time Series", PKDD '97, 1997.

[9] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "The KDD process for Extracting Useful Knowledge from Volumes of Data", Comm. of the ACM, vol.39, No. 11, Nov 1996.

[10] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advanced in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[11] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and Pirahesh, "Data Cube" A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals", Data Mining and Knowledge Discovery, Kluwer Academy Publishers, Vol.1, No.1, pp. 29-53, 1997.

[12] D. Heckerman, "Bayesian Networks for Data Mining", Data Mining and Knowledge Discovery, Kluwer Academy Publishers, Vol.1, No.1, pp. 79-119, 1997.

[13] C. Hidber, "Online Association Rule Mining", http://elib.stanford.edu, Technical Report CSD-98-1004, 1998.

[14] R.J. Hilderman, H.J. Hamilton, and N. Cercone, "Data Mining in Large Databases Using Domain Generalization Graphs", Dept. of CS, Univ. of Regina, submitted for publication, 1998.

[15] M. Holsheimer and A. Siebes, "Data Mining : The Search for Knowledge in Databases", Report CS-R9406, ISSN 0169-118X, CWI(Centrum voor Wiskunde en Informatica), The Netherland, 1994.

[16] D. Koller, "From Knowledge to Belief", http://sunsite.berkeley.edu, Technical Report CS-TR-94-1527, Univ. of California, Berkeley, Oct 1994.

[17] O. Lee, "MTLS : A Tool for Extending and Refining Knowledge Bases", Tech nical Report CS-TR-97-, George Mason University, May 1997.

[18] H. Mannila, H. Toivonen, and I. Verkamo, "Efficient Algorithms for Discovering Association Rules", AAAI Workshop on Knowledge Discovery in Databases, Eds. Usama M.Fayyad and Ramasamy Uthurusamy, pp. 181-192, Seattle, Washington, July 1994.

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

[20] G. Nakhaeizadeh, C. Taylor, and C. Lanquillion, "Evaluating Usefulness for Dynamic Classification", 4th Int. Conf. on Knowledge Discovery and Data Mining New York, 1998.

[21] B. Padmanabhan, and A. Tuzhilin, "A Belief-Driven Method for Discovering Unexpected Pattern", 4th Int. Conf. on Knowledge Discovery and Data Mining New York, 1998.

[22] R. Srikant, R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", Proc. ACM SIGMOD, Quebec, Jun. 1996.

[23] K. Wang, S.H. William Tay, and B. Liu, "Interestingness-based Interval Merger for Numeric Association rules", 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, 1998.

[24] J. Yoon and L. Kerschberg, "Query-Initiated Discovery of Interesting Association Rules", To appear in the 1st Int' l Conference on Discovery Science, Fukuoka, Japan, 1998.

[25] T. Zhang, "Data Clustering for Very Large Datasets Plus Applications", A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Dept. of CS, Univ. of Wisconsin, 1997.

윤 종 필

1981 연세대학교 전기공학 학사
1987 University of Florida 컴퓨터 공학 석사
1993 George Mason University 전산학 박사
1994~현재 숙명여자대학교 전산학과 조교수
관심분야 : Data Mining, 분산 데이타베이스, 멀티미디어 데이타베이스, 에이전트 기반 데이타베이스, 능동 데이타베이스
E-mail : jyoon@sookmyung.ac.kr

김 희 숙

1989 성신여자대학교 전산학과 학사
1991 숙명여자대학교 전산학과 석사
1997~현재 숙명여자대학교 전산학과 박사과정
관심분야 : Data Mining, OLAP, 분산 데이타베이스
E-mail : hskim@cs.sookmyung.ac.kr

최 옥 주

1987 숙명여자대학교 전산학과 학사
1990 숙명여자대학교 전산학과 석사
1990~1996 LG생산기술원 주임연구원
1997~현재 숙명여자대학교 전산학과 박사과정
관심분야 : Data Warehousing, Data Mining, Time-Series
E-mail : ojchoi@cs.sookmyung.ac.kr