데이터마이닝을 이용한 웹 데이터 분석

채승경* · 서용무**

Analysis of Web Data Applying Data Mining

요 약

인터넷의 확산으로 웹 구조, 웹 로그 등을 분석하는 웹마이닝(Web Mining)에 대한 연구가 활발히 진행되고 있다. 그러나 웹에서 발생하는 데이터에 대한 분석은 아직 미약한 상태이다. 웹에서 획득된 데이터는 신뢰도가 낮아 통계와 같은 기존의 분석 방법을 적용하기에 많은 어려움이 따른다. 또한 대용량 데이터와 실제 데이터에 유연한 분석을 제공하는 데이터마이닝은 아직까지 적용 분야가 매우 한정되어 있다.

본 논문에서는 인터넷 사이트의 실제 데이터를 이용하여 데이터마이닝 과정에 따라 데이터 정제, 데이터 선택, 데이터 변환 등 효과적인 데이터 전처리 방법을 제시한다. 또한 이렇게 전처리된 데이터로 고객 세분화, 우수 고객 분류를 위한 데이터마이닝 기법을 적용한 후 수행 결과를 분석한다. 마지막으로 분석의 한계점을 지적하고 보다 양질의 데이터마이닝을 위한 시스템 및 사이트 설계 방안을 제시한다.

Key words : 데이터마이닝, 지식발견 프로세스, 연관규칙, 인공신경망, 의사결정나무

I 서론

인터넷의 확산으로 웹상에서 생성되는 데이터가 증가함에 따라 데이터에 대한 분석 요구가 증가하였고 최근에 웹 구조, 웹 로그 등을 분석하는 웹마이닝에 대한 연구가 활발히 진행되고 있으나 웹에서 입력되는 데이터에 대한 분석은 아직도 미약한 상태이다.

웹에서 획득된 데이터는 신뢰도가 낮아 좋은 분석 결과를 기대하기 힘들며 통계와 같은 기존의 분석 방법을 적용하기에는 많은 어려움이 따르며 대용량 데이터와 실제 데이터에 유연한 분석을 제공하는 데이터마이닝은 아직까지 적용 분야가 매우 한정되어 있다.

* 고려대학교 대학원 경영학과 석사과정 (sgchae@korea.ac.kr)
** 고려대학교 경영대학 교수 (ymsuh@yahoo.com)

본 연구에서는 웹에서 입력된 데이터에 데이터마이닝 기법을 적용한 후 결과에 대한 분석과 실제 적용 가능성, 문제점 등을 알아 본다. 또한 보다 양질의 분석을 위한 사이트 및 시스템 설계 방안을 제시하려 한다.

본 논문의 구성은 다음과 같다.

제 2 장에서는 데이터마이닝에 대한 개념, 데이터마이닝 과정, 데이터마이닝 수행 업무에 대해 살펴본다. 또한 연관규칙 탐사, 사례기반추론, 자동군집추출, 의사결정나무, 인공신경망 등의 데이터마이닝 기법에 관한 내용을 알아본다.

제 3 장에서는 인터넷 경매 사이트의 실제 데이터를 가지고 고객 세분화, 우수고객 분류 등의 데이터마이닝을 수행하기 위한 연구 절차, 데이터 처리, 모델 구축에 관해 설명한다.

제 4 장에서는 데이터마이닝 수행 결과를 살펴보고 의미를 분석한 후 한계점과 효과적인 데이터마이닝을 위한 사이트 및 시스템 설계에 관한 방안을 제시한다.

제 5 장에서는 연구 결과를 종합하고 전체적인 연구의 한계점과 향후 연구 방향을 제시한다.

II 데이터마이닝

2.1 데이터마이닝의 정의

데이터마이닝(Data Mining)이란 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 90년대 초반부터 지식발견(Knowledge Discovery in Databases), 정보발견, 정보수확의 이름으로 소개되어 왔는데 일반적으로 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하여 의사 결정에 활용하는 작업이라 정의된다[장남식외 2인 1999, 조재희와 박성진 1999].

데이터마이닝과 지식발견이란 용어의 정확한 의미에 대해서는 의견이 분분하다. 1995년 몬트리올 국제 지식발견 학술대회에서 지식발견은 데이터로부터 지식을 추출하는 전 과정을 설명하는 뜻으로 해석되었다. 이때 지식이란 데이터 요소들 간의 연관성이나 패턴을 의미한다. 그리고 데이터마이닝은 지식발견 과정에서 탐사 단계만을 뜻하는 의미로 사용한다고 제안되었다[장남식외 2인 1999, 조재희와 박성진 1999].

2.2 데이터마이닝 과정

데이터마이닝은 일반적으로 데이터 선택, 데이터 정제, 데이터 변환, 데이터마이닝, 패턴 평가, 지식 표현의 6단계로 되어 있다[그림1 참조]. 데이터마이닝 과정, 순서, 용어 등은 문헌에 따라 분분하나 본 논문에서는 몇몇 기존의 연구 내용을 종합하여 정의하였다[Bigus 1996, Fayyad *et al* 1996, Famili *et al* 1997, Han 1999].

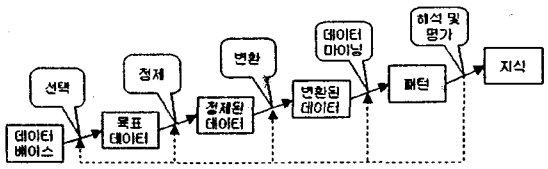

그림1: 데이터마이닝 과정

데이터 선택, 데이터 정제, 데이터 변환 등의 데이터 준비 과정은 반자동화된 많은 시간을 요구하는 작업이다. 데이터 준비 과정이 필요한 이유는 데이터 분석 수행을 방해하는 데이터의 문제를 해결하고 데이터의 특성을 이해하여 더 의미 있는 데이터 분석을 수행할 수 있음으로 인해 주어진 데이터 집합에서 더욱 의미 있는 지식을 추출할 수

있기 때문이다[Famili *et al* 1997].

2.2.1 데이터 선택(Data Selection)

데이터마이닝 수행 시 어떤 데이터가 가장 중요한 지 결정해야 하는데 데이터 선택은 두 차원에서 이루어진다. 첫째, 데이터마이닝의 대상이 되는 열(Column) 혹은 모수(Parameter)의 선택이고 둘째, 각 필드값에 기반한 행(Row) 혹은 레코드의 선택이다[Bigus 1996].

데이터 항목(Attribute)의 수가 많을 경우 데이터마이닝 수행 시 모델 구축 및 결과를 도출하는데 많은 시스템 자원과 시간이 필요하므로 상관분석, 카이제곱 검정, t-통계량, 단계적 변수 선정(Stepwise Variable Selection) 등을 통해 적절한 항목을 선택해야 할 필요도 있다[김영만 1998, 장남식외 2인 1999].

2.2.2 데이터 정제(Data Cleaning)

데이터 정제는 부정확한 값, 결손값(Missing Values), 불일치(Inconsistency), 잡음(Noisy) 등을 제거하고 데이터의 범위를 벗어난 데이터 및 특이값을 추출하는 단계이다[Bigus 1996].

시각화(Visualization)는 대용량의 데이터에서 범위를 벗어난 데이터를 쉽게 찾을 수 있게 한다. 또한 통계적 정보를 이용하여 결손 필드값이나 오류가 있는 필드값을 중간값이나 적절한 값으로 대체할 수 있다[Bigus 1996].

데이터의 수가 적고 상대적으로 결손값이 많을 때엔 결손값 제거는 데이터 분석을 복잡하게 하고 정확성을 저하시킨다. 또한 결손값은 그 자체로 중요한 정보를 포함하고 있는 경우가 많다. 결손값을 중간값 등으로 대체하는 경우 결손값의 부정확하고 불일치한 결과를 초래할 수 있다[Famili *et al* 1997, Wright 1998]. 결손값을 무시하는 경우 모든 결손값이 동일한 것으로 가정되어 군집화 등의 기법 수행 시 악영향을 주게 된다[Wright 1998]. 결손값 문제를 효과적으로 해결하기 위해 결손값의 비율을 계산하고 20%이상의 결손값을 가지는 레코드를 제거하는 휴리스틱 레코드 사용성(Heuristic Record Usability)이 제시되었는데 이 방법은 가치 있는 데이터까지 제거할 수 있는 단점을 가지고 있다[Famili 1997, Wright 1998]. Famili *et al*은 레코드의 사용성을 결정하기 위해 각 필드에 가중치를 할당함으로써 휴리스틱 방법보다 더욱 개선된 방법을 제시하였다[Wright 1998].

2.2.3 데이터 변환(Data Transformation)

데이터 변환 단계에서는 이미 존재하는 필드로부터 새로운 데이터 필드를 생성하거나 더 많은 정보를 포함하도록 몇 개의 필드를 하나의 필드로 변환하는 등의 작업을 수행하고 선택된 데이터가 특정 데이터마이닝 알고리즘에 수행에 적당하도록 데이터 값을 변형한다[Bigus 1996]. 예를 들어, 일반적으로 인공신경망은 데이터가 0에서 1 사이, 혹은 -1에서 1 사이일 때에 가장 잘 수행된다. 많은 데이터마이닝 도구들이 내부적으로 데이터 변환 기능을 제공하지만 그런 기능이 없을 경우 사용자가 데이터의 범위를 변환시켜 주어야 한다.

2.2.4 데이터마이닝(Data Mining)

데이터 패턴을 추출하기 위해서 실제 데이터마이닝 알고리즘이 적용되는 단계이다. 마이닝 하는 작업의 유형에 따라 연관규칙, 군집화, 의사결정나무, 인공신경망 등의 알

고리즘이 사용될 수 있으며 하나 이상의 기법들이 사용되는 것이 일반적이다[조재희와 박성진 1999].

2.2.5 패턴 평가

도메인 전문가와 협력하여 마이닝된 결과를 해석하고 발견된 패턴이 가치가 있는 지, 실제로 적용 가능한지를 평가하는 단계이다. 경우에 따라서는 앞 단계로 돌아가 마이닝을 재수행 하기도 한다.

2.2.6 지식 표현

실제로 발견된 패턴을 지식화 하거나 문서화 하고 관련 부서에 보고하는 단계이다. 또한 기존에 추출된 지식과 상충하는 지 검사하고 해결하는 과정도 포함된다.

데이터마이닝 과정에는 응용 도메인, 데이터, 환경적 특성에 관한 도메인 지식(Domain Knowledge)도 매우 중요한데 Anand *et al*은 도메인 지식을 세가지로 분류하고 데이터마이닝에서의 도메인 지식의 역할에 대해 논의하였다[Anand *et al* 1995].

2.3 데이터마이닝 수행작업

일반적으로 실용적인 데이터마이닝은 제한된 환경 하에서 제한된 작업을 수행하는데 이러한 작업은 분류, 추정, 예측, 세분화, 설명으로 압축될 수 있다[Berry & Linoff 1997].

2.3.1 분류(Classification)

분류는 가장 일반적인 데이터마이닝 작업으로 새로이 제시된 데이터를 조사하고 그 데이터를 미리 정의된 클래스에 지정하는 작업 유형이다. 분류는 부류 값이 포함된 과거의 데이터로부터 부류별 특성을 찾아내어 분류모형을 만들고 이를 토대로 새로운 레코드의 부류 값을 예측하는 것을 의미한다. 분류에 사용되는 데이터마이닝 기법으로는 의사결정나무, 사례기반추론 등이 있다.

2.3.2 추정(Estimation)

추정은 미리 입력된 데이터를 가지고 알려지지 않은 연속형 변수를 추정하는 것이다. 추정은 종종 분류의 작업을 수행하는데 분류가 이산적인 결과를 다루고 있는 반면 추정은 연속적인 결과를 다룬다. 추정에는 인공신경망이 잘 적용된다.

2.3.3 예측(Prediction)

예측은 미래의 행동을 예견하거나 미래의 가치를 평가하는 것만 제외하고는 분류나 추정과 같다. 예측은 주로 과거의 데이터를 통해 수행되는데 사례기반추론, 의사결정나무, 인공신경망 등의 데이터마이닝 기법이 사용된다.

2.3.4 세분화(Segmentation)

데이터마이닝에서 세분화와 군집화(Clustering)의 용어가 비슷한 개념으로 사용되어 왔으나 Berry[1998]는 세분화를 군집화와 연관규칙(Association Rule)을 포함하는 상위 개념으로 정의하였다[Westphal & Blaxton 1998].

연관규칙은 동시에 발생하는 사건 그룹 내에서 사건들 사이에 존재하는 친화성(Affinity)이나 패턴을 발견하는 작업을 말한다. 연관규칙을 발견하는 작업이란 데이터 안에 존재하는 항목간의 종속(Dependency)관계를 찾아내는 작업이며, 마케팅 분야에서

는 장바구니 분석(Market Basket Analysis), 기계학습(Machine Learning) 분야에서는 규칙 유도(Rule Induction)라고도 한다.

연관규칙에 시간과 관련된 제약이 포함된 형태를 순차패턴(Sequential Pattern)이라고 하는데 순차패턴은 연관규칙과 달리 시간적 순서가 고려된다.

군집화(Clustering)란 다양한 특성을 지닌 관찰대상을 유사성을 바탕으로 동질적인 집단으로 분류하는 데 사용되는 기법이다. 분류가 어느 한 레코드가 미리 정의된 계층에 지정되는 것에 비해 군집화는 사전에 정의된 계층이 없이 레코드의 유사성에 기반하여 그룹화 되는 것이다. 분류규칙이 불명확하거나 또는 집단의 수를 미리 정하지 않는 경우에는 군집화가 매우 유용하며 다른 데이터마이닝 작업을 위한 선행작업으로서의 역할을 수행하는 경우가 많다. 군집화에 효과적으로 적용될 수 있는 데이터마이닝 기법에는 k-평균(k-Means) 기법 등과 인공신경망의 한 종류로 코호넨 망(Kohonen Network)이 있다.

2.3.5 설명(Description)

데이터마이닝의 목적이 단순히 사람, 상품, 프로세스에 대한 이해를 높이기 위해 복잡한 데이터베이스에 대한 묘사가 될 수도 있는데 설명은 데이터에 대한 간결한 요약 또는 다른 데이터와의 구별을 수행한다. 설명에는 데이터에 대한 요약을 수행하는 특성화(Characterization), 데이터 집합간의 비교를 나타내는 비교(Comparison) 혹은 판별(Discrimination)이 있다[Han 1999]. 설명은 빈도, 합계, 평균 등의 요약 특성 뿐만 아니라 데이터의 분산, 사분위수(Quartile) 등의 분포 특성도 포함하여야 한다[Han 1999].

2.4 데이터마이닝 기법

데이터마이닝에 특정 작업에 적용하는 기법이 정해진 것은 아니고 하나의 기법만으로는 모든 문제를 해결할 수 있는 것도 아니다. 데이터마이닝 기법의 선택은 데이터마이닝이 수행하는 작업과 목적, 분석에 이용되는 데이터의 특성, 발견된 패턴의 설명력, 사용의 용이성 등에 따라 달라질 수 있다. [표1]은 데이터마이닝 기법과 그 기법이 지원하는 데이터마이닝 작업을 나타낸 것이다[Berry & Linoff 1997].

표1: 데이터마이닝 작업과 기법

수행 작업 / 기법	분류	추정	예측	세분화		설명
				연관규칙	군집화	
통계	√	√	√	√	√	√
연관규칙 탐사			√	√	√	√
사례기반 추론	√		√	√	√	
자동군집 검출					√	
의사결정 나무	√		√		√	√
인공 신경망	√	√	√		√	

이 절에서는 데이터마이닝의 기법들과 그 기법들이 어떠한 작업을 수행하는 지에 대해서 설명하려 한다.

2.4.1 연관규칙(Association Rules) 탐사

연관규칙이란 데이터로부터 항목(Item)의 연관성 정도를 측정하여 연관성이 많은 항목을 그룹화하는 세분화(Segmentation)의 일종이며 하나의 거래(Transaction)에서 함께 일어나는 항목들의 그룹을 찾아내는 자율학습(Unsupervised Learning)의 한 형태이다. 연관규칙은 마케팅 분야에서는 장바구니 분석(Market Basket Analysis), 기계학습(Machine

Learning) 분야에서는 규칙 유도(Rule Induction)라고도 한다.

연관규칙은 'If X then Y (X→Y)'와 같은 형태로 표현되고 '조건부(Condition) 항목집합 X가 거래에 나타난 경우 결과부(Result) 항목집합 Y도 나타난다'와 같은 방법으로 해석된다.

하나의 연관규칙 A → B에 대하여 지지도(Support)란 전체 거래 중에서 조건부 A와 결과부 B 항목을 모두 포함하는 거래의 수를 나타낸다.

$$Supp(A \rightarrow B) = p(A \cap B)$$

신뢰도(Confidence)는 조건부 A 항목을 포함하는 거래 중에서 결과부 B항목이 포함된 거래는 어느 정도인가를 나타낸다.

$$Conf(A \rightarrow B) = \frac{p(A \cap B)}{p(A)}$$

리프트(Lift)란 신뢰도를 결과부 B항목을 포함하는 거래수로 나눈 값으로 리프트가 1보다 크면 클수록 품목간에 양의 상관관계가 많으므로 유용한 연관규칙이라고 말할 수 있다. *Berry & Linoff*은 리프트 대신 개선도(Improvement)라는 용어를 사용하기도 하였다[Berry & Linoff 1997].

$$Lift(A \rightarrow B) = \frac{p(A \cap B)}{p(A)p(B)}$$

연관규칙 탐색은 먼저 최소지지도(Minimum Support)를 만족하는 규칙을 찾은 후 그 규칙들의 신뢰도를 산출하고 최소신뢰도(Minimum Confidence)를 만족하는 규칙 중에 일반적으로 리프트가 1보다 큰 규칙을 유용한 연관규칙으로 채택한다.

2.4.2 순차패턴 탐사

순차패턴 탐사는 한 거래 안에서 발생하는 항목들간의 연관 규칙에 시간의 제약 조건을 추가한 것이다. 즉, 연관규칙이 같은 거래 레코드 안에서의 항목들에 대한 관계를 발견하는 내부 거래 연관규칙(Intra-Transaction Association Rules)인 반면 순차패턴은 다른 거래 레코드 사이의 항목들에 대한 관계를 발견하는 상호 거래 연관규칙(Inter-Transaction Association Rules)이다[Lu *et al* 1998, Feng *et al* 1999].

"IBM과 SUN의 주가가 오르면 같은 날 HP의 주가도 오를 확률이 높다."가 연관규칙의 예라면 순차패턴은 "IBM과 SUN의 주가가 오르면 *다음 날* HP의 주가도 오를 확률이 높다."와 같다. 또한 "IBM과 SUN의 주가가 오르면 *3일 이내*에 HP의 주가도 오를 확률이 높다."의 규칙도 발견할 수 있을 것이다.

2.4.3 사례기반추론(Case-Based Reasoning)

사람들은 문제해결을 위해 종종 자신의 과거 경험을 활용하는 경우가 많다. 사례기반추론은 과거의 데이터를 바탕으로 새로운 레코드를 분류하거나 예측하는 기법이다. 먼저 사례기반추론에서는 동일한 유형의 레코드들은 서로 가까이 있을 것이라고 가정한다. 다시 말하면 동일한 타입의 레코드는 데이터 공간에서도 서로 가까이 있다는 것이다. 사례기반추론의 가장 일반적인 기법으로는 k-최단 인접 이웃(k-Nearest Neighbor) 기법이 있다. k-최단 인접 이웃 기법은 새로운 데이터가 나타났을 때 각각의 모든 데이

터와의 거리를 계산한 후 새로운 데이터와 가장 가까이 있는 데이터에 기반하여 새로운 데이터를 분류하거나 예측한다. 사례기반추론은 수치형 값과 순서형 값에 대해서는 그리 좋은 결과를 얻을 수 없는데 이 때 양질의 결과를 얻기 위해 사용되는 방법이 통계의 회귀분석이다.

2.4.4 자동군집검출

자동군집검출(Automatic Cluster Detection)은 데이터마이닝의 군집화 작업에 주로 사용되는 기법으로 가장 일반적인 기술로는 k-평균군집화(k-Means Clustering)가 있는데 k-평균군집화는 N개의 속성으로 구성되는 각각의 레코드를 벡터로 표시하여 N차원의 데이터 공간에 나타낼 때, 유사한 특성을 갖는 레코드들은 서로 근접하여 위치한다는 가정에 근거한다. 자동군집검출은 군집 분석 외에도 분류, 예측을 위한 선행작업, 특히 오류값이나 결손값 처리작업 등 다양한 분석에 사용할 수 있다.

2.4.5 의사결정나무(Decision Trees)

의사결정나무는 데이터마이닝의 분류 작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류 모형을 나무의 형태로 만드는 것이다. 그리고 이렇게 만들어진 분류 모형은 새로운 레코드를 분류하고 해당 부류 값을 예측하는 데 사용된다.

의사결정나무 분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법은 통계에 기반을 하고 있다. 의사결정나무 분석의 대표적인 알고리즘에 CHAID가 있는데 CHAID는 카이제곱-검정 또는 F-검정을 이용하여 다원 분리(Multiway Split)를 수행하는 알고리즘이다. CHAID는 목표변수가 이산형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량을 분리기준으로 사용한다.

의사결정나무의 또 다른 대표적인 알고리즘 중에 C4.5가 있는데 C4.5는 가지치기를 하는 데 있어서 베르누이의 이항분포를 사용한다. 즉, 주어진 신뢰수준에 대한 신뢰구간과 기대값의 범위를 가지고 아직 알려지지 않은 데이터에 대한 에러율을 계산함으로써 가장 적합한 수준의 의사결정나무를 구축하는 것이다.

2.4.6 인공신경망(Artificial Neural Network)

데이터마이닝에 이용되는 한 기법으로서의 인공신경망은 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 모방하여 자신이 가진 데이터로부터의 반복적인 학습과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써 특히 향후를 예측하고자 하는 문제에 있어서 유용하게 적용되는 기법이다.

신경망은 [그림2]와 같이 입력계층(Input Unit), 출력계층(Output Unit), 그리고 은닉계층(Hidden Unit)으로 이루어져 있다.

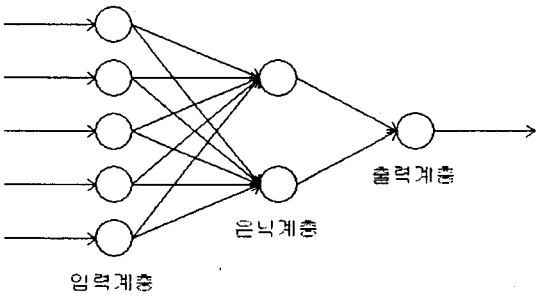

그림2: 인공신경망의 구조

입력계층은 결과변수를 설명하는데 이용하고자 하는 입력변수들이고 출력계층은 예측치를 얻고자 하는 결과변수이다. 두 개 이상 놓여질 수 있는 은닉계층은 인간의 신경망을 모형화한 몇 개의 은닉마디로 이루어져 있다. 각 은닉마디와 은닉계층에는 활성함수(Activation Function)를 지정하게 되는데 활성함수는 조합함수(Combination Function)와 전송함수(Transfer Function)로 이루어져 있다. 조합함수가 입력으로 들어오는 모든 데이터의 값을 하나의 값으로 통합시키면 전송함수가 통합된 값을 변환시켜 다른 마디나 계층으로 전송한다. 전송함수에는 선형이동함수(Linear Transfer Function)와 비선형이동함수(Non-linear Transfer Function)가 있다. 선형이동함수에는 선형회귀분석이 사용되고 비선형이동함수에는 로지스틱 회귀분석, Hyperbolic Tangent 함수 등이 사용된다.

Ⅲ 인터넷 경매 사이트 데이터마이닝

3.1 태스크의 목적

인터넷 경매 사업을 하고 있는 ○사는 다른 많은 사이트와 제휴를 통해 새로운 방식의 네트워크 경매 모델을 제시한 사이트이다. 이 사이트는 2000년 6월 당시, 28,000여 명의 회원을 가지고 있으며 매일 5,000~10,000여 명이 방문한다. 등록된 전체 물품 수는 40,000여 개이고 매일 300~500여 개의 물품이 새로 등록되고 있다. 또한 매일 등록된 물품에 참여하는 입찰은 1,000여 회에 달하며 낙찰로 이어지는 물품은 10~40여 개이다.

인터넷 경매 사이트의 데이터 특성을 고려하여 볼 때 다음과 같은 몇 개의 데이터 마이닝 작업이 가능하다.

첫째, 회원 데이터를 통한 고객 세분화이다. 회원 데이터에 군집화 기법을 적용함으로써 회원의 정보를 보다 자세하게 파악하여 마케팅에 적용할 수 있으며 어느 한 군집에 속한 회원들이 주로 참여하는 경매 물품을 추천하는 등 개인화(Personalization)와 추천 시스템(Recommender System)에도 적용이 가능하다.

둘째, 회원 데이터와 입찰/낙찰 회수, 낙찰 금액 등의 데이터 분석을 통한 우수고객의 분류이다. 우수 고객의 특성을 분석함으로 잠재 고객 마케팅에 적용할 수 있으며 고객의 충성도(Loyalty) 프로그램 개발에 도움이 될 수 있다.

셋째, 판매/입찰 항목간의 관계 분석이다. 고객의 거래 데이터를 분석하여 연관규칙을 발견하고 관심을 갖을 만한 물품을 추천함으로써 고객이 경매에 참여하도록 유도할 수 있게 한다. 또한 이를 통해 시간적인 요소를 고려하여 이벤트를 실시할 수도 있다.

본 논문에서는 k-평균기법, 코호넨 망을 이용한 고객 세분화와 의사결정나무, 인공신경망 기법을 이용한 우수고객 분류 작업만을 수행하였다.

3.2 데이터마이닝 절차

본 연구를 위해서 회사로부터 제공 받은 데이터는 회원 데이터, 물품 데이터, 입찰 데이터, 주문 데이터로 구성되어 있다.

먼저 각 데이터의 기본키(Primary Key) 및 외래키(Foreign Key)를 찾아내 필드의 속성을 규명하여 데이터베이스의 구조를 파악한 후 이 데이터를 가지고 가능한 데이터마이닝 작업을 정의하고 세부 계획을 수립하였다.

다음으로 텍스트 형식으로 된 데이터를 관계형 데이터베이스로 구축하고 데이터 선택, 데이터 정제, 데이터 변환 등의 전처리 과정을 거쳐 데이터마이닝을 수행하였다.

마지막으로 수행한 데이터마이닝 결과에 대한 해석과 실제 적용 가능성, 문제점 등과 더욱 효과적이고 효율적인 데이터마이닝을 위한 사이트 및 시스템 설계에 대한 방안을 제시하였다.

3.3 데이터 준비

3.3.1 데이터베이스 구축

보다 용이한 데이터 전처리와 데이터마이닝을 위해 텍스트 형식으로 되어 있는 데이터를 관계형 데이터베이스에 로딩하여 데이터베이스를 구축하였다.

○사의 데이터는 회원 데이터, 물품 데이터, 입찰 데이터, 주문 데이터로 구성되어 있다.회원 데이터와 입찰 데이터는 각 필드가 '|' 파라미터로 구분되어 있어 별다른 변환 과정 없이 데이터베이스에 각각 테이블로 로딩하였다.

그러나 물품 데이터와 주문 데이터는 각 레코드들이 '+' 기호로 분리되고 각 필드는 한 줄에 표현되어 있다. 특히 물품테이블은 텍스트 상자 입력 형식의 필드와 HTML 소스 코드 필드를 포함하고 있어 복잡한 변환 과정을 필요로 하였다.

두 데이터는 C언어로 프로그래밍해서 관계형 데이터베이스에 로딩(Loading)이 가능한 형태로 변환시켰다.

데이터베이스에 로딩 후 각 테이블의 기본키와 외래키를 지정하여 테이블간의 관계를 구축하였다.

3.3.2 데이터 선택

데이터마이닝에 적용될 데이터를 선택하기 위해 각 테이블에 SQL로 질의하여 새로운 테이블을 생성하였다. 그 결과 27,598개의 레코드를 갖는 테이블이 생성되었다.

3.3.3 데이터 정제

일반적으로 텍스트 형식을 갖는 데이터는 데이터마이닝에 사용하기에 적합하지 않으므로 회원테이블의 성명 필드, 물품테이블의 물품요약정보, 교환정보, 물품상세설명 등의 필드, 주문테이블의 모델명제조업자, 주문시 요청사항, 배송시 요청사항 등의 필드, 입찰테이블의 설명 필드 등은 제거하였다.

결손값(Missing Value)을 갖는 14,628개의 레코드도 제거하여 12,970개의 레코드를 갖는 테이블을 구축하였다. 결손값을 가지는 레코드 수가 많은 이유는 데이터마이닝에 유용할 만한 필드가 회원정보 입력 시 선택항목인 경우가 많아서 사용자들이 일반적으로 잘 입력하지 않기 때문이다.

3.3.4 데이터 변환

결손값을 가지는 레코드를 제거했다고 할지라도 웹사이트에서 입력된 데이터, 특히 선택항목으로 입력된 데이터는 부정확하고 이 데이터를 이용한 마이닝의 결과도 유용하지 않을 확률이 높다. 따라서 비교적 정확한 정보를 가지고 있는 다른 필드에서 유용한 정보를 추출해야 한다. 현재 대부분의 사이트에서 유효 주민번호 프로그램을 채택하고 있어서 주민등록번호 정보는 대부분 정확하다고 가정할 수 있는데 주민등록번호를 이용하면 회원의 정확한 성별, 나이 등을 추출할 수 있다. 또한 입찰테이블과 주문테이블에 기반하여 각 회원별로 입찰회수,

낙찰회수, 낙찰총액 등의 필드를 생성하였다.

텍스트 형식의 주소 필드는 데이터마이닝에 사용될 수 없으므로 주소 필드에서 범주형 값을 갖는 거주지역 필드를 생성하였고 연속형 값을 갖는 나이 필드를 통해 범주형 값을 갖는 연령대 필드도 생성하였다.

3.4 모델구축

본 연구에서는 인터넷 경매 사이트 데이터로 고객 세분화와 우수고객 분류 작업을 수행하였다. 이를 위해 사용된 데이터마이닝 도구는 SPSS Clementine 5.2.1이며 Windows 2000 플랫폼(Platform)과 메모리 256M, Intel Pentium II CPU를 사용하였다

3.4.1 고객 세분화

고객의 특성을 파악하고 고객 세분화를 위해서 사용한 기법은 군집화로 k-평균 기법과 인공신경망의 일종인 코호넨 망(Kohonen Network)을 사용하여 모델을 구축하였다.

모델 구축을 위해 결혼여부, 차량소유여부 등의 이산형 변수, 직업유형, 수입, 주거형태, 인터넷경력, 성별, 연령대, 거주지역 등의 범주형 변수, 입찰회수, 낙찰회수, 낙찰총액 등의 연속형 변수가 입력값으로 사용되었으며 훈련용 데이터(Training Data)로 50%의 레코드를 임의로 추출하였다.

좀 더 효과적인 분석을 위해 k-평균 기법은 결과로 나타날 초기 군집 수를 5개로 지정하고 점차적으로 군집수를 증가시키고, 코호넨 망 또한 5×5 형태의 망을 기본으로 차원수를 증가시켰다.

3.4.2 우수고객 분류

우수고객 분류를 위해 사용한 기법은 C5.0 알고리즘으로 구현된 의사결정나무와 인공신경망 기법을 사용하여 모델을 구축하였다.

우수고객의 분류를 위한 모델 구축을 위해 출력값으로 사용될 목표변수가 필요한데 선택된 데이터에는 목표변수로 적용할 만한 데이터가 없다. 따라서 입찰회수, 낙찰회수, 낙찰총액 필드에서 목표변수 필드를 생성하였다.

입찰회수가 5회 이상, 낙찰회수가 3회 이상이고 낙찰총액이 10,000원 이상인 고객을 우수고객이라고 가정하고 이진값(Binary)을 갖는 우수고객 필드를 생성하여 출력값으로 하고 직업유형, 수입, 결혼여부, 차량소유여부, 주거형태, 인터넷경력, 성별, 연령대, 거주지역 등을 입력값으로 하여 모델을 구축하였는데 훈련용 데이터(Training Data)는 긍정적 데이터(Positive Data)와 부정적 데이터(Negative Data)의 비율을 고려하여 1,600여 개의 레코드를 갖는 데이터를 생성하였다.

Ⅳ 연구결과 및 연구방향 제시

4.1 연구 결과

4.1.1 고객 세분화

군집화 기법은 주로 대용량 데이터가 가지고 있는 정보를 파악하는 데 사용된다. 군집화 기법의 문제점은 신뢰성과 타당성에 관한 것인데 군집화 기법의 결과에 대한 신뢰성과 타당성을 검증하기가 어렵다.

군집의 수를 증가시켜 적은 구성요소를 갖는 군집을 분석해 보았는데 몇몇 군집의 정보가 부정확한 것으로 나타났다. 예를 들면, 어떤 군집의 직업이 주부가 40%인데 미혼이 80%, 남자가 70%로 나왔고 또 다른

군집은 월 수입이 500만 이상인데 직업이 학생 등과 같은 결과가 나왔다.

군집화를 통한 고객 세분화는 단지 고객의 유형을 파악하는 데 그치고 있는 한계를 가진다. 따라서 각 군집의 정보 만을 가지고는 실제 마케팅 전략에 적용하기는 어렵다.

4.1.2 우수고객 분류

훈련 데이터(Training Data)로 모델을 구축하여 전체 레코드에 적용해 본 결과, 의사결정나무의 분류 정확도는 64.1%로 상당히 낮았다.

인공신경망 모형의 예측 정확도는 68.75%였고 분류 정확도는 65.04%이었다. 우수고객, 비우수고객 각각의 분류 정확도를 살펴보면 비우수고객의 경우 80% 이상의 적중률을 나타내었으나 우수고객은 30%대로 상당히 낮았다. 따라서 우수 고객 분류의 목적으로는 이 모형을 적용할 수 없다. 우수고객 분류 정확도가 낮은 이유는 입력값의 속성이 턱없이 부족하여 출력값에 미치는 영향이 극히 미약하기 때문이다.

인공신경망은 민감성 분석(Sensitivity Analysis)을 통해 신경망 결과에 영향을 미친 입력값에 대한 상대적 중요도(Relative Importance)을 알 수 있는데 분석결과 거주지역, 수입, 성별, 인터넷 경력, 연령대 등의 순으로 상대적 중요도가 높은 것으로 나타났다.

4.3 데이터마이닝 적용을 위한 제안

○사의 데이터를 가지고 데이터마이닝을 수행하는 데 따르는 어려움은 다음과 같다.

첫째, 다수의 텍스트 필드 사용으로 인해 데이터마이닝에 이용할 수 있는 데이터 필드가 감소한다는 것이다. 일반적으로 텍스트 필드는 데이터마이닝에 사용될 수 없으므로 제거해야만 한다.

또한 경매 물품을 사용자가 직접 입력하게 되어 있어 같은 종류의 물품이라도 이름이 다를 수 있다. 그 결과 수작업을 통해 각 물품을 일일이 분류하고 매핑(Mapping)하여야 하는 어려움 때문에 연관규칙 탐사 적용이 거의 불가능하다.

둘째, 고객 정보의 선택적 입력으로 인해 여러 가지 문제가 발생한다.

데이터마이닝에 도움에 될 만한 고객 정보의 대부분이 고객이 선택적으로 입력하게 되어 있어 값을 갖지 않는 데이터가 많다. 또한 입력된 데이터 또한 정확한 정보가 아닐 확률이 높다. 보다 효과적인 데이터마이닝을 위해서는 이벤트 등을 통해 고객이 선택적 데이터를 정확히 입력하도록 유도하여야 한다.

셋째, 데이터마이닝 수행을 위한 변수 또는 속성의 부족이다.

데이터마이닝 방법에는 크게 지도방법(Supervised/Directed Method)과 자율방법(Unsupervised/Undirected Method)이 있다. 분류 작업을 수행하는 의사결정나무, 인공신경망 등이 전자에 속하고 세분화 작업을 수행하는 Apriori 알고리즘 같은 연관규칙과 k-평균, 코호넨 망 등의 군집화가 후자에 속한다.

특히, 지도방법은 목표변수가 반드시 있어야 하며 수 많은 입력 변수를 필요로 한다. ○사의 데이터에는 목표변수로 사용될 만한 변수가 없고 입력값으로 사용할 변수 또한 부족하다.

다른 변수에서 우수 고객임을 판별할 수 있는 변수를 입찰 회수, 낙찰 회수, 낙찰 금

액 등에서 유도하여 우수 고객을 판별하는 방법도 한계를 지닌다.

이 절에서는 보다 양질의 데이터마이닝 수행을 위한 방안을 제안하려 한다.

4.2.1 데이터마이닝 수행 목표 및 데이터 수집 계획 설정

대부분의 기업에서 데이터마이닝을 단지 하나의 기술로 인식하고 도구의 선택, 기술적 지원에만 과다하게 치중함으로 성공적인 결과를 내지 못하고 있다.

데이터마이닝이 대용량의 데이터와 덜 정제된 데이터 분석에 더 유연한 분석 기법을 제공하는 것은 사실이지만 그렇다고 충실도가 낮은 데이터에도 좋은 결과를 도출하는 마법 상자는 아니다.

과거에 목적 없이 수집된 데이터를 기반으로 데이터마이닝을 시도할 때 발생하는 문제는 다음과 같다.

첫째, 데이터의 양적 문제이다. 양질의 데이터마이닝을 수행하기 위해서는 대용량의 데이터와 많은 변수를 필요로 한다. 계획 없이 수집된 마이닝 목적에 맞는 변수가 부족할 수 있다.

외국 사례의 경우, Mozer *et al*[2000]은 이동전화의 이탈 고객을 예측하기 위해 134개의 변수를 사용했으며 Viveros *et al*[1996]은 의료보험 사기를 검출하기 위해 550GB 달하는 데이터베이스와 120개의 변수를 사용하였다[Mozer *et al* 2000, Viveros *et al* 1996].

한국의 경우 최종후외 2인[1999]은 12개의 변수를 갖는 15,000개의 데이터를 사용하여 보험회사 이탈 고객을 예측하였다[최종후외 2인 1999]. 모형의 전체 분류 정확도는 78% 높았으나 이탈 고객의 분류 정확도는 30% 대로 상당히 낮았다. 따라서 최종후 등이 구축한 모형은 이탈 고객 예측 목적에는 부적합하다.

김영만[1998]도 통신서비스 이탈 고객을 예측한 분석을 수행하였다[김영만 1998]. 43개의 변수와 3,000개의 데이터를 사용하여 70%의 전체 분류 정확도를 얻었으나 이탈 고객의 분류 정확도는 41%로 낮게 나타났다.

둘째, 데이터의 질적 문제이다. 웹상에서 입력된 데이터는 결손값을 가지는 경우가 많고 그 결과 복잡한 데이터 전처리 과정이 요구되며 많은 시간을 소비하여 마이닝하여도 좋은 결과를 얻을 수 없다.

데이터의 수가 적고 상대적으로 결손값이 많을 때엔 결손값 제거는 데이터 분석을 복잡하게 하고 정확성을 저하시킨다. 또한 결손값을 중간값 등으로 대체하는 경우 그 값으로 인해 부정확하고 불일치한 결과를 초래할 수 있다[Famili *et al* 1997, Wright 1998]. 결손값을 무시하는 경우 모든 결손값이 동일한 것으로 가정되어 군집화 등의 기법 수행 시 악영향을 주게 된다[Wright 1998].

또한 웹상에서 입력된 데이터는 결손값을 전부 제거할 수 있을 지라도 입력된 데이터에 대한 신뢰성을 측정할 수 없다.

셋째, 적용할 수 있는 데이터마이닝 기법이 한정된다. 수집된 데이터에 목표변수로 사용할 만한 변수가 없으면 분류 등의 태스크와 인공신경망 같은 지도방법을 적용할 수 없다. 그 결과 적용할 수 있는 데이터마이닝 기법이 한정되어 좋은 결과를 얻을 수 없다.

따라서 데이터마이닝 시스템을 도입하기 이전에 기업의 사업 목표를 제대로 이해하고 해당 목표를 달성하기 위해 필요한 성공 요소들을 조사하여야 한다. 데이터마이닝을

수행할 때 목표가 분명하지 않고 성공요소가 파악되지 못하면 필요한 데이터 선택이 어려우며 발견된 정보의 신뢰도도 저하된다.

데이터마이닝은 주어진 데이터에 대한 정보를 찾는데 마이닝 목표에 맞는 데이터, 변수 선정과 어떠한 경로를 통해서 수집할지에 대한 계획이 필요하다. 필요한 데이터 선택이 잘못되거나 데이터 수집에 문제가 있는 경우, 오류값, 특이값, 결손값 등으로 인해 의미 있는 결과를 얻을 수 없다.

또한 마이닝으로부터 발견된 지식에 대한 구체적이고 현실성 있는 적용 방안에 대한 계획이 필요하다.

4.2.2 시스템 구축

효과적인 데이터마이닝을 수행하기 위한 시스템은 그림과 같이 원천데이터, 데이터웨어하우스, 응용 프로그램 등으로 구성된다[그림3 참조].

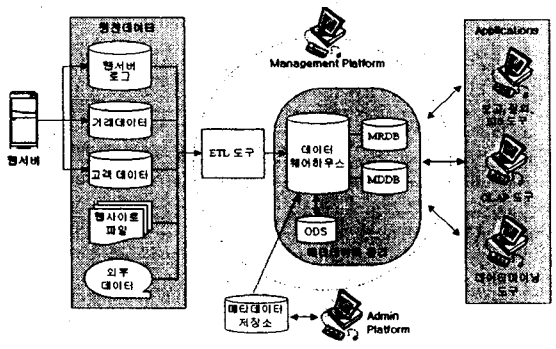

그림3: 데이터마이닝 시스템의 구조

데이터마이닝 수행을 위해 데이터웨어하우스가 반드시 필요한 것은 아니지만 데이터웨어하우스 도입을 통해 여러 소스에서 추출 통합되어 일차 걸러진 깨끗한 데이터를 얻을 수 있으므로 마이닝을 위한 데이터의 질과 일관성이 보장된다.

따라서 일단 데이터웨어하우스가 구축된 후 마이닝을 수행한다면, 데이터와 관련된 많은 문제들이 상당부분 해결될 수 있다. 즉, 데이터웨어하우스는 마이닝에 필요한 형태의 정제된 데이터를 가지고 있으며 마이닝을 위한 좋은 기반을 제공한다. 또한 데이터마이닝 작업에 초점을 맞춘 데이터마트를 별도로 구축할 수도 있다.

웹마이닝을 위해 웹 로그를 데이터베이스에 저장하면 데이터 전처리 과정을 보다 쉽게 할 수 있다. 특히 웹 사용 마이닝(Web Usage Mining)은 사용자 및 세션 구분이 가장 중요한데 엠베디드 세션 아이디(Embedded Session ID)를 이용하면 쉽게 사용자 세션을 구분할 수 있다[Mobasher *et al* 2000].

4.2.3 사이트 설계

웹사이트에서 사용자에 의해 입력된 데이터의 신뢰도가 낮기 때문에 군집화, 의사결정나무, 인공신경망 등을 적용한 데이터마이닝 작업을 수행하기는 어렵지만 보다 양질의 데이터를 수집할 필요는 있다.

사이트 설계를 통해 데이터마이닝에 적용할 수 있는 형태의 데이터를 수집할 수 있는데 예를 들어, 사용자가 마음대로 입력하는 텍스트 필드보다 선택 메뉴(Selection Menu), 라디오 버튼(Radio Button), 체크 박스(Check Box) 등을 이용하는 것이 좋다. 이러한 기술을 이용하면 설계자가 마이닝에 필요한 데이터의 범위를 지정할 수 있다.

또한 주민등록번호에서 성별과 연령을, 우편번호에서 지역을 자동으로 추출하여 데이터베이스에 저장하면 사용자가 입력하는 것보다 정확한 데이터를 수집할 수 있다.

인공신경망 등 지도방법을 이용하여 데이

터마이닝을 수행하기 위해서는 대용량의 데이터와 많은 수의 변수가 필요하고 의사결정나무는 데이터의 크기와 값에 지나치게 민감하여 서로 상이한 값은 갖는 양질의 입력 데이터를 필요로 한다.

웹 사이트에서 수집되는 데이터는 양적·질적으로 한계가 있으며 아직 우리나라의 상황에서는 외부 소스를 통해 데이터를 보완하는 것도 불가능하다. 따라서 웹 데이터를 이용하여 유용한 결과를 얻기 위해 적용할 수 있는 데이터마이닝 기법은 연관규칙 밖에 없는데 인터넷에서 발생하는 거래 데이터는 정확하고 데이터 수집 또한 용이하기 때문이다.

연관규칙 탐사를 위한 데이터를 수집하기 전에 [그림4]과 같이 분류(Taxonomy)를 정의할 필요가 있다.

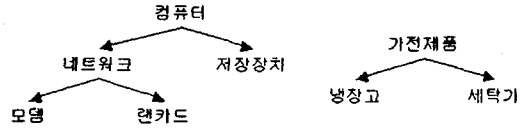

그림4: 분류의 예

한 사이트가 다루고 있는 물품이 수천 가지에 달하는 경우 지지도가 작아지게 되기 때문에 하위수준의 규칙은 최소지지도를 갖지 않을 수 있다. 또한 분류(Taxonomy)의 특성을 이용하면 그 분류에 의해 확장된 거래에서 일반화된 연관 규칙(Generalized Association Rules)을 탐사할 수도 있다 [Srikant & Agrawal 1995].

분류 없이 [그림5]와 같이 설계된 사이트는 사용자가 물품명을 직접 입력하는 경우 물품명이 통일되지 않고 연관규칙 발견을 위해서는 물품명을 수작업으로 일일이 매핑하여야 한다.

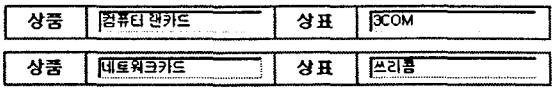

그림5: 텍스트 형식의 물품 등록

따라서 기업은 미리 정의된 분류표에 기반해 물품을 등록해야 하며 고객이 직접 물품을 등록할 경우 [그림6]과 같이 분류별로 입력이 가능하도록 설계한다.

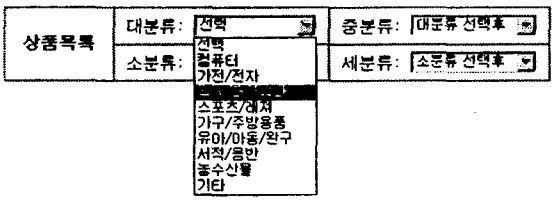

그림6: 분류를 이용한 물품 등록

4.2.4 데이터베이스 설계

연관규칙을 발견하기 위한 데이터베이스 구조는 [표2]와 같다. 여기서 TID는 고객 ID이다.

표2: 연관규칙 탐사를 위한 데이터베이스

TID	I_1	I_2	…	I_m
1	0	1	…	0
2	1	0	…	0
…	…	…	…	…
N	0	1	…	1

순차패턴을 발견하기 위한 데이터베이스는 시간적 요소가 추가되어 [표3]과 같다.

표3: 순차패턴 탐사를 위한 데이터베이스

TID	고객 ID	시간	I_1	I_2	…	I_m
1	1	0	0	1	…	0
2	1	1	1	0	…	1
3	1	3	0	1	…	0
4	2	0	0	1	…	1
5	2	2	1	1	…	0
…	…	…	…	…	…	…

시간 제약 요소는 인접 시퀀스 간의 최대/최소간격(Maximum/Minimum Gap), 하나의 항목집합에 속한 최대 거래 시간과 최소 거래 시간 간의 차이를 제한하는 이동 시간 윈도우(Sliding Time Windows) 등이 있는데 대부분의 상용 데이터마이닝 도구에서는 이동 시간 윈도우 기능만 제공한다[Srikant & Agrawal 1996].

[그림7]은 최대/최소 간격, 이동 윈도우를 나타낸 것이다. 시퀀스의 각 구성요소는 하나의 거래(Transaction)를 의미한다.

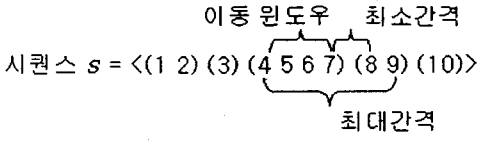

그림7: 최소/최대간격과 이동 윈도우

[표2, 표3]와 같은 데이터베이스 구조는 용량의 낭비를 초래하기 때문에 거래 데이터는 [표4]와 같은 구조로 저장한다.

표4: 저장된 거래 데이터베이스

고객	대분류	중분류	소분류	세분류	항목명	시간
1	1	4	2	2	모니터	00-MAY-12
1	1	4	6	3	마우스	00-MAY-12
8	3	2	4	1	장갑	00-MAY-12
…	…	…	…	…	…	…
2	1	3	2	1	모뎀	00-MAY-31

일부 상용 데이터마이닝 도구는 내부적으로 다양한 구조의 데이터를 연관 규칙 탐사에 적합한 구조로 변환시키는 기능을 제공하고 있다. 그러한 기능을 제공하지 않더라도 간단한 질의를 통해 연관규칙 탐사에 적합한 구조로 변환시킬 수 있다.

전체 거래의 수와 최소지지도를 고려한 후 대분류, 중분류, 소분류, 세분류 중 하나를 택하여 연관규칙·순차패턴을 발견하거나 또는 전체 분류를 선택하여 일반화된 연관규칙·순차패턴을 발견할 수 있다.

V 결론

인터넷 경매 사이트 데이터로 고객 세분화를 수행하고 군집의 특성을 해석해 본 결과 결손값을 가지는 데이터를 전부 제거했음에도 불구하고 데이터의 상당 부분이 부정확한 정보를 가지고 있는 것으로 나타났고 또한 이러한 부정확한 정보를 제거하기 위한 판별 기준이 없다는 것이다. 부정확하고 부족한 데이터로 인해 의사결정나무, 인공신경망을 통한 우수고객 분류 수행의 정확도가 상당히 낮은 것으로 나타났다.

전자상거래 초기 단계인 우리나라의 현실에서 온라인 상에서 입력되는 대부분의 데이터가 부정확한 정보를 가지고 있으며 이러한 데이터를 가지고 데이터마이닝을 수행한다는 것은 어려운 일이다. 또한 미국과 같이 대량의 데이터를 보유하고 있는 대형 데이터 벤더들도 없는 상황에서 다른 소스를 통해 데이터를 보완하는 것도 거의 불가능하다고 볼 수 있다.

Freechal.com, Wowcall.com 등 많은 사이트에서 보다 정확한 고객 데이터를 얻기 위해 이벤트를 실시하여 데이터를 제공하는 고객에게 보상하고 있으나 이렇게 입력된 데이터 역시 정확한 정보를 가지고 있다고 판단할 수 없다. 아직 전자상거래가 제도적으로 확립되지 않았고 인터넷 사용자의 전자상거래에 대한 의식 수준도 낮아 온라인 상에서 정확한 데이터를 기대하기는 힘들다. 일방적인 단순한 보상 외에 고객의 사생활 보호

에 대한 제도적 장치와 사용자들의 자발적인 참여를 유도할 수 있는 전략을 구사해야 보다 정확한 데이터를 수집할 수 있을 것이다.

또한 더 좋은 데이터마이닝 작업을 위해서는 데이터마이닝 태스크와 이를 위해 필요한 데이터를 정의한 후 사이트 및 데이터베이스를 재설계할 필요가 있다. 인구통계학적인 데이터에 의존하기 보다는 보다 정확하고 유효한 값을 갖는 데이터와 데이터마이닝 작업에 관련된 데이터를 수집해야 할 것이다.

참고문헌

김영만, "통신서비스 시장에서 데이터마이닝을 이용한 이탈고객 분석", 한국과학기술원 석사논문, 1998.

장남식 · 홍성완 · 장재호, 데이터마이닝, 대청, 1999.

조재희 · 박성진, OLAP 테크놀로지, 시그마컨설팅, 1999

최종후 · 한상태 · 김은석 · 강현철, "클레멘타인을 이용한 보험회사 이탈고객 관리분석", SPSS 사용자 사례 발표회, 1999.

Agrawal, R., T. Imielinsk, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington D.C., pp. 207-216, May, 1993.

Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules", In *Proceedings of the VLDB Conference*, Santiago, Chile, September, 1994.

Agrawal, R. and R. Srikant, "Mining Sequential Patterns", In *Proceedings of the 20th International Conference on Data Engineering*, Taipei, Taiwan, March, 1995.

Anand, S., D. Bell, and J. Hughes, "The Role of Domain Knowledge in Data Mining", *CIKM 95*, 1995.

Berry, J. and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.

Bigus, J., *Data Mining with Neural Network: Solving Business Problems from Application Development to Decision Support*, McGraw-Hill, 1996.

Famili, A.W Shen, R. Weber, and E. Simoudis, "Data Preprocessing and Intelligent Data Analysis", *Intelligent Data Analysis* 1, January, 1997.

Fayyad, U., G. Piatetsky-Shapito, P. Smyth, and R. Uthurusamy, "From Data Mining to Knowledge Discovery: An Overview", *Advance in Knowledge Discovery and Data Mining*, AAAI/MIT Press, CA, 1996.

Feng, L., H. Lu, and J. X. Yu, J. Han, "Mining Inter-Transaction Association Rules with Templates", *CIKM'99*, 1999.

Han, J., "Data Mining", Encyclopedia of Distributed Computing, Kluwer Academic Publishers, 1999.

Lu, H., J. Han, and L. Feng, "Stock Movement and N-Dimensional Inter-Transaction Association Rules", In *Proceedings of 1998 SIGMOD '96 Workshop on Research Issues on Data Mining and Knowledge Discovery*, Seattle, Washington, pp. 12:1-12:7, June, 1998.

Mobasher, B., R. Cooley, and J. Srivastava

"Automatic Personalization Based on Web Usage Mining", *Communications of the ACM*, Vol. 43, No. 8, August, 2000.

Mozer, M., R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky, "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry", *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, May, 2000.

Srikant, R. and R. Agrawal, "Mining Generalized Association Rules", In *Proceedings of the 21st International Conference on VLDB*, Zurich, Switzerland, September, 1995.

Srikant, R. and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", In *Proceedings of the 5th International Conference on Extending Database Technology*, Avignon, France, March, 1996.

Viveros, M., J. Nearhos, and M. Rothman, "Applying Data Mining Techniques to a Health Insurance Information System", In *Proceedings of the 22nd VLDB Conference*, Mumbay, India, 1996.

Westphal, C. and T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, John Wiley & Sons, 1998.

Wright, P., "Knowledge Discovery Preprocessing: Determining Record Usability", In *Proceedings of the 36th Annual Conference on Southeast Regional Conference*, 1998.