碩士學位論文

데이터마이닝 시스템의 커스터마이징 방법론에 관한 연구

濟州大學校 經營大學院

經營情報學科 經營情報學 專攻

吳 京 訓

碩士學位論文

데이터마이닝 시스템의 커스터마이징 방법론에 관한 연구

指導敎授　金 根 亨

濟州大學校 經營大學院

經營情報學科 經營情報學 專攻

吳　京　訓

2004 年　8 月

데이터마이닝 시스템의

커스터마이징 방법론에 관한 연구

指導敎授 金 根 亨

이 論文을 經營學 碩士學位 論文으로 提出함.

2004 年 8 月

濟州大學校 經營大學院

經營情報學科 經營情報學 專攻

吳 京 訓

吳京訓의 經營學碩士學位論文으로 認准함.

2004 年 8 月

심 사 위 원 장 _____ 印

심 사 위 원 _____ 印

심 사 위 원 _____ 印

<목 차>

<그림 차례>

Ⅰ. 서론

1. 연구배경 및 목적

오늘날 정보기술의 눈부신 발전은 업무의 자동화를 촉진시켜 엄청난 양의 데이터를 전자적으로 수집하고 보관하는 기능을 가능하게 했다. 데이터 수집과 저장 기술의 발달, 데이터베이스 관리 시스템과 데이터웨어하우스 기술의 광범위한 사용은 기업내부에 대량의 데이터를 축적할 수 있도록 하였으며, 기업들도 축적된 데이터를 의사결정에 필요한 새롭고 가치있는 정보와 지식으로 변환 하고 활용할 수 있는 잠재적인 원천으로 인정하고 있다[김종훈,2000]. 실제로 기업들이 얻고자 하는 정보의 원천은 축적된 데이터 자체에 존재하는 것이 아니라, 축적된 데이터에서 효과적으로 찾아낸 정보에 있으며, 이러한 정보 추출 방법론으로써 데이터마이닝(data mining)이 각광받게 되었다.

데이터마이닝은 기업경영 활동의 결과로서 수집된 대량의 데이터를 분석하여 기업의 이윤 추구에 도움이 될 수 있는 정보와 지식을 획득할 수 있는 기술이다[장남식, 1999]. 데이터마이닝은 이러한 비즈니스적 요구로 인해 등장한 정보 추출 방법론으로써 통신, 은행, 소매, 의료, 제조, 유통, 항공 등 다양한 산업 분야에서 널리 사용 되고 있다.

데이터마이닝의 기술에는 연관규칙 탐사(association rule mining), 항목분류(classification), 클러스터링(clustering), 요약(summarization), 순차패턴 탐사(sequential pattern) 등이 있다[김정자, 1998].

더욱이, 데이터웨어하우스가 구축되어 있다면 데이터마이닝 기술을 적용하기 위한 최적의 환경이 될 수 있다.

이러한 기술들은 다양한 응용 영역들에서 활용될 수 있고 데이터마이닝의 사용자들에 따라 필요한 응용영역 및 데이터마이닝 기술이 다양하다. 즉, 사용자들에 따라 데이터마이닝에 대한 요구사항은 다르게 나타날 수 있다.

상용화된 데이터마이닝 제품들은 대부분 완성품의 형태로 데이터마이닝 기능들의 일부 또는 전부를 지원한다. 완성품은 구입 즉시 사용할 수 있다는 장점은 있으나 사용자의 요구사항이나 기업의 환경에 유연하게 적응할 수 없는 문제가 있다. 즉, 데이터마이닝의 기능들 중 일부만을 제공하는 완성품의 경우 다양한 응용영역의 다양한 사용자들을 만족시킬 수 없게 되고, 데이터마이닝의 모든 기능을 제공하는 완성품의 경우는 다양한 사용자들에 의하여 사용되어 질 수 있지만 제품가격 대 효용 측면에서 비효율성이 생기게 된다. 따라서 데이터마이닝 기능을 제공하는 패키지를 커스터마이징 가능한 형태로 제공하고 사용자들의 요구사항에 적합하게 커스터마이징 하여 사용할 수 있으면 비용 대 효용 측면에서 효율화를 기할 수 있을 것이다[Hwang,1991].

본 논문에서는 데이터마이닝 시스템을 완성품이 아닌 커스터마이징의 형태로 도입 및 구축하기 위한 방법론을 고찰해보고, 실제 오라클 제품을 이용하여 커스터마이징을 하는 사례를 살펴본다. 그리고 구축된 데이터마이닝 시스템을 이용하여 판매 데이터를 분석한 사례를 제시하였다.

2. 연구 내용 및 논문의 구성

본 논문에서는 데이터마이닝 시스템을 완성품이 아닌 커스터마이징의 형태로 도입 및 구축하기 위한 방법론을 고찰해보고 성공적인 커스터마이징 전략을 제안한다. 실제 오라클 제품을 이용하여 커스터마이징을 하는 사례를 살펴보고 구축된 데이터마이닝 시스템으로 판매 데이터를 이용한 연관규칙 분석을 제시하였다.

본 논문의 구성은 다음과 같다.

1장에서는 서론 부분으로 연구배경 및 목적을 제시하였다.

2장에서는 데이터마이닝 개념과 기법 및 응용과 데이터웨어하우스에 대해서 정리하였다.

3장에서는 커스터마이징 관련 선행연구와 함께 데이터마이닝 시스템의 커스터마이징 필요성을 고찰하였으며 데이터마이닝 시스템의 커스터마이징 방법론을 제안하였다.

4장에서는 제안한 커스터마이징 방법론의 타당성 검토 차원에서, 제안한 커스터마이징 방법론을 이용하여 실제 오라클 데이터마이닝 시스템을 커스터마이징 하는 과정을 소개하였다.

5장에서는 오라클 데이터마이닝 시스템을 이용한 판매 데이터의 분석하였다.

6장에서는 결론과 향후 연구 방향을 제시하였다.

Ⅱ. 데이터마이닝

1. 데이터마이닝의 정의

데이터마이닝이란 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 90년대 초반부터 지식발견, 정보발견, 정보수확 등의 이름으로도 소개되어 왔는데 일반적으로 “대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업”이라 정의한다[장남식, 1999].

데이터마이닝은 ‘데이터베이스 내에서의 지식발견(Knowledge Discovery in Database, KDD)′과 유사어로 언급[pieter, 1996]되기도 하지만, 지식발견은 데이터로부터 유용한 정보를 발견하는 프로세스의 전 과정이며, 데이터마이닝은 지식발견 중에서도 데이터로부터 정보를 추출하기 위해서 기법을 적용하는 특정단계라고 정의할 수 있다[장남식, 1999].

KDD의 전체적인 프로세스는 연구자들마다 다소 표현에 차이는 있으나 이를 종합하여 보면, 1) 비즈니스 환경과 문제를 바탕으로 분석 목적 설정, 2)가용한 자료 확인, 3)자료의 탐색적 분석, 4) 데이터마이닝을 통한 지식추출, 5) 결과에 대한 평가와 해석, 6)전략결정 및 실행, 7)실행결과 모니터링의 7단계로 나누어 볼 수 있다 [Adrianns & Zantinge, 1996; Pyo, Uyal. & Chang, 2002].

이러한 KDD 과정에서 데이터마이닝을 통한 지식 추출과정이 차지하는 역할이 가장 결정적인 단계이기 때문에 결국 어떠한 비즈니스의 목적을 달성하기 위해 어떠한 데이터마이닝 기법을 적절히 적용하느냐가 중요한 문제가 되며, 이러한 이유로 KDD와 데이터마이닝을 유사어처럼 이용하는 것이다.

2. 데이터마이닝 기법

데이터마이닝 기법에는 연관규칙 탐사(association rule mining), 연속규칙 탐사(sequence rule mining), 분류(classification), 군집화(clusteing) 등이 있다. 이들 데이터마이닝 기법은 특정 업종에만 국한된 것이 아니라 비즈니스의 환경과 목표, 사용 가능한 데이터의 속성에 따라 적용될 수 있는 적합한 기법들이 달라지게 된다[Berry&Linoff, 1997]. 최근에는 유전자 서열 데이터베이스 등과 같은 생물학 분야에서도 응용되고 새로운 생물학적 지식의 창출에 이용되고 있다[김양석, 2000].

각 기법을 간략히 설명하면 다음과 같다.

1) 연관규칙(association rule)

많은 데이터마이닝 기법중 가장 활발히 연구가 이루어지는 분야로서 데이터안에 존재하는 항목간의 종속관계를 찾아내는 작업으로, 마케팅에서는 손님의 장바구니에 들어 있는 품목간의 관계를 알아본다는 의미에서 장바구니 분석이라고 한다[박종수, 1998].

연관규칙의 표현은 데이터베이스의 데이터 항목 집합 I에 대한 부분 집합 X,Y ($X \subset I$, $Y \subset I$, $X \cap Y \neq \emptyset$)에 대한 연관규칙이 존재할 때 $X \rightarrow Y(c\%)$로 표현하며, 그 규칙은 신뢰도가 c%임을 의미한다. X에 대한 지지도는 전체 트랜잭션에서 X를 접근하는 트랜잭션들의 비율로써 sup(X)로 표현하며 규칙 $X \rightarrow Y$의 신뢰도는 X를 포함하는 트랜잭션 중에서 X와 Y를 동시에 포함하는 트랜잭션의 비율로 conf($X \rightarrow Y$)로 표현한다. 사용자는 규칙 탐사 시 지지도와 신뢰도에 대하여 탐사기준의 임계값으로 최소 지지도(minsup)와 최소 신뢰도(minconf)를 지정하며, 규칙의 신뢰도와 지지도는 이 최소 지지도와 최소 신뢰도를 만족해야 채택된다.

즉, 규칙 R이 sup(R) ≥ minsup 이고 conf(R) ≥ minconf의 조건을 만족할 때 연관규칙으로서 의미가 있게 된다. 지지도를 만족하는 데이터 항목들은 빈발하다(large, frequent)라고 하며 빈발항목들에 의해 구성된 규칙들의 신뢰도를 검사하여 최종적인 연관규칙을 탐사한다.

연관규칙 탐사 알고리즘의 기본적인 골격은 다음 2단계로 구성된다.

1단계 : 사용자가 정의한 최소 지지도를 만족하는 데이터 항목의 집합을 탐사하는 단계이다. 이 단계에서는 각각의 데이터 항목에 대한 지지도를 계산하여 미리 정의된 최소 지지도를 만족하는 데이터 항목들을 추출해낸다.

2단계 : 연관규칙을 생성하는 단계이다. 이 단계에서는 1단계에서 탐사된 데이터 집합을 이용하며, 데이터의 부분 집합에서 생성된 규칙 중 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 최종 규칙으로 탐사한다[하단심, 2001].

2) 연속규칙(sequence rule)

연속규칙은 연관규칙에 시간관련 정보가 포함된 형태로서, 시간상에 순차적으로 나타나는 사건이나 거래의 종속관계를 찾아내는 작업이고 순차패턴 발견은 순서대로 일어난 데이터를 분석해 빈도수가 높은 순차패턴을 찾아내는 기술을 말한다[Rakesh, 1995].

응용 분야를 살펴보면 홈쇼핑 회사에서 소비자가 구매한 물건을 보고 다음에 살 것으로 예상되는 물건들의 쿠폰이나 카탈로그를 발송하는데 사용할 수 있고, 학습지 회사에서는 국어학습지를 구독하는 학생들이 그 다음에 어떤 다른 과목을 주로 더 구독하는지 알아내 판매를 촉진하는데 사용할 수도 있다. 우편 주문이나 전자상거래 사이트에서 고객이 미래에 구매할 물건을 예측하는 데 사용할 수 있고 웹 페이지 방문자들의 액세스 로그를 분석해 웹 페이지를 고객에 따라 다른 구조

를 갖게 하는데 사용할 수 도 있다. 또 병원에서 진료 받은 환자들의 진료기록을 보고 과거의 어떤 증상이나 치료 과정(또는 결과) 이 지금 현재 걸린 병을 유발하는 원인이었는지 찾아내는 데 이용할 수 있다[MinosN, 1999].

3) 분류(classification)

데이터마이닝에서 가장 많이 사용되는 작업으로, 부류값이 포함된 과거의 데이터로부터 부류별 특성을 찾아내어 분류모형수립, 이를 토대로 새로운 레코드의 부류값을 예측하는 것이다.

예를 들어 어떤 동네의 대형슈퍼마켓에서 생성된 구매 데이터를 보면 그 동네의 아이와 젊은 여성, 노인의 비율을 대략적으로 알 수 있다. 이러한 데이터를 이용해 과거 다른 동네의 데이터와 비교해 성공했는지 실패했는지를 나타내는 클래스를 사용하면 새로운 동네에 새 슈퍼마켓을 만들려고 할 때 성공 가능성을 예측하는데 사용할 수 있다[Rajeev, 1998]. 또 새로운 의약품을 개발했을 때 여러 부류의 사람들, 즉 연령, 인종, 성별, 체중, 키 등이 서로 다른 사람들에게 임상 실험을 한 후, 그 약품이 효능이 좋았는지, 부작용이 있었는지를 나타내는 정보를 클래스로 만들어 입력한 후 각각의 클래스의 특징을 결정 트리 기법을 사용해 만들어 의사가 약을 처방할 때 주의 깊게 사용하도록 감독할 수 있다. 또 전자상거래 사이트에서 고객들의 구매 데이터를 보고 어떤 특징이 있는 고객이 비싼 수입 명품을 구입하는지 예측하는 데도 사용할 수 있다[Jone Shafer, 1996].

4) 데이터 군집화(clustering)

레코드를 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업으로서 분류작업과 흡사하나, 분석하고자 하는 데이터에 부류가 포함되어 있지 않다는 점에 차이가 있다. 이 기법은 어떤 목적 변수를 예측하기 보다는 고객 수입, 고객연령과 같은 속성이 비슷한 고객들을 묶어서 몇 개의 의미 있는 군집으로 나타내는 것을 목적으로 한다. 숲이 너무 복잡해서 전체를 파악할 수 없을 때 나무들을 살펴보아

야 하듯이, 대용량이 데이터가 너무 복잡할 때는 이를 구성하고 있는 몇 개의 군집을 우선 살펴봄으로써 전체에 대한 윤곽을 잡을 수 있을 것이다.

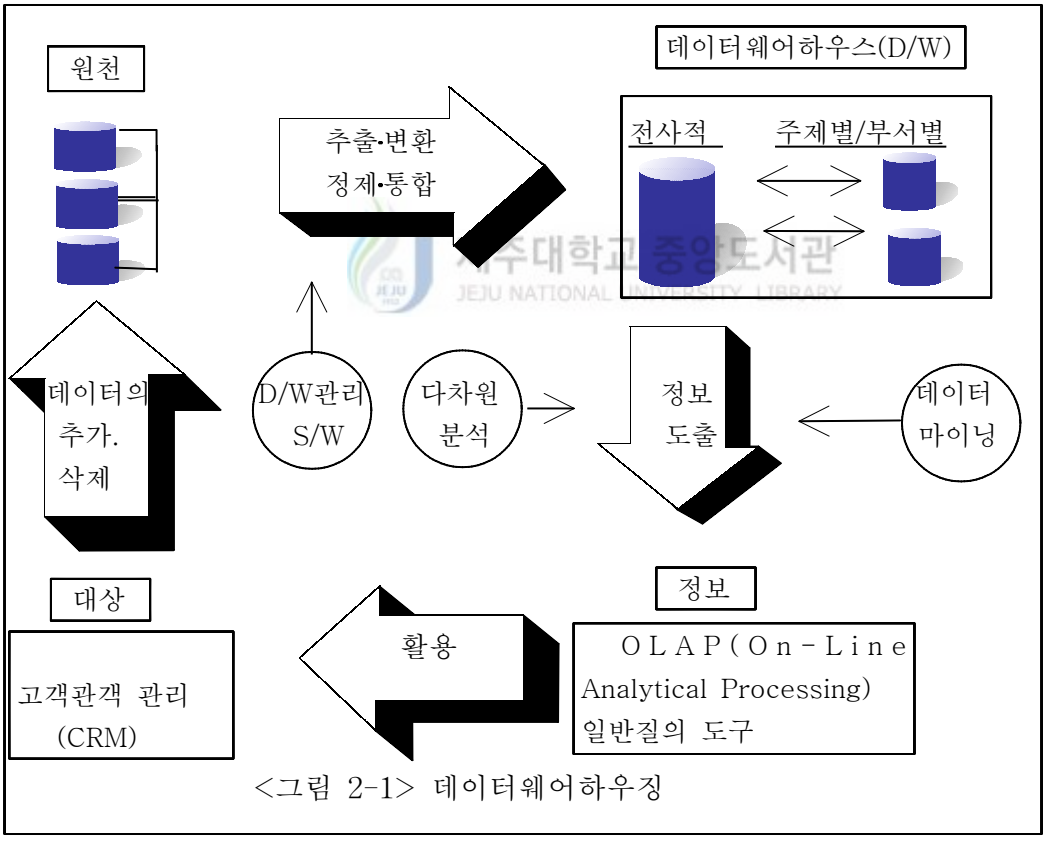
유사한 특성을 갖는 클래스를 함께 그룹화하고 분할하는 방법으로, 어떤 그룹이 사전에 정의되어 있지 않다는 점에서 분류방법과 차이가 있다.

주어진 여러 고객의 구매 데이터를 바탕으로 그 구매 상품의 특징에 따라 고객을 여러 그룹으로 나누는 것이라 할 수 있다[Sudipto Guha, 1999]. 또 모든 고객의 신상정보를 이용해 그 유사성에 따라 그룹을 나누는 데 사용할 수 있다. 인터넷 검색엔진 회사에서는 웹 페이지의 내용에 따라 그룹을 만드는 데 사용할 수 있다.

군집화에는 여러 가지 알고리즘이 개발됐는데 알고리즘에 따라 다른 군집화를 만들어 낸다. 그러므로 모든 알고리즘의 특성을 잘 알고 있어야 자기 응용 분야에 맞는 것을 잘 사용할 수 있다. 숫자 형태의 데이터인가 범주 형태의 데이터인가에 따라 다른 형태의 알고리즘이 존재한다[Tan Zhang, 1996].

3. 데이터마이닝과 데이터웨어하우스

‘데이터웨어하우스’란 용어는 단순히 데이터가 보관되어 있는 거대한 저장고를 의미한다. 이에 반해 ‘데이터웨어하우징(data warehousing)’ 이란 데이터의 수집 및 처리에서 도출되는 정보의 활용에 이르는 일련의 프로세스라고 정의할 수 있다. 데이터웨어하우징은 개략적으로 <그림 2-1>에서와 같이 데이터 추출·변환·정제·통합, 정보도출, 활용, 데이터의 추가·갱신·삭제의 4단계로 구성된다. 데이터마이닝은 데이터웨어하우징의 정보도출 단계에서 적용될 수 있는 하나의 기법이라고 할 수 있다.



<그림 2-1> 데이터웨어하우징

・데이터의 추출・변환・정제・통합 단계

업무에 따라 다양한 운영시스템별로 산재해 있는 데이터나 외부 원천으로부터 데이터를 추출하고, 사용자의 요구 상에 맞게 변환하여 통합하며, 필요에 따라 정제하는 과정으로 데이터웨어하우징 프로세스 중에서 가장 힘들고 많은 인원과 시간을 필요로 한다. 작업에 필요한 도구는 3GL(3rd Generation Language)와 같은 프로그래밍 언어를 이용하여 제작할 수 있으나 작업의 효율성을 고려하면 상용화된 관리 소프트웨어를 이용하는 것이 바람직하다.

데이터웨어하우스를 구축하는 데에는 '하향식(top-down)'과'상향식(bottom-up)' 방식이 있다. '하향식' 방식은 전사적 차원의 데이터웨어하우스를 먼저 구축한 후 필요에 따라 하나의 주제 또는 부서 중심의 데이터마트(data mart)를 구축하는 방법이고 '상향식' 방식은 반대로 데이터마트를 먼저 구축한 후 점차적으로 이들 데이터마트를 통합하고 확장하는 방법이다. 하향식 방식을 통한 데이터웨어하우스 구축은 전사 차원의 일관된 정보를 제공한다는 장점을 지닌 반면, 장기간에 걸친 대규모 인원 투입이 필요하기 때문에 시간 및 비용 측면에서 경영자에게 많은 부담이 될 수 있다. 상향식 방식은 이러한 문제를 줄여준다는 장점을 가지고 있지만 데이터 모델, 아키텍처, 전사적 데이터웨어하우스 등을 포함한 전체 시스템 구성도의 청사진이 준비되지 않은 상황에서 진행될 경우에는 향후 데이터의 일관성을 유지하며 데이터마트들을 통합하거나 확장하는데 상당한 어려움이 따른다.

・정보도출 단계

일반질의나 보고서 작성도구를 이용하여 필요한 정보를 조회하고 정리하는 과정이다. 그러나 보다 심도 있는 정보를 찾아내기 위해 OLAP 도구를 이용하는 것이 보편화되어 있으며, 새로운 정보의 필요성에 대한 요구가 증대함에 따라 데이터마이닝을 적용하는 사례가 점차 늘어나고 있다.

· 정보활용 단계

추출된 정보를 실제 현장에서 활용하는 과정이다. 예를 들어 고객 관련 정보는 고객 직접대면, 콜센터, 통신이나 인터넷, DM(Direct Mail) 등의 고객접점에서 다양한 고객관계관리(CRM: Customer Relationship Management)를 위해서, 제품 관련 정보는 새로운 제품의 설계나 불량요인 분석 등에 활용할 수 있다.

· 데이터의 추가 · 갱신 · 삭제 단계

운영시스템 상의 데이터를 변경하는 과정으로 활용대상을 상대로 정보를 이용하여 조직의 업무를 수행함에 따라 부가적으로 진행되곤 한다. 데이터웨어하우스는 궁극적으로 기존의 운영시스템의 데이터나 외부 데이터를 이용하여 만들어지기 때문에 원천 데이터의 품질은 데이터웨어하우스 데이터의 품질에 직접적인 영향을 준다. 따라서 이 단계는 원천 데이터의 고품질을 유지하는데 기여하는 바가 크다.

효과적인 데이터마이닝 작업을 위해 데이터웨어하우스가 지원하는 이점은 다음과 같다.

· 양질의 데이터 제공

데이터마이닝을 통해 발견되는 정보의 신뢰도는 입력되는 데이터의 품질에 의존한다. 의사결정을 지원하는 목적으로 구축된 데이터웨어하우스는 이미 다양한 데이터 원천으로부터 통합되어 변환 및 정제의 과정을 거쳤기 때문에 데이터의 양과 질적인 면에서 신뢰성이 확보되어 있다. 따라서 데이터웨어하우스는 데이터마이닝을 위한 최적의 시스템이다.

· 시간 및 노력 절감

데이터웨어하우스 환경 하에서 데이터마이닝을 위한 지식발견 프로세스를 수행할 경우 프로세스 중에서 가장 많은 시간과 노력이 요구되는 데이터의 선택, 정제, 보완, 변환 단계에서의 작업을 상당히 줄일 수 있다. 따라서 빠른 시간 내에 최소한의 노력으로 필요한 정보를 획득할 수 있다.

· 다양한 형태의 데이터 제공

데이터웨어하우스에서 제공하는 상세 데이터와 요약 데이터는 동일한 내용을 담고 있으면서도 시간과 요약수준에 따른 다양한 분석을 가능하게 한다. 따라서 데이터마이닝을 통해 산출되는 정보의 깊이와 폭 또한 다양해진다.

· 데이터 지도(map) 제공

데이터웨어하우스의 메타데이터는 데이터마이닝을 통해 찾고자 하는 정보의 원천이 되는 데이터에 대한 위치와 구조를 알려주는 이정표 역할을 한다. 따라서 데이터마이닝 분석가나 사용자가 필요한 데이터의 종류나 소재를 용이하게 파악할 수 있다.

4. 데이터마이닝 응용분야

다양한 산업 분야에 속한 많은 기업들은 그들이 구축한 세부적인 거래정보 데이터베이스에 데이터마이닝 기법을 적용하여 얻은 유용한 정보를 활용함으로써 경쟁우위를 확보하고 있다. 데이터마이닝은 앞으로 자사의 경쟁력 향상에 상당한 역할을 할 것으로 기대하고 있다.

데이터마이닝은 특정한 업종에만 국한된 것이 아닌 모든 분야에 적용할 수 있는 수평적 응용기술이다. 단지 아직까지는 데이터마이닝의 성공사례가 주로 데이터의 충실도가 상대적으로 높은 은행, 카드, 보험들의 금융분야와 통신분야에서 주로 발표되고 있다.

1) 은행업

은행에서 데이터마이닝은 사기 행위 색출(fraud detection), 고객집단 분류(customer segmentation), 라이프 싸이클에 따른 고객가치 관리(predictive life cycle management) 등 다양한 분야에 데이터마이닝을 이용할 수 있다.

은행이 신용카드 사업에서 신용카드 사기로 인한 피해는 매우 커서 데이터마이닝 기법을 이용하여 과거에 사기행위에 사용된 신용카드거래를 분석함으로써 사기행위의 패턴을 찾아낼 수 있다. 신용카드 사기 행위의 전형적인 사례는 전자상거래에서 짧은 기간에 많은 거래가 일어나는 경우이며, 이러한 경우를 사기행위의 가능성을 알려주는 경고 신호로 인식함으로써 피해를 줄일 수 있다. 또한 은행은 어떤 고객의 구매행위가 사기행위 패턴과 비슷할 경우 그 거래를 승인하지 않도록 시스템을 구성하는데 이용할 수 있다.

고객집단 분류는 특정 고객집단을 찾아내고 이 집단만을 겨냥한 차별화된 서비스를 제공할 수 있다.

예를 들어 어떤 상품은 자주 여행을 다니는 고객집단에게, 어떤 상품은 언제나 결제일을 잘 지키는 고객들에게 중점적으로 판매할 수 있다. 또한 은행은 고객집단 편성에 관한 지식을 이용하여 특정 판촉활동을 위하여 가장 많은 효과와 혜택을 얻게 될 지점을 찾는 데에도 사용할 수 있다.

라이프 싸이클 예측 관리(predictive life-cycle management)에 의한 지식발견은 은행이 고객의 시간에 따른 가치(lifetime value)를 예측하고 이에 따라 개개인의 고객집단에 알맞은 서비스를 제공하는데 도움을 준다. 은행은 수익성이 높은 고객집단을 정의하고 지식발견 기법을 이용하여 이들의 공통된 특성을 발굴한다. 이러한 특성을 지닌 현재의 고객들을 찾아낼 수 있는데 이들은 가까운 장래에 수익성이 높은 고객이 될 가능성이 매우 높다. 은행은 이들에게 특별한 상품거래를 제안하거나 수수료를 면제해 주는 것과 같은 고객 이탈방지 프로그램 같은 것을 실시할 수 있다.

2) 통신산업

전 세계적으로 점점 치열해져가는 경쟁에 직면하고 있는 통신회사들은 기존 고객을 유지하고 새로운 고객을 끌어들이기 위해 적극적인 마케팅 정책과 가격 정책을 실시하고 있다. 이러한 통신 산업 분야에 통화기록 분석, 고객충성도 등 지식발견에 적용되고 있다.

통신사업자들은 고객의 자세한 통화기록을 가지고 있다. 통화기록 분석을 통하여 비슷한 통화 사용패턴을 가진 집단을 찾아내어 그들에게 유리한 가격정책이나 기능들을 개발할 수 있다.

어떤 고객은 계속 통신서비스 제공자를 바꾸면서 각 통신회사가 제공하는 인센티브를 이용한다. 통신회사는 지식발견 기술을 이용하여 한 번 고객이 되면 오랫동안 지속적인 거래를 하게 될 고객과 그들의 특성을 찾아내고 이들을 중심으로 가장 이익이 많은 곳에 마케팅 투자를 할 수 있다.

3) 보험업

보험회사는 오랜 기간에 걸쳐 집적된 방대한 데이터를 가지고 있는데 이것은 효과적인 계획을 세우는 지렛대로 활용될 수 있다. 사기색출, 상품설계, 위험분석 등 보험 산업 분야에 지식발견 기술이 다양하게 이용되고 있다.

예를 들면, 허리부상과 같은 높은 보험 청구율을 가진 분야의 청구자, 의사, 변호사들 사이의 관련성 또는 보험청구 패턴을 찾아냄으로써 보험사기를 줄일 수 있다. 또한, 보험업자는 가장 수익성이 좋은 상품 구성 즉, 보험가입 신청자의 특성, 보험증권의 보장범위, 및 보험증권의 특약의 최적의 결합을 알고 싶어 한다. 보험업자들은 이 정보를 새로운 상품을 설계하고 장래의 판매를 위하여 기존의 상품을 고부가가치화 하는데 이용한다. 보험업자는 보험지급액과 관련된 여러 요인들을 찾아냄으로써 지급부담 위험을 줄일 수 있다. 예를 들어 미국의 대형 보험회사는 최근 지난 2년간의 중요한 보험청구건을 검토한 결과 기혼자의 청구금액이 미혼자의 청구금액의 두 배에 달한다는 사실을 발견하였다. 이 지식을 바탕으로 이 회사는 기혼자에게 일률적으로 적용, 할인하여 주는 정책을 조정하였다.

4) 유통업

유통업자들은 자사가 발행한 신용카드와 컴퓨터화 된 결제시스템을 통하여 고객들의 매일 매일의 자세한 구매정보를 보유할 수 있게 되었다. 이러한 정보는 유통업자들로 하여금 여러 다른 성격의 고객 집단을 보다 잘 이해하는데 도움을 주고 있다. 바구니 분석, 시계열 패턴 조사, 예측모델의 개발 등은 유통업에 대표적인 지식발견 분야 이다.

바구니 분석은 일명 친화성 분석이라고도 하는데 고객들의 구매 행위시 어떤 상품들이 같이 구매되는가를 밝혀낸다. 이와 같은 지식은 상점의 진열 전략이나 재고 전략, 판매촉진 등의 성과 제고에 활용할 수 있다.

시간에 따른 구매행위에 대한 지식은 유통업자들의 재고에 관한 의사결정에 많은 도움을 준다. 예를 들어 “ 오늘 한 고객이 캠코더를 구매하였다면 이 고객은 언제쯤 별도의 건전지와 추가적인 테이프를 구매할 것인가?” 와 같은 질문의 해답을 구하는데 많은 도움을 줄 수 있다.

유통업자들은 고객의 구매행위, 예를 들어 어떤 상품의 구매행위나 할인 행사에 참여하는 행위 등을 통하여 특성을 파악할 수 있으며 이러한 지식을 통하여 특정 고객집단을 겨냥한 효과적이고 경제적인 판매촉진책을 구사할 수 있다.

5) 제조업

최종 생산품의 품질에 영향을 미치는 요인발견, 경쟁사의 입찰액 예측, 제품의 수요 예측, 대리점 여신평가 모형 개발 등 여러 분야들에 적용될 수 있다.

6) 항공업

항공사는 자사 비행기를 더 자주 이용할 수 있도록 인센티브를 제공할 고객 집단을 찾아내는데 데이터마이닝을 이용할 수 있다. 예를 들면 한 항공사의 경우 비행거리 누적 혜택을 받는 데는 별 도움이 되지 않을 정도의 짧은 거리를 매우 자주 여행하는 고객집단이 존재함을 발견하였다. 따라서 이 항공사는 비행거리 뿐만 아니라 비행횟수에 의해서도 항공사가 제공하는 혜택을 받을 수 있도록 규칙을 변경하였다 [엄용환, 2000].

이처럼 데이터마이닝 기법을 이용하여 발굴되는 정보를 통해 기업은 선점 효과를 누릴 수 있는 경우가 많다. 즉 먼저 그 정보를 확보하여 활용함으로써 고객을 모으고 서비스를 제공하는 데 유리한 위치를 확보하게 되는 것이다.

Ⅲ. 커스터마이징 방법론

본 절에서는 커스터마이징을 통하여 보다 효율적이고 효과적인 데이터마이닝시스템의 구현을 위한 방법론을 제안한다.

1. 커스터마이징 관련 연구 고찰 및 분석

조직 내에서 정보시스템을 자체개발하는 경우 많은 시간과 비용을 소모하게 된다. 그런데 이렇게 개발된 시스템이 꼭 사용자의 모든 요구사항을 만족시키는 것도 아니다. 자체개발의 어려움에 대한 해결방안으로 패키지 소프트웨어의 사용이 제안되기도 하는데, 패키지 소프트웨어를 제대로 구축하여 성공으로 이끄는 데에는 여러 요인이 작용하게 된다.

Lucas[1998] 등의 논문에서는 패키지 소프트웨어를 구축하는 과정을 연구하였는데, 우선 패키지 소프트웨어를 2가지 유형으로 구분하였다. 하나는 사용자가 문제를 해결하는데 직접적으로 활용할 수 있는 Lotus 1-2-3과 같은 일반목적의 패키지이고 다른 하나는 사용자의 요구사항에 맞게 커스터마이징(customizing)하여 활용할 수 있는 Dedicated 패키지이다. 이 논문에서 커스터마이징의 의미는 이미 개발된 패키지를 사용자의 요구에 맞게 시스템의 기능을 조정하고 수정하여 사용자의 욕구를 충족시켜주는 일련의 활동이라고 정의한다. 커스터마이징의 유형은 변경, 추가, 확장의 3가지 형태로 구분하여 살펴볼 수 있다.

첫째, 변경(Modify)이라 함은 패키지에서 제공하는 기능이 업무처리상에서 필요한 기능과 상이하여 패키지의 프로그램 자체를 변경하는 활동을 말한다. 커스터마이징의 유형 중 가장 어렵다고 알려져 있다.

둘째, 추가(add-on)라 함은 패키지에서 제공되지 않는 별도의 기능을 개발하여 기존 기능들과 연계하여 사용할 수 있도록 하는 것을 말한다. 추가의 경우는 기존의 전통적인 개발방법과 거의 유사한 형태이다.

셋째, 확장(extension)의 경우는 변경과 추가의 중간 형태로서 기존의 패키지에 사용자가 원하는 기본적인 기능이나 모듈은 있으나 더 확장된 추가 기능이 필요하며 기본 기능이나 모듈에 부가적으로 프로세스나 기능을 첨가하여 개발하는 것을 말한다[Oracle 9i, 2000].

일반목적의 패키지는 완성된 형태로 제공되기 때문에 즉각적인 활용이 가능하지만 그 기능들이 사용자의 요구사항과 완전하게 일치하지 않을 수 있으므로 사용자의 불편과 불만이 생길 수 있고, 또한 사용자가 패키지의 기능들을 익히고 적응해야 한다.

Dedicated 패키지는 정제되지 않은 기본적인 기능들을 제공하지만 사용자의 필요에 따라 커스터마이징해서 사용할 수 있으므로 일반목적 패키지에 비하여 보다 융통성이 있다고 할 수 있다. 그러나 커스터마이징 작업에 대한 비용이 필요하고 어떻게 커스터마이징 했느냐에 따라서 패키지가 제공하는 기능에 대한 질이 결정되므로 효율적이고 효과적인 커스터마이징 방법론이 적용되어야 한다.

Dedicated 패키지의 대표적인 예로써 ERP(Enterprise Resourse Planning) 패키지가 해당된다고 할 수 있다. ERP패지지의 커스터마이징 방법론에 대해서는 몇몇 문헌에서 제시되고 있다. 김병곤[2000] 등의 연구에서는 ERP 패키지의 성공적인 커스터마이징을 위한 고려 요인으로 “ERP에 대한 정확한 이해, 목표의 명확한 설정, 철저한 도입준비, 최고 경영자의 강력한 의지, 커스터마이징의 최소화”를 들고 있다. 커스터마이징 작업 중 변경을 위해서는 원래의 패키지가 가지고 있는 기능이나 프로그램의 내용을 잘 알아야 하는데, 통상 일반 사용자나 개발자들에게는 패키지의 내용이 블랙박스화 되어 있기 때문에 시스템 전체를 파악하는 것이 그리 쉬운 일이 아니다.

Soh[2000] 등의 연구에서는 ERP시스템을 커스터마이징 해야 하는 경우, 추가는 하더라도 변경은 자제할 것을 제안한다. 일반적으로 패키지의 소스코드를 변경하는 것은 버전 상향시 유지보수의 어려움 때문에 피하는 것이 좋다. 조직의 중요한 기능을 위해 커스터마이징이 필요한 경우에도 소스코드의 변경보다는 추가모듈의 개발을 권고한다. 정승민[2002] 등의 연구에서는 커스터마이징을 많이 할수록 사용자 만족도 및 조직의 경쟁우위를 높일 수 있으며, 소스코드를 변경하는 것이 아니라 추가의 형태로 필요한 기능을 지원하여야 한다.

이상의 문헌고찰의 결과로 Dedicated 패키지의 커스터마이징은 사용자의 만족을 가져다 줄 수 있지만 패키지의 기능을 추가하는 방향으로 적용될 필요가 있음을 알 수 있다. 따라서 Dedicated 패키지가 제공하는 기본적인 기능을 API(Application Programming Interface)의 형태로 제공하고 API를 이용하여 필요한 기능들을 추가할 수 있는 형태의 Dedicated 패키지가 바람직할 수 있겠다.

2. 데이터마이닝의 커스터마이징 필요성

데이터마이닝 기능의 종류와 데이터마이닝 사용자의 유형은 다양하다. 즉, 데이터마이닝을 지원하는 제품은 다양한 기능을 제공해야 하고 다양한 사용자들에 의하여 편리하게 사용될 수 있어야 한다. 그러나 실제 기업에서는 기업에서 필요로 하는 고급정보를 만들기 위하여 데이터마이닝의 모든 기능을 필요로 하지는 않을 것이다. 예를 들어, 판매매장에서는 판매되는 제품간의 연관성 정보는 필요할 수 있지만 제품들을 군집화 하는 기능은 필요로 하지 않을 것이다. 또한, 데이터마이닝 제품은 데이터마이닝 전문가가 이용할 수 있어야 할 뿐만 아니라 판매점포의 점원도 편리하게 사용할 수 있어야 좋은 정보추출 도구가 될 수 있다. 결과적으로, 데이터마이닝 제품에 대한 요구사항은 사용자 또는 기업들에 따라 다양하다고 할 수 있으며, 이러한 다양한 요구사항을 하나의 일반목적 패키지의 형태로 제공하는 것은 비효율적일 수 있다. 패키지 용량이 커지면서 가격이 비싸지게 되고 모든 기능을 사용할 필요가 없는 사용자는 지불한 비용만큼 회수가 되지 않을 것이기 때문이다.

따라서, 데이터마이닝 기능을 제공하는 패키지의 형태도 Dedicated 패키지가 바람직할 수 있다. 데이터마이닝 기능을 제공하는 Dedicated 패키지는 데이터마이닝의 모든 기능을 API의 형태로만 제공하고, 사용자들은 필요한 기능들만을 API를 이용하여 추가 개발하는 방식으로 이용할 수 있다. 이러한 방식은 사용자들이 불필요한 기능들을 이용하지 않게 할 뿐만 아니라 사용자들의 수준에 적합하게 인터페이스(Interface)를 구현할 수 있기 때문에 사용자들의 만족도를 높여줄 수 있다.

3. 데이터마이닝 시스템의 커스터마이징 방법론

데이터마이닝 시스템을 구현하기 위하여 데이터마이닝 기능이 내장된 Dedicated 패키지를 효율적이고 효과적으로 커스터마이징 하기 위한 방법론을 제안한다.

선행연구에서 살펴보았듯이, 커스터마이징은 추가 형태로 적용될 때 그 성공 가능성이 높다고 할 수 있다. 추가유형의 커스터마이징은 전통적인 개발방법인 SDLC(System Development Life Cycle)와 유사하면서도 커스터마이징 이라는 특성이 반영되어야 하므로 단계별 실행지침들에서 차이점들이 존재한다. 추가 유형의 커스터마이징을 성공적으로 수행하기 위한 전략을 다음과 같이 제안한다.

전략 : 구축단계별로 자체개발과 커스터마이징의 차이점을 비교분석한다.

일반적으로 기존의 개발방법은 업무분석, 설계, 구축, 테스트, 이행, 유지보수의 개발 사이클을 가진다. 그러나 Dedicated 패키지의 구축에 있어서는 기존의 방법과는 다르게 보다 강화되어야 할 단계가 프로젝트의 준비단계이다. 프로젝트의 시작시점에 준비단계를 신설하여 이에 적합한 활동을 우선 실시하여야 한다. 일반적인 개발방법에 있어서는 앞에서 열거한 단계들을 진행하면서 하나의 상품화된 시스템이 탄생하게 된다. 그러나 Dedicated 패키지는 상품화된 시스템이 이미 존재하고 있으며, 문제는 자사의 요구사항을 정확하게 지원해줄 수 있는 기능이 제공되게 커스터마이징하기 위하여, Dedicated 패키지의 전반적인 이해를 위한 교육 및 연구 과정이 필요하다는 것이다. 또한, SDLC의 각 단계별로 이행되어져야 할 세부 시행지침들도 기존의 개발방법과 다른 관점에서 파악되어져야 한다.

<표 3-1>은 기존의 전통적인 개발방법과 커스터마이징 방법의 구축단계 별 차이점을 비교 분석하고 있다.

<표 3-1> 기존 개발 방법론과 커스터마이징의 단계별 비교분석표

구분	기존개발방법	커스터마이징
프로젝트 준비단계	• 별도로 정의된 사항이 없음	• dedicated 패키지의 기능 이해를 위한 전반적인 교육 또는 연구
업무분석 단계	• 현업 사용자와 면담 후 업무분석	• 업무분석을 통한 데이터마이닝의 응용 영역 이해 • 사용자의 요구사항 분석
설계 단계	• 분석 산출물을 중심으로 업무설계와 기술설계	• 사용자의 요구사항을 해결하기 위한 기능설계 • Dedicated 패지지에서 제공하는 기능을 이용하여 설계
구현 단계	• 개발툴과 DB관리 툴을 이용하여 설계내용에 따라 엔진기능을 개발	• Dedicated 패키지와 호환될 수 있는 개발툴을 이용하여 설계내용을 구현
테스트 단계	• 자동화된 테스트툴을 이용하거나 테스트 계획서를 작성하여 실시	• 테스트 시나리오를 작성하여 실시
운영 및 유지보수 단계	• 사용자의 요구사항에 맞도록 관리	• 추가개발 요청 시 Dedicated 패키지의 기능을 고려하여 결정

각 단계의 구체적인 차이점들과 이행지침들은 다음과 같다.

· 프로젝트 준비단계

기존개발방법	기존 개발 방법에서는 정의된 사항이 없다.
커스터마이징	프로젝트 참여자 전원에게 Dedicated 패키지에 대한 교육을 실시한다. Dedicated 패키지의 기능과 사용법, 호환 가능한 개발툴 등에 대하여 교육을 실시한다. 교육 여건이 어려운 중소기업들의 경우는 프로젝트 참여자 스스로 매뉴얼 등을 이용하여 연구함으로써 관련 지식들을 습득하도록 한다.

· 분석단계

기존개발방법	현업 업무 위주로 실사용자와 인터뷰 후 업무내용을 상세하게 기록한다.
커스터마이징	데이터마이닝의 응용 영역에 대한 이해를 정확히 함으로써 사용자의 요구사항을 최종 시스템에 정확히 반영할 수 있으므로 데이터마이닝이 활용될 업무분석을 수행한다. 업무에 대한 지식을 바탕으로 현업의 실 사용자와 인터뷰를 한 후, 사용자의 요구사항을 파악하고 필요한 데이터마이닝의 유형을 결정한다. 사용자가 요구하는 데이터마이닝 기능이 Dedicated 패키지에서 제공되고 있는 것인지 아니면 새롭게 개발되어져야 할 부분인지를 파악한다. 또한, 데이터마이닝 사용자의 유형을 결정한다. 즉, 데이터분석 전문가인지 또는 일반 사용자인지를 파악한다.

· 설계단계

기존개발방법	분석시의 산출물을 근거로 업무설계와 기술설계를 수행해 나간다.
커스터마이징	분석단계에서 요구되어진 데이터마이닝 기능들을 수행할 수 있는 모듈들의 논리(logic)를 Dedicated 패키지의 API 또는 라이브러리 등을 이용하여 설계한다. 필요한 경우 ERD(Entity Relationship Diagram)와 DFD(Data Flow Diagram) 등을 이용한다. 또한, 사용자의 요구사항과 사용자의 유형이 고려된 입출력 화면을 설계한다.

· 구축단계

기존개발방법	선정된 개발툴과 데이터베이스 관리툴을 이용하여 설계된 내용대로 구축작업을 진행한다. 본격적인 기능의 개발에 앞서서 각종 라이브러리나 API 등 엔진 기능을 먼저 개발하여야 한다.
커스터마이징	설계된 내용대로 구축작업을 진행한다. 개발툴은 Dedicated 패키지와 호환될 수 있는 것을 사용한다. 엔진 기능은 대부분 dedicated 패키지에서 API나 라이브러리 형태로 제공되므로 그대로 사용하면 된다.

· 테스트단계

기존개발방법	자동화된 테스트 툴이 있을 경우 이를 이용하고, 테스트툴이 없을 경우 테스트 계획서를 작성하고 이에 따른 테스트 시나리오와 테스트 데이터를 개발하여 이를 근거로 테스트를 실시하고 결과를 기록한다.
커스터마이징	자동화된 테스트툴을 이용하는 경우 Dedicated 패키지에서 제공하는 API에 대한 트레이스(trace)는 불가능하므로 테스트 시나리오를 만들 때 이러한 사항을 고려한다.

· 운영 및 유지보수단계

기존개발방법	현업의 요구대로 개발된 내용이 원래의 의도나 시나리오대로 가동이 되는가를 면밀히 모니터링하고 실 사용자의 요청에 의거, 지속적으로 시스템을 개선해 나간다.
커스터마이징	사용자의 추가개발 요구 시 우선적으로 Dedicated 패키지 내에서 해결책을 찾아야 한다. 그러기 위해서는 전 유지보수 요원이 패키지의 기능에 대해서 숙달되어 있어야 한다.

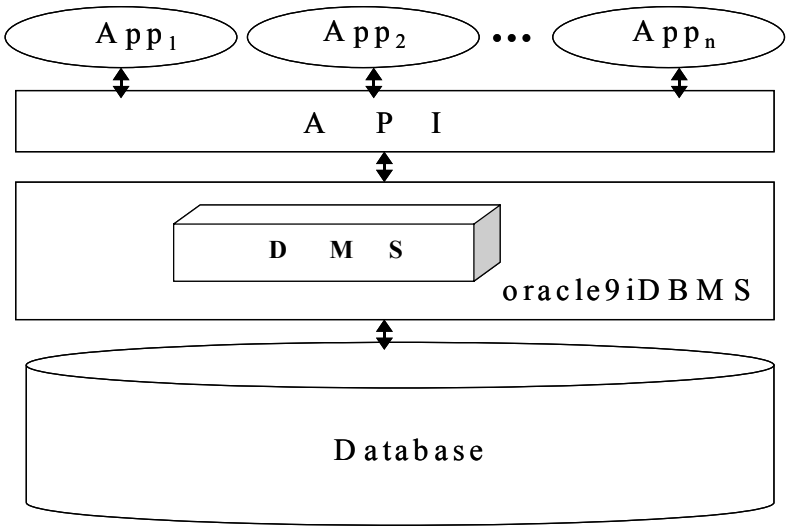
Ⅳ.오라클 데이터마이닝 시스템의 커스터마이징

본 절에서는 Dedicated 패키지의 일종인 오라클 데이터마이닝을 이용하여 앞에서 살펴 본 커스터마이징 방법론에 따라 데이터마이닝 시스템을 구축하는 사례를 제시한다.

1. 오라클 데이터마이닝 시스템 구성도

오라클 데이터마이닝은 오라클 9i DBMS(DataBase Management System)와 연동되어서 데이터마이닝 기능을 제공할 수 있는 Dedicated 패키지이다. 오라클 데이터마이닝에는 연관규칙탐사 기능과 분류기능을 지원하기 위한 API들이 포함된다.

오라클 9i DBMS와 오라클 데이터마이닝이 결합 되어져서 구성된 오라클 데이터마이닝시스템의 구성도는 <그림 4-1>과 같다.



<그림4-1> 오라클 데이터마이닝 시스템의 구성도

<그림 4-1>에서 API(Application Programming Interface)부분은 오라클 데이터마이닝을 설치하면 이용할 수 있는 라이브러리 함수이다. API의 각 기능들은 DMS(Data Mining Server)의 서비스를 통하여 수행된다. DMS는 Oracle9i DBMS 내에 포함되어 있고 데이터마이닝 기능을 수행하는 엔진이라고 할 수 있다. Appi는 데이터마이닝 기능을 포함하는 응용프로그램들인데 사용자의 요구사항에 따라 API를 이용하여 개별적으로 만들어질 수 있다. 데이터마이닝 프로젝트의 결과물은 사용자가 최종적으로 이용할 수 있는 App_i들이라 할 수 있다.

2. 커스터마이징에 의한 오라클 데이터마이닝 시스템 구축

커스터마이징 방법론을 이용하여 오라클 데이터마이닝 시스템을 구축할 때, 각 단계에서 실행되어져야 할 사항들을 사례를 통하여 구체적으로 살펴본다. 본 사례는 앞에서 제시한 커스터마이징 방법론의 타당성 검토를 위한 수단으로서 프로토타입(prototype)형태로 구축되었다.

· 프로젝트 준비 단계

오라클 데이터마이닝의 매뉴얼을 분석하여 오라클 데이터마이닝에서 제공하는 기능, 특징들을 고찰한다. 오라클 데이터마이닝에서 제공하는 API들의 종류들과 개발도구에 대하여 학습한다. API는 자바 클래스 및 메소드의 형태로 제공되고 오라클 데이터마이닝과 연동될 수 있는 개발도구는 자바언어이므로 자바언어에 대하여 학습한다.

· 분석 단계

사용자의 요구사항은 판매매장에서 POS(Point of Sales)시스템에 의하여 생성된 판매데이터를 분석하려는 것이고, 사용자의 유형은 데이터마이닝에 대한 전문지식이 있는 데이터 분석 전문가로 가정하였다. 입출력화면은 편의상 커맨드방식으로 구현할 것이다.

· 설계 단계

판매 데이터들 사이의 연관성을 분석할 수 있는 모듈을 API의 기능을 고려하여 설계한다. 모듈의 로직은 다음과 같다.

```
/* 연관규칙 탐사 프로그램의 절차 */

- 파일입력을 통하여 환경변수 값을 설정한다.
- DMS와 연결한다.
- 분석용 데이터화일의 형태를 변환한다.
- 연관규칙을 탐사한다.
- 탐사된 규칙들을 파일을 통하여 출력한다.
```

<그림 4-2> 연관규칙 탐사 프로그램의 설계

· 구현 단계

설계단계의 모듈을 자바 언어를 이용하여 코딩한다. Edit plus나 기타 자바 개발 툴을 이용한다.

· 테스트 단계

API의 입출력 데이터가 포함된 테스트 시나리오를 작성하여 시스템의 오작동 여부, 사용자 요구사항의 만족여부를 테스트한다.

바. 운영 및 유지보수 단계

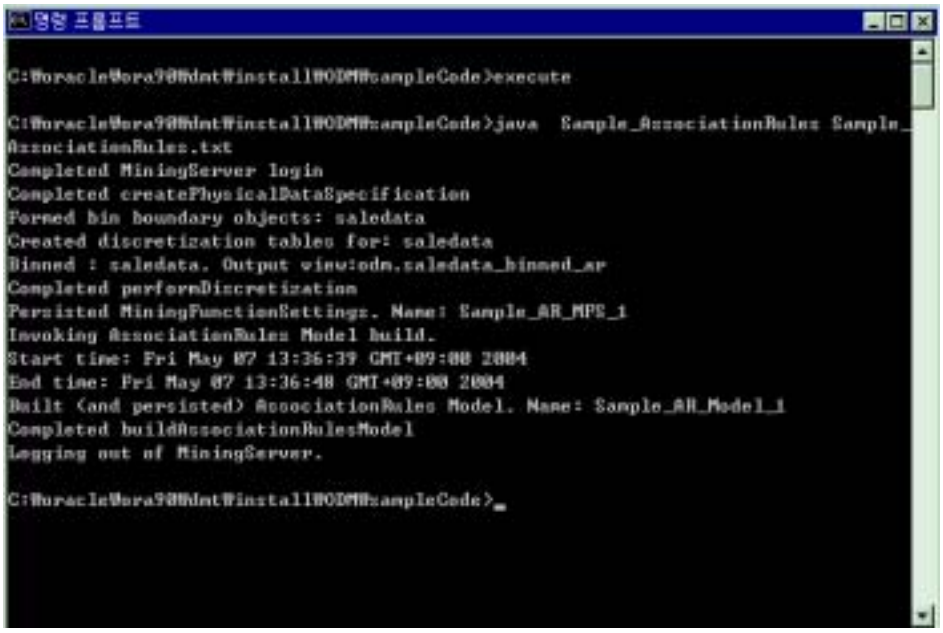
판매데이터를 이용하여 데이터마이닝 프로세스에 따라 연관규칙 탐사를 한다. Ⅴ장에서 실제의 판매데이터를 분석한 사례를 제시한다.

V. 오라클 데이터마이닝을 이용한 데이터 분석

본 절에서는 구축된 오라클 데이터마이닝 시스템을 이용하여 대형 할인매점의 판매데이터들 사이의 연관성을 분석한다. 판매 데이터의 트랜잭션(transaction) 수는 494건으로 데이터양이 작다고 할 수는 있지만 시험용 분석이라는 측면에서 타당성이 있다고 하겠다. <그림 5-1>는 494건의 판매 트랜잭션 데이터를 담고 있는 관계형 파일이다. 판매 물품들의 종류 즉, 항목들의 종류는 17가지이고 한 트랜잭션의 최대 항목 수는 10개로 설정하였다.

<그림 5-1> 판매 트랜잭션 데이터

<그림 5-2>는 오라클 데이터마이닝을 이용하여 판매데이터 분석을 실행하는 화면이다. 실행시간은 9초 걸리는 것으로 나타나고 있다.



<그림 5-2> 오라클 데이터마이닝의 실행화면

<표 5-1>은 연관규칙탐사를 위한 환경변수 값들에 따라 도출된 규칙들을 정리한 것이다.

<표 5-1>도출된 규칙들

환경변수	규칙	지지도	신뢰도
최소지지도 : 0.01 최소신뢰도 : 0.5 규칙의 길이 : 2	rule : 잡화 → 식품	0.060728744	0.51724136
	rule : 빙과 → 제과	0.030364372	0.6818182
	rule : 담배 → 제과	0.020242915	0.5263158
	rule : 수산물 → 야채	0.01417004	0.5833333
	rule : 야채 →청과	0.01417004	0.5833333
	rule : 수산물 → 식품	0.012145749	0.5
최소지지도 : 0.01 최소신뢰도 : 0.5 규칙의 길이 : 3	rule : 식품, 제과→ 잡화	0.018218623	0.6
	rule : 빙과, 음료→제과	0.016194332	1.0
	rule : 잡화, 음료→제과	0.016194332	0.53333336
	rule : 잡화, 음료→ 식품	0.016194332	0.53333336
최소지지도 : 0.01 최소신뢰도 : 0.5 규칙의 길이 : 4	rule : 음료, 제과, 식품→잡화	0.012145749	0.85714287

최소지지도는 1%로 하고 최소신뢰도는 50%로 설정하였다. 도출된 규칙들은 우리가 일반적으로 기대했던 결과를 반영하고 있다.

예를 들면, [rule : 잡화 → 식품]의 의미를 해석해 보면, 잡화와 식품을 포함하는 트랜잭션 건수는 전체 트랜잭션 494건 중 30건(즉, 0.06 0728744 = 30/494)이고, 잡화를 포함하는 트랜잭션이 식품도 포함할 확률은 0.51724136, 즉 약 50% 라는 말이다.

달리 해석해 보면, 잡화와 식품은 많이 팔리는 제품이고 서로 연관성도 높다는 의미이다. 특히, 주목할 규칙은 [rule : 빙과, 음료→제과]로써 빙과와 음료를 구매하면 제과도 함께 구매될 확률이 100%로 나타나고 있다. 전반적으로 볼 때, 기대하지 않았던 특이한 규칙은 도출되지 않았음을 알 수 있다. 이는 데이터마이닝이 의미 있는 의외의 결과를 항상 제공하지는 않음을 알 수 있다.

Ⅵ. 결 론

데이터마이닝은 축적된 대규모의 데이터들을 분석하여 기업의 경영활동에 도움이 될 수 있는 정보와 지식을 획득할 수 있는 기술이다. 데이터마이닝 기술에 의하여 다양한 정보 분석 및 지식추출 기능을 활용할 수 있는데, 사용자는 이러한 다양한 기능들을 전부 필요로 하지 않고 업무특성에 따라 제공 기능의 일부분을 이용한다. 또한, 데이터마이닝 기술은 전문가뿐만 아니라 비전문가도 편리하게 활용할 수 있어야 보다 폭넓은 공헌을 할 수 있다. 결과적으로, 데이터마이닝은 사용자에 따라서 다양한 요구사항이 존재할 수 있으므로 커스터마이징에 의한 데이터마이닝 시스템 구축의 필요성이 생기게 된다.

본 논문에서는 데이터마이닝 시스템을 커스터마이징에 의하여 효율적이고 효과적으로 구축하기 위한 방법론을 제시하였다. 기존의 개발 방법론과의 비교분석 전략을 통하여 체계적인 단계별 실행지침을 제시하였다. 커스터마이징 방법론은 기존의 개발방법에 비하여 프로젝트 준비단계의 비중을 중요시 할 필요가 있고 미리 제공되는 라이브러리 기능을 정확히 파악하여 설계, 구현 단계에서 효과적으로 이용하여야 한다.

본 논문에서 제안한 커스터마이징 방법론에 대한 타당성 검토를 위하여 커스터마이징 방법론의 각 단계별 실행지침에 따라서 오라클 데이터마이닝 시스템을 구축해보았다. 또한, 대형 할인매장의 판매 분석 데이터를 확보하여 구축된 데이터마이닝 시스템을 이용하여 연관규칙 탐사를 실행하였다.

커스터마이징 방법론을 도출하는데 있어서 커스터마이징 전문가들의 의견을 수렴하여 반영할 필요성이 있지만, 이를 수행하지 못한 점은 본 연구의 한계임을 밝혀둔다.

향후, 데이터마이닝 시스템의 커스터마이징에 관한 연구는 본 연구를 시발점으로 하여 커스터마이징에 관한 많은 전문가들의 다양한 의견을 수렴하여 신뢰성과 타당성을 입증할 수 있는 실증적 연구가 이루어져야 할 것이다.

<참고 문헌>

국내문헌

김병곤, 오재인(2000), “ERP 패키지의 성공적인 커스터마이징 전략”, 경영정보학연구, 10권 3호, pp. 123~137.

김양석(2000), “게놈 프로젝트를 위한 생물 정보학” 정보과학회지 제 18권, pp. 551.

김정자, 이도헌(1998), “데이터마이닝 기술 및 연구동향”, 정보과학회지 제16권 9호, pp. 6~14.

김종훈(2000), “데이터마이닝에 기초한 DB 마케팅 분석 방법과 사례” 월간 경영과 컴퓨터 10월호.

박종수, 유원경, 홍기형(1998), “연관 규칙 탐사와 그 응용”, 정보과학회지 제 16권 pp. 37.

엄용환(2000), “새로운 정보기술, 데이터마이닝, 행정과 전산” 82집 pp. 78~79.

장남식(1999), 데이터마이닝, 대청미디어, pp. 19~26, 125~130.

정승민, 김준석(2002), "ERP 시스템 도입 시 커스터마이징 정도가 사용자 만족도와 조직의 경쟁우위에 미치는 영향", 한국경영정보학회 2002 추계 학술대회, pp. 529-540.

하단심, 황부현(2001), "상대 지지도를 이용한 의미 있는 희소 항목에 대한 연관 규칙 탐사 기법", 정보과학회지, 제 28권 4호, pp. 578.

외국문헌

Adrianns, P.Zantinge(1996), Data mining, New York NY:Addison-Wesley.

Berry.M.J.A. & Linoff.G.(1997), Data mining techniques: For marketing sales and customer support, New York, NY : John Wiley & Sons.

Hwang, K T(1991), Evaluating The Adoption, Implementation, and Impact of Electronic Data Interchange Systems, Unpublished Ph. D. Dissertaton, State University of New York at Buffalo.

John Shafer Rakesh Agrawal & Marish Mehta(1996), "SPRINT: A scalable parallel classifier for data minning", the VLDB Conference, Bombay, India.

Lucas, H.C., Walton, E.J., and Ginzberg, M.J.(1988), "Implementing Packaged Software", MIS Quarterly, pp. 537-549.

Minos N(1999), Garofalaklis, Rajeev Rastogi and Kyuseok Shim " SPIRIT: Sequential Patten Minning with Regular Expression Constraints", the VLDB Conference, Edinburgh, Scotland, UK.

Oracle 9i Data Mining Administrator's Guide in oracle manual, 2000.

Oracle 9i Data Mining Concepts in oracle manual, 2000.

Pieter Adriaans, Dolf Zantige(1996), " Data minning" Addison Wesley.

Pyo,S.P.,Uysal. M.&Chang, H.S(2002), "Knowledge discovery in database for tourist destinations", Journal of Travel Research, 40, May.

Rakesh Agrawal and Ramakrishnan Srikant(1995), " Minning sequential patterns", International I conference on Data Engineering Taipei, Taiwan, March.

Rajeev Rastogi and Kyuseok Shim(1998), "PUBLIC : A decision tree classifier that integrates building and Pruning, the VLDB Conference New York.

Soh, C., Kien, S.S., and Joanne Tay-Yap.(2000), "Cultural Fits and Misfits: Is ERP a Universal Solution?", Communications of ACM, pp. 47-51.

Sudipto Guha, Rajeev Rastogi and Kyuseok Shim(1999), ROCK : A Rovost Clustering Algorithm for Categorical A tributes" the 15th International Conference on Data Engineering, Sydney.

Tan Zhang, Raghu Ramakirshnan, and Miron(1996), "BIRCH: An efficient data clustering method for very large databases" the ACM SIGMOD Conference on Management of Data.

ABSTRACT

Study on Methodology for Customizing Data Mining System

Kyoung-Hun Oh
Department of Management Information Systems
Graduate School of Business Administration,
Cheju National University
Supervised By Professor Keun-Hyung Kim

Data mining is a technology to analyse accumulated data and sort out the information and knowledge which are vital for companies to pursue profits. The technology of Data mining includes association rule mining, classification, clustering, summarization and sequential pattern.

As a finished product, most commercialized products of Data mining are supposed to support some or all of these functions. These products have an advantage to be available as soon as they are purchased. But, they have some limit to meet users' needs and companies' requirements and to be adaptable to different settings.

This study is focused on how to introduce Data mining system and implement it as a form of customizing, instead of a form of finished product. Also, this study try to present several cases which actually used customizing with Oracle products and some other cases which used constructed data mining to analyse sales data.