

## 성공적인 e-Business를 위한 인공지능 기법 기반 웹 마이닝

이장희

한국기술교육대학교 산업경영학부  
(ianghlee@kut.ac.kr)

박상찬

한국과학기술원 산업공학과  
지식기반시스템 연구실  
(sangpark@cais.kaist.ac.kr)

유성진

한국과학기술원 산업공학과  
지식기반시스템 연구실  
(genesis@major.kaist.ac.kr)

웹 마이닝은 e-Business 환경下에서 존재하는 대량의 웹 데이터에 데이터 마이닝 기법을 적용하여 유용하고 이해 가능한 정보를 추출해내는 과정을 의미하는데, 성공적인 e-Business 전개를 위한 핵심적인 기술이다. 본 논문은 인공지능 기법에 기반한 웹마이닝 기술을 활용하여 e-Business 상의 온라인 고객의 특성을 분석할 수 있는 data visualization system과 구매 판매 예측시스템의 효과적인 구조와 핵심적인 분석절차를 제안하였다.

### 1. 서론

인터넷의 활성화와 더불어 컴퓨터가 저렴해지고 계산능력이 비약적으로 향상됨에 따라 중소기업도 기업간 (business to business, BtoB) 전자상거래 실행을 위한 기본적인 하드웨어 인프라를 구축할 수 있게 되었다 (Alonso, G. et al., 1999). 기업간 전자상거래는 기업간 상품의 구매 및 판매에 대한 제반 비용이 거의 필요하지 않으므로 많은 기업이 관심을 가지게 되었고, 그 요구에 의해 최근 수 많은 기업간 전자상거래 업체가 생겨나고 있다.

기업간 전자상거래는 일반적으로 2 가지 실행 방법이 있는데, 첫번째로 전자화된 문서형태로 미리 선정된 공급자들로부터 구매를 처리하는 대기업들의 전통적 접근 방법이 있다. 다른 접근

방법은 전사적인 네트워크의 개념을 기반으로 한 것으로 여러 개의 기업이 그들의 서비스를 협작하여 더 복잡하고 가치 있는 상품을 제공하는 것이다 (Zhong Tian et al, 1999). 이러한 방법은 중소/중견 기업이 선택할 수 있는 가장 효과적인 접근 방법으로, 인터넷의 관점에서 자연스러운 모델로서 가상 기업(virtual enterprise), 가상 비즈니스 프로세스(virtual business process), 그리고 거래 공동체(trading community)의 개념과 함께 모델링되어 질 수 있다.

최근 특화된 상품을 팔거나 인터넷 쇼핑몰의 형태를 가지는 기업 대 소비자간(business to customer, BtoC) 전자상거래 업체는 물론이고 기업간 전자상거래 업체도 많이 생겨나고 있고 또한 벌써 많은 업체들이 도산하고 있는 실정이다. 이들 기업들은 인터넷을 기반으로 하기 때문에

기존 고객이 이탈 하기가 쉽고 또 새로운 서비스업체를 찾기가 쉽기 때문에 고객들의 충성도가 오프라인상의 기업들에 비해서 많이 낮은 편이다. 그럼에도 불구하고 아직까지 고객의 특성을 분석하고 고객 만족을 위한 전략을 제시할 수 있는 분석도구나 방법론에 대한 연구가 아직 미미하거나 초보적인 수준에 불과하다.

실제로 인터넷을 통한 상거래에서 발생하는 데이터는 매우 많다. 웹 사이트에 접속할 때마다 생기는 Web Log, 고객들의 기본적인 인구통계 데이터, 고객들이 상품을 구매 혹은 판매한 정보 등 아주 많은 데이터들이 있다. 일반적으로 쇼핑몰에서 하루에 저장되는 Web Log 데이터의 양이 200메가 바이트가 넘는데, 이와 같이 수많은 양의 데이터로부터 고객들의 특성 정보를 찾아내고 고객의 패턴을 알아내기가 쉽지 않다. 대부분의 web log 분석 도구는 1차원적 분석만을 수행하므로 고객의 패턴이나 접속 행태를 알아내기는 어렵다. 더욱이 많은 경우, 고객에 대한 분석을 통해 고객 만족도를 향상하려는 노력보다는 web traffic 분석에만 치중하여 웹·호스팅의 안정성 유지에만 중점을 두고 있는 실정이다. 그러므로 현재와 같이 인터넷을 기반으로 하는 전자상거래가 일반화 되고 있는 상황에서 고객 만족도 향상을 위한 기본 전략을 제시할 수 있는 고객에 대한 분석과 접속행태에 대한 분석 방법의 개발이 필요하다 하겠다.

또한, 기업 대 소비자간 전자상거래와는 다르게 기업간 전자상거래에서는 기업의 판매량 및 구매량에 대한 정확한 예측이 필요할 뿐만 아니라, 나아가 고객 기업의 재고 정보와 생산 일정을 포함하는 정보를 바탕으로 생산 계획 및 판매 계획을 정확하게 수립해 주는 것도 매우 중요하다. 본 논문에서는 기업간 전자상거래의 고객 정

보 및 접속행태에 관련한 데이터를 바탕으로 효과적인 e-Business를 위한 고객의 특성 및 패턴을 분석하는 방법론과 고객 기업에 필요한 구매 및 판매에 대한 예측정보를 제공하기 위한 방법론을 제시하고자 한다.

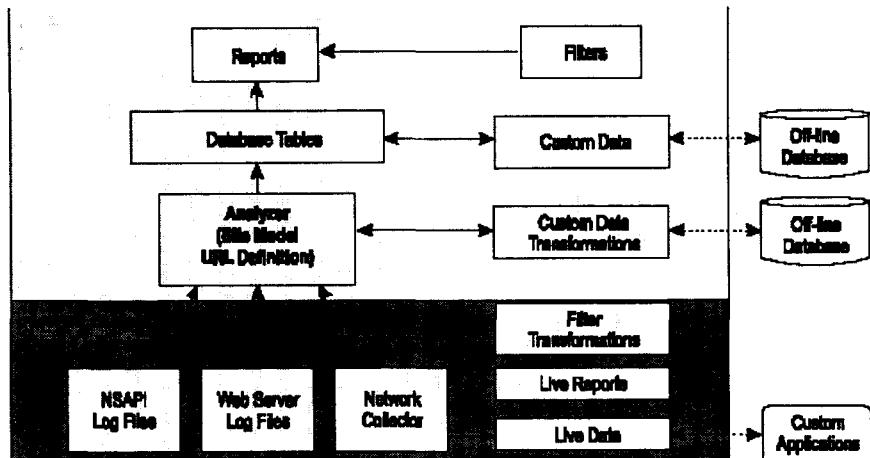
## 2. 기존연구 고찰

앞에서 언급한 바와 같이, 인터넷을 기반으로 한 e-비즈니스 상에서 기업의 고객들은 자신의 구매 기호에 맞고 원하는 서비스를 제공할 수 있는 기업을 인터넷에서 쉽게 찾을 수 있는 이점을 가지기 때문에 오프라인 상의 고객들에 비해 상대적으로 기업에 대한 로열티가 낮다 (Dussart, C., 2001). 그러므로, e-비즈니스를 수행하는 기업은 항상 고객의 관심을 유도하고 고객의 특성에 맞는 서비스를 제공해야만 고객의 이탈을 막을 수가 있고, 이러한 서비스를 제공하기 위해서는 고객의 특성을 파악하여 그 특성에 따라서 차별화된 서비스를 제공하여야 한다.

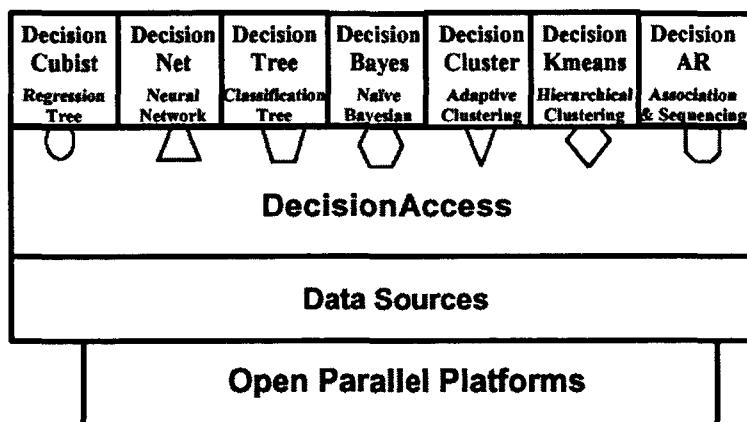
반면에, 오프라인 상의 서비스 기업에 비해 e-비즈니스를 수행하는 기업이 가질 수 있는 강점은 고객의 특성을 알아내기 위한 기본 데이터를 쉽게 얻을 수 있다는 것이다. 즉, <그림 2.1>과 같이 웹 데이터가 생성되어 이를 활용할 수 있다. 웹 데이터는 Netscape 웹서버 및 모든 웹서버의 로그 파일과 Network Collector로부터 Analyzer가 분석후 DB Table을 생성함으로써 활용할 수 있다.

효과적인 e-비즈니스의 전개를 위해 <그림 2.1>과 같은 웹 데이터를 이용한 3 단계의 웹 활용법을 고려해 볼 수 있다.

- 1) Web Searching : 임의 질의나 일반 사용자



<그림 2.1> Data Pipeline



<그림 2.2> 마이닝 분석엔진 (Mining Engines)

에 초점을 두며 일시적인 검색 프로세스이다.

2) Web (Content) Mining: 특정 주제를 가진 질의나 특정 사용자에게 초점을 두며 상호 작용적인 마이닝 프로세스이다.

3) Web (Data/Traffic) Mining: 데이터 마이닝을 단순히 웹에 확장한 것으로 많은 양의 웹 데이터에 숨겨진 패턴을 분석하여 효과적인 비즈니

스 전략적 지식을 찾는 일련의 과정을 의미한다.

이중에서 고객의 정보를 e-비즈니스에서 전략적으로 활용하기 위해서는 Web Traffic 분석이 필요한데, 이러한 Web Traffic 분석은 Web Log 분석과 Web Mining과 같은 2가지로 분류할 수 있다. Web Log 분석은 일반적인 Web Traffic 분석에 해당하는 것으로 웹 서버 로그 파일을 기

초로 하여 기초적인 traffic 정보를 분석하여 웹 사이트에서 일어나는 것에 대한 일반적인 정보를 제공한다. Web Mining은 일반적인 Web Traffic 분석과는 다르게 데이터 베이스와의 결합을 통한 분석으로 하부의 데이터 베이스나 하드웨어의 종류에 관계없이 모든 데이터 소스(Source)를 통합하여 분석하여 특정 패턴이나 상호연관 관계를 찾아낸다 (R. Cooley et al., 1997). 이러한 마이닝을 위한 도구는 <그림 2.2>와 같다(Osmar R et al., 1998).

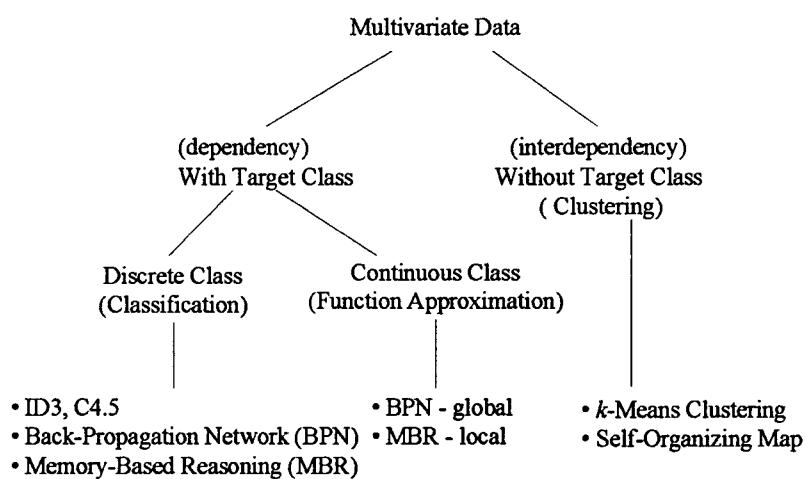
<그림 2.2>에서 볼 수 있듯이, 다변량의 데이터에 숨겨진 지식(Knowledge)을 추출하기 위한 많은 인공지능 도구들이 개발되어 왔는데, 그 중에서 가장 많이 사용되는 도구는 의사결정나무(Decision Tree)를 생성하는 C4.5와 신경회로망의 역전파 네트워크(back-propagation network, BPN)이다 (Irani, K. B. et al., 1993).

이 도구들은 사용이 간단하고 광범위한 적용 범위를 가지는 장점이 있는데도 각각의 알고리즘이 가지는 본질적인 제한 때문에 지식기반시스템

을 개발하는 목적에는 개별적으로 사용되지 않는 실정이다. 예를 들면, C4.5는 연속적인 값을 가지는 데이터에 직접적으로 적용될 수 없고(Aha, D.W., 1992), BPN은 추출된 지식을 해석하기가 불가능하다 (Benitez, J. et al., 1997). 최근 이러한 한계를 극복하고 더욱 효과적인 지식을 얻기 위해서 다음과 같은 혼합 전략(hybrid strategy)이 연구되고 있다.

### (1) 혼합전략 1: SOM (Self-Organizing Map) Inductive Learning

일반적으로 Inductive Learning은 분류 문제(classification task)에 사용된다. Inductive Learning에 의해 추출된 지식은 일반적인 사용자도 쉽게 인식할 수 있는 형태인 의사결정나무로 표현되고 학습 시간이 신경회로망 보다 훨씬 짧기 때문에 지식 추출(Knowledge Extraction)에 매우 유용하다. 그러나 Inductive Learning은 출력 변수 (Output)가 이산적이어야 하는 제약이



<그림 2.3> 여러 가지 데이터 형태에 대한 인공지능 분석도구

있고 학습 알고리즘이 정보 이론 (Information Theory)에 기반하므로 출력 변수가 여러 개 존재하는 문제에는 적용하기가 힘들다.

그러나, 이러한 출력 변수가 연속적이어야 하는 문제점은 SOM 신경회로망의 도움으로 극복 할 수 있다(Kohonen, T., 1982). SOM은 연속적인 입력 공간을 이산적인 출력공간으로 변환하는 역할을 하는 알고리즘으로, 연속적인 값을 갖는 출력 변수와 이의 독립변수들로 구성된 데이터 집합을 SOM으로 학습시키면 학습결과에 근거하여 이산적인 클래스로 변환할 수 있고 Inductive Learning을 위한 학습자료로 재구성할 수 있다. Kang et al. (1998)은 앞에서 언급한 바와 같이 출력 변수인 반도체 수율 데이터에 SOM을 적용하여 이산적인 클래스로 변환하였고, 변환된 데이터에 Inductive Learning을 적용하여 수율 행태에 대한 지식을 의사결정나무 형태로 표현하는 연구를 하였다.

## (2) 혼합전략 2: Inductive learning SOM

혼합전략 1과는 반대의 순서로 적용하는 것으로 Inductive learning을 통해 다수의 변수간의 차이점이나 가장 중요한 변수를 찾아내고 SOM으로 학습하여 학습데이터 그룹을 동일한 특성을 가지는 클러스터로 그룹핑하여 그룹별 특성을 파악할 수 있다. 이러한 방법은 이동통신 산업에서 고객들의 인구 통계적 데이터를 분석하는 데 사용되었다 (Yu, S. J., Park, S. C, 1999). 인구 통계적 데이터는 고객의 생활패턴을 기술한 것이므로 출력변수의 타겟 클래스는 존재하지 않는다. 이러한 데이터를 Inductive learning 과 SOM을 연속적으로 적용하여 고객들에게 나타나는 특징적인 몇 가지 생활 패턴을 추출하고, 추출된 생

활 패턴별로 중요한 특성/변수를 파악하는 것이다.

## (3) 혼합전략 3: BPN SOM Inductive learning

위 세가지 분석 도구를 혼합하여 사용한 예를 Yu et al.(1998)이 제안한 인공지능 기반 반도체 제조공정에서의 수율 예측 시스템에서 확인할 수 있다. 일반적인 인공 신경회로망에 기반한 예측 시스템은 선처리 (Preprocessing) 과정 모듈과 학습 (Learning) 모듈로 구성되어 있는데 비해, Yu et al.(1998)이 제안한 시스템은 선처리 과정 모듈, 특성 변환 (Feature Transformation) 모듈, 학습 모듈의 세가지 모듈을 가진다.

다수의 변수들이 서로 복잡하게 연결되어 있는 상황에서는 관리 변수와 출력간에 직접적인 상관관계는 거의 보이지 않고, 인자/주성분 분석 (Factor/ Principal Component Analysis)의 인자/주성분과 같이 보이지 않는 어떤 매개체, 즉 Feature가 존재해서 이것이 관리 변수와 출력간의 연결고리 역할을 한다. 이 Feature들은 입력 공간의 차원과는 다른 차원을 가지는데, 이들은 다시 SOM에 의해 간단한 차원을 가지는 새로운 차원 공간으로 변환된다. Yu et al. (1998)은 BPN의 은닉 층으로부터 Feature들을 추출했다. 즉, 복잡한 관련성을 가지는 학습 자료의 공간이 BPN과 SOM을 사용하여 간단한 관계를 가지는 새로운 공간, 즉 Feature의 공간들로 맵핑된 것이다. Yu et al. (1998)은 추출된 Feature와 이산적인 Class를 가지는 수율과의 상호관계를 나타내는 지식을 Inductive Learning을 통해 추출하여 이를 기반으로 예측하였고, 그들은 제안된 시스템이 입력과 출력의 관계가 비선형적이고 불연

속적일 경우에 잘 작동한다고 보고하였다.

#### (4) 혼합전략 4: BPN MBR

BPN과 Memory-based reasoning (MBR, instance based learning이나 case based reasoning으로도 알려져 있다)은 타겟 클래스를 가지는 문제를 처리할 때 가장 많이 적용하는 도구이다. 이를 알고리즘은 시스템에 대한 모델링 없이 학습이 가능하다는 특성 때문에 적용분야에 대한 사전 정보 (*priori* information)가 없는 경우에도 적용될 수 있는 반면 몇 가지 문제점을 가지고 있다. MBR은 과거의 사례(Case)를 단순히 메모리에 저장시키므로 입력 변수들이 출력 값에 미치는 영향에 대한 차이점을 고려하지 않고 같은 중요도를 가진다고 가정하는 문제점이 있다. BPN은 사전정보를 전혀 이용할 수 없기 때문에 최적의 네트워크를 형성하는데 데 걸리는 시간이 긴 문제점이 있다.

이러한 문제점을 극복하기 위해 최근에는 MBR과 BPN을 혼합하는 전략이 연구되고 있는

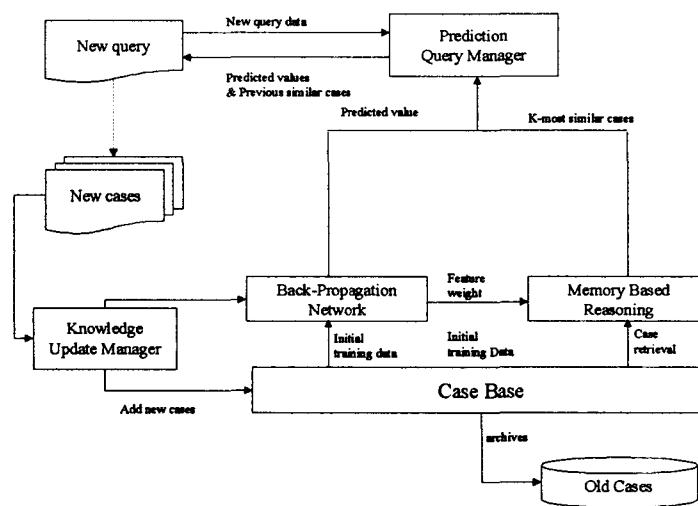
데, Shin et al. (1999)은 이러한 BPN후 MBR 적용이라는 혼합전략을 도입하여 반도체 제조공정에서의 수율 관리 시스템을 제안하였다(<그림 2.4> 참조).

Shin et al. (1999)은 이러한 혼합전략을 사용함으로써 수율 예측이 더욱 정확해졌고, 추출된 지식이 안정된 수율 유지를 위한 공정관리에 활용될 수 있다고 발표했다.

### 3. 연구방법론

#### 3.1 SOM 분석에 기반한 e-Business상의 고객 특성 분석

Web Log와 같은 웹 데이터(<그림 2.1> 참조)와 기업내의 고객DB, 매출DB, 상품 DB를 연계하여 인공지능 기법을 통해 e-Business상의 고객 특성을 효과적으로 분석하고자 하는 본 연구의 프레임워크는 <그림 3.1.1>과 같다.



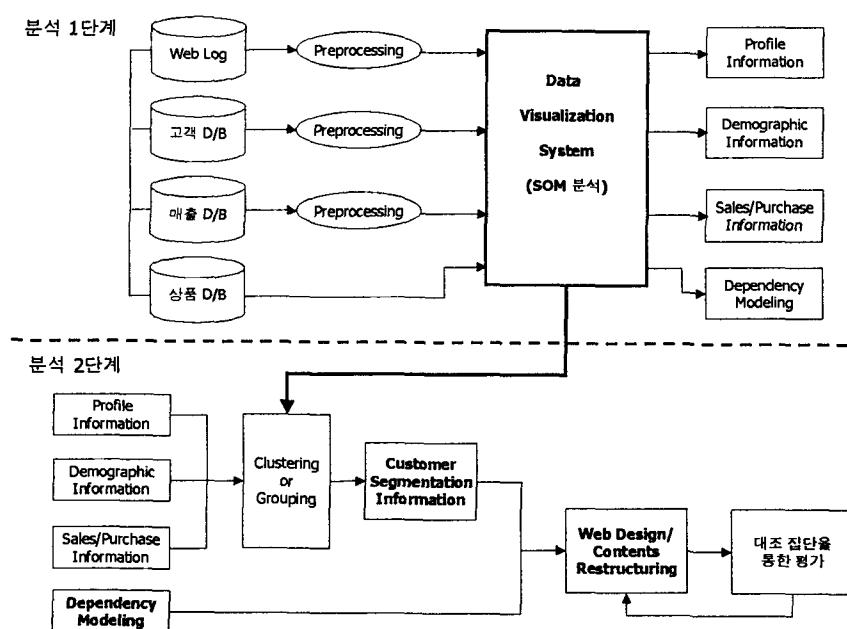
<그림 2.4> BPR & MBR 혼합전략을 가지는 시스템

고객이 웹 사이트에 접속하면 웹 서버의 로그 파일에 단순히 접속이라는 데이터로 기록되며, HTML 문서 한 페이지를 검색해도 접속은 해당 페이지에 포함된 다양한 형태의 데이터마다 기록을 남긴다. 일반적으로 표준적인 로그파일은 접속 당 7개의 필드 값인 Host, RFC931, AuthUser, Time, Request, Status, Volume을 가지는 데이터로 구성되는데, 이러한 자료를 이용하여 7가지의 대략적인 1차원적 분석은 수행할 수 있다: 1) 기간별 분석, 2) 시스템 분석, 3) 에러 분석, 4) 사용자(방문자) 분석, 5) 페이지 분석, 6) 디렉토리 분석, 7) 메뉴 분석.

그러나, 이러한 1차원적인 분석은 시스템의 관리적인 측면에는 많은 도움을 줄 수 있지만, 고객의 접속 패턴 분석에는 아무런 도움을 주지 못한다. 고객의 접속 패턴이나 고객의 접속 선호도

(preference)를 알기 위해서는 웹 로그 데이터의 선처리(preprocessing)가 필요하고(<그림 3.1.1>의 분석 1단계 참조), 다차원의 자료를 한꺼번에 분석할 수 있는 분석 도구가 필요한데 본 연구에서는 SOM에 근거한 Data Visualization System의 활용을 제시하였다. 즉, Data Visualization System은 Web Log 데이터로부터 SOM에 의한 다차원 분석을 통해 고객의 접속행태와 고객이 관심을 가지는 상품 품목에 관한 정보를 제공한다.

또한 기업내 고객 DB로부터 얻을 수 있는 고객의 인구통계적 데이터를 분석하여 고객을 그룹핑(grouping) 한다(<그림 3.1.1>의 분석2단계 참조). 이것은 고객의 인구통계적 특성에 따라 접속 패턴이나 매출 형태가 크게 달라질 수 있으므로 인구통계적 데이터를 고객 특성 분석에 사용하였고, 모든 고객의 고유한 특성에 맞는 서비스



<그림 3.1.1> 효과적인 e-Business를 위한 웹상의 고객 특성 분석 프레임워크

를 제공하는 것이 실질적으로 불가능하기 때문에 고객을 몇 개의 그룹으로 그룹핑하고 각 그룹의 특성에 맞는 서비스를 제공하기 위함이다.

매출 DB에는 고객이 전자상거래를 통해 상품을 판매 혹은 구매한 정보가 저장되는데, 이것은 실제로 기업간의 거래가 체결되어 거래에 대한 모든 정보가 저장되므로 고객 분류 (customer segmentation)에 중요한 자료가 된다. 본 연구에서 제안한 Data Visualization System은 매출 데이터에 대한 SOM 분석을 통해 고객의 판매 및 구매 패턴 정보를 제공한다. 분석된 고객 특성, 즉, 고객의 접속 행태와 판매 및 구매 패턴을 기준으로 Data Visualization System은 고객을 분류한다(<그림 3.1.1>의 분석 2단계 참조).

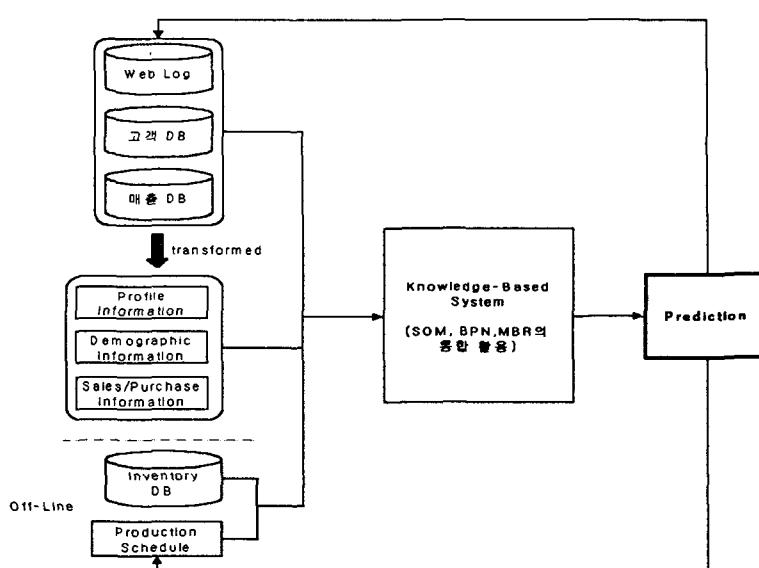
또한 Data Visualization System은 매출 DB와 상품 DB를 결합한 자료로부터 Dependency Model을 도출한다(<그림 3.1.1>의 분석 2단계 참조). 즉, 특정 상품의 구입이 특정 기간 후에 다른 상품의 구입에 영향을 주는 것이 존재하는데,

예를 들면, A라는 상품을 구매한 고객이 3개월 후에 B라는 상품을 구매할 확률이 높다면 A와 B 사이에는 dependency가 있다고 한다. 이러한 관계를 상품 군에 대해 규명한 Dependency Model을 도출하는 것이다.

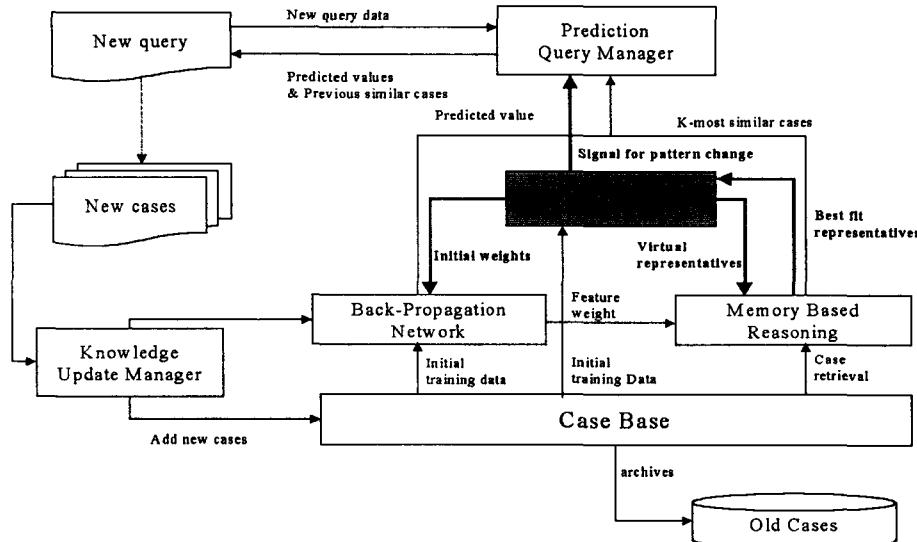
이와 같이 분석된 고객 분류 정보와 상품에 대한 Dependency Model을 바탕으로 각각의 고객 특성에 맞게 웹 화면을 재설계하고 웹Content를 재배치하는 것이 본 연구의 마지막 절차이다. 고객의 특성에 맞게 재구성한 웹 화면/Content는 고객에게 제공한 후 고객의 행태에 대한 정기적인 분석을 실시하여 행태 변화를 지속적으로 평가하면서 고객의 요구에 맞는 화면으로 지속적으로 변경하여 가는 것이다.

### 3.2. SOM, BPN, MBR의 통합활용에 기반한 B2B EC상의 구매 판매 예측 시스템

3.2장에서는 기업간 전자상거래의 고객 기업에



<그림 3.2.1> 기업간 전자상거래에서의 구매 판매 예측시스템의 프레임워크



<그림 3.2.2> SOM, BPN, MBR이 통합 활용된 구매 판매 예측시스템내의 지식기반시스템

제 그 기업의 구매 및 판매에 대한 예측 정보를 제공할 수 있는 구매 판매 예측 시스템의 프레임 워크를 제안한다(<그림 3.2.1> 참조).

<그림 3.2.1>에서 보듯이, 구매 판매 예측 시스템은 기업간 전자상거래에서 고객 기업의 구매 및 판매 정보, 접속 행태 그리고 각 기업의 재고현황, 생산 일정과 같은 데이터를 이용하여 고객에게 구매 및 판매에 대한 예측 정보를 제공 할 수 있다. 즉, 전자상거래 고객 중 A라는 기업 고객의 구매 및 판매 데이터, 인구통계 및 프로파일 데이터를 이용하여 미래의 판매 변동 및 안정도 등에 대한 예측을 지원해 주게 된다. 기업 고객 A가 오프라인 상으로 재고 현황 자료와 생산 일정과 같은 데이터를 제공한다면 추가적으로 원자재의 공급 및 판매 전략에 대한 예측이 가능하다.

<그림 3.2.1>에서 보듯이, 구매 판매 예측시스

템에서는 Shin et al. (1999)이 제안한 MBR & BPN 혼합 전략에 추가적으로 SOM을 활용하는 지식기반시스템을 구축한다. 지식기반시스템은 학습된 BPN으로부터 특성치의 최적 가중치 (feature weight)를 가지는 MBR을 수행하고, 또한 SOM을 활용하여 동적(Dynamic)인 외부환경에 순응적으로 적응하며 구매/판매의 예측 정확도를 향상시킨다.

본 연구에서 제안하는 지식기반시스템의 구조는 <그림 3.2.2>와 같다.

<그림 3.2.2>에서 보듯이, 먼저 SOM을 사용하여 입력 자료의 패턴분석을 수행한 후 기존의 지식 기반을 수정하여야 하는지를 결정한다. 이것은 구매/판매의 예측의 정확도를 향상할 수 있는 더 좋은 지식 기반을 만들기 위한 것으로, 적용 영역(application domain)에서 외부 환경이나 내부 프로세스의 변화로 인한 특성치들의 변화가

발생할 경우 지식기반시스템의 기준의 지식기반은 더 정밀한 예측을 위하여 정제(refine)되어야 하고 SOM이 이러한 목적에 사용되는 것이다.

SOM에 의한 패턴 분석을 통해 새로운 자료의 패턴 변화 여부를 판정할 수 있는데, 변화가 있다고 판정되면 SOM은 BPN과 MBR의 예측력(correct prediction ratio, CPR)이 떨어질 것이라고 경고하면서 동시에 BPN이 새로운 자료와 함께 재학습되어야 하고 MBR에 대해서는 Case Base로부터 오래된 자료를 삭제해야 한다고 알려준다. 이와 같이, 구매 판매 예측시스템에서 지식기반시스템의 예측 정확도를 유지하기 위해 입력 패턴의 변화를 지속적으로 모니터링하는 SOM의 역할은 아주 중요하다.

또한 SOM을 사전에 활용하여 특성 지도를 분석함으로써 BPN의 학습 부담을 줄여주고, MBR에게는 원 자료의 의미있는 특성치(feature value)들인 가상적인 Case를 제공한다. 즉, BPN은 적절한 네트워크가 얻어질 때까지 탐험적인 실험을 반복적으로 수행하여야 하는데, 학습할 때 네트워크의 초기 가중치(initial weight)는 0에 가까운 임의의 값으로 설정되기 때문에 지역적인 최소화(local minimum)를 피하기 위해서는 더 많은 실험이 요구된다. 이러한 때 SOM을 통해 입력 패턴의 특성 지도를 분석함으로써 BPN의 초기 가중치와 한계 값(threshold value)을 결정하는 데 도움을 줄 수 있고, 입력 패턴의 특성에 의해 나누어진 부분적 자료의 학습이 모든 자료를 한꺼번에 학습시키는 것보다 더 좋은 지식을 얻을 수 있는 것인지를 알 수 있다.

지식기반시스템에서의 MBR과 BPN은 새로운 Case가 발생할 때마다 그들의 지식을 쉽게 수정 할 수 있다. BPN은 많은 자료가 새로이 도입될 때 가중치 집합을 갱신할 수 있고 MBR은 온라

인 학습(on-line learning)이 가능하다. SOM이 새로운 자료의 특성이 과거의 자료 특성과 다르다고 경고할 때 사람이 Case Base에서 오래된 자료를 제거하고 새로운 자료로 Case Base를 갱신할 수 있다.

본 연구에서 제시한 지식기반시스템에서 구매/판매의 예측 결과의 통합은 예측 질의 관리자(prediction query manager, PQM)에 의해 수행된다. 즉, PQM은 새로운 자료를 받아서 BPN과 MBR에 동시에 의뢰를 한다. BPN과 MBR의 두 예측치가 서로 일치하는 경우에는 PQM이 예측된 값을 제시해 주지만 그 결과가 서로 많이 차이 나는 경우에는 PQM은 가장 비슷한 과거의 Case에 대한 결과를 제시해 주고 의심스러운 결과라고 응답해 준다. 두 방법의 불일치가 자주 발생하는 경우, PQM은 위험 신호를 보내는데 이 위험 신호는 지식 기반이 자료의 전체 내용을 잘 반영하지 못한다는 것을 의미하는 것으로 지식기반시스템에 잘못이 있다는 것을 나타낸다.

지식기반시스템에서 지식 갱신 관리자(knowledge update manager, KUP)는 지식의 정제를 위해 새로운 자료를 제공하는 역할을 한다. MBR의 경우 오래된 Case들은 Case Base에서 단지 불필요한 저장 공간만을 차지하고 비슷한 Case를 찾는 시간을 지연시킬 뿐만 아니라 동적인 변화가 심한 환경하에서 MBR이 적절한 예측을 하는 것을 어렵게 만든다. 또한 BPN의 오래된 지식은 변화된 환경하에서의 새로운 Case에는 적절하지 않을 가능성이 많다. 이러한 문제점의 해결을 위해, SOM이 어떠한 변화의 시그널을 잡아낼 때마다 KUP는 BPN의 재학습과 MBR의 Case Base 수정을 위한 새로운 Case를 제공하는 것이다.

## 4. 결론

전자상거래 서비스를 제공하는 웹 사이트에서 고객의 접속에 의해 발생되는 데이터는 무수히 많다. 이러한 데이터에는 고객의 접속 행태, 고객의 관심 분야와 같은 비즈니스 선점을 위한 중요 정보를 포함하고 있지만 이러한 정보를 추출하기 위한 분석 방법론이 충분히 연구되지 않았다. 본 논문에서는 웹상의 많은 데이터로부터 고객의 특성을 파악하기 위한 SOM에 근거한 Data Visualization System과 다수의 인공지능 도구를 통합 적용하여 고객이 원하는 구매 및 판매 정보를 추출할 수 있는 예측시스템을 제시하였다.

Data Visualization System은 Web Log로부터 다차원적인 분석을 통해 고객의 접속행태와 고객이 관심을 가지는 상품 품목 등에 관한 정보를 제공한다. 또한 고객 정보 데이터베이스와 매출 정보 데이터베이스로부터 B2B 전자상거래에서 고객 기업의 특성, 구매/판매 행태를 파악할 수 있다. 파악된 고객 기업의 특성을 기반으로 고객 기업을 분류하고 특성에 맞는 다양한 정보 제공, web 화면 디자인, web contents 재구성을 실시함으로써 고객 기업의 만족도를 높이게 되고 따라서 충성도(loyalty) 향상에 중요한 역할을 할 수 있다.

일반적으로 중소/중견 기업이 특정 시장에 대한 현황을 분석하여 체계적으로 구매/판매 전략을 수립하는 것은 소요 비용과 인력을 고려할 때 거의 불가능한 일로서, 전자상거래 업체에서 고객 기업의 판매 및 구매 전략을 수립할 수 있는 지식기반시스템을 구축하여 고객 기업에게 필요한 전략을 제공하는 것은 전자상거래의 고객 충성도와 고객 유지 관점에서 매우 중요한 일이다.

지식기반시스템을 개발할 때, 좋은 지식기반을

구축하는 것은 새로운 Case에 대한 정확한 예측을 하기 위해서 가장 중요하다. 일단 지식기반이 개발되면, 그것의 유지가 핵심이 되고 대부분의 유지 작업은 지식을 정제(refine)하는 능력을 가지고도록 하는 것이다. 지식기반시스템이 개발된 임의의 특정분야에서 발생하는 지식의 변화와 지식기반의 진화사이에 시간 차이가 존재하는데 이것은 대부분의 지식기반시스템이 환경이나 지식의 변화에 사전 warning 정보를 제공하지 못하기 때문이다. 즉, 일반적인 지식기반시스템은 새로운 Case의 입력패턴으로부터 나오는 신호를 해석할 수 있는 능력을 가지지 못하기 때문에 부정확한 예측 결과만을 제시한다. 그러나 본 연구에서 제안한 구매 판매 예측시스템은 그 자체로서 외부 환경 변화에 적응할 수 있다. 이것은 변화에 대한 시간 차이를 거의 0에 가깝게 줄여주는 것으로 변화가 일어날 때마다 제안한 시스템은 빨리 경고를 주고, 지식 정제를 위해 학습을 시작한다.

본 논문에서 제안한 구매 판매 예측시스템은 BPN 신경회로망으로부터 가중치 절삭 알고리즘과 함께 가중치 문제에 대해 혼합 접근 방법을 사용함으로써 MBR의 특성치에 대한 가중치 결정의 약점을 보완하였다. 또한 제안된 시스템은 BPN이 최적의 네트워크를 구성하는데 학습시간이 많이 소요되는 것과 초기 가중치와 한계 값이 임의의 값으로 설정되기 때문에 로컬 미니멈에 빠질 가능성을 가지고 있는 점을 보완하였다. 즉, 자료로부터 BPN의 학습 부담을 줄여주기 위해서 SOM분석을 통해 얻을 수 있는 학습자료의 특성을 이용하였다.

## 5. 부록

### 5.1 SOM

본 논문에서 제시한 SOM은 여러 종류의 자기 조직화(self-organizing) 신경회로망중에서 SOFM (self-organizing feature map)이라고 알려진 신경 회로망으로 이것은 경쟁 학습(competitive learning)을 기초로 한다. 즉, SOM네트워크의 출력 뉴런이 활성화되기 위하여 서로 경쟁한 후 어떤 시점에서 단지 하나의 출력 뉴런만을 활성화 시킨다. 경쟁에서 이긴 출력 뉴런을 승자획득뉴런 (winner-takes-all neuron)이라고 한다. 출력 뉴런들 중에서 승자획득뉴런을 유도하는 한가지 방법은 출력 뉴런 사이에 측면억제연결(lateral inhibitory connection)을 사용하는 것이다.

SOM에서는 뉴런이 N-차원으로 구성된 격자(lattice)의 노드에 위치한다. 뉴런은 경쟁적 학습 과정에 있는 입력 패턴이나 입력 패턴의 출력 클래스에 따라서 선택적으로 조정(tune)된다. 조정된 뉴런의 위치는 다수의 입력 패턴의 특징들에

대해서 의미 있는 좌표체계가 격자(lattice)에 생성되는 방법으로 상호 연관되어 순서를 이루게 되는 경향이 있다: 따라서 격자에 있는 뉴런의 위치는 입력 패턴의 특징에 대응한다. 즉, SOM은 격자에 있는 뉴런의 공간의 위치가 입력 패턴의 고유의 특성과 일치하도록 입력 패턴의 지형적인 지도(map)를 형성한다.

SOM의 알고리즘은 <그림 5.1.1>과 요약할 수 있다.

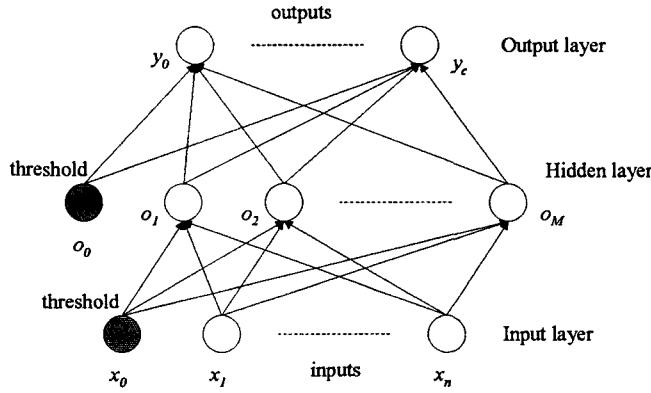
### 5.2 학습된 BPN을 이용하는 MBR의 Feature Weight Algorithms

본 논문의 구매 판매 예측시스템에서 활용한 방법은 학습된 신경 회로망으로부터 기호적인 규칙(symbolic rule)을 추출하려는 것(Towell, 1994, Craven, 1997, Lu, 1996)이 아니고, MBR의 가중치들의 벡터  $\{w_1, w_2, \dots, w_n\}$ 를 얻으려고 하는 것이다 (여기서  $n$ 은 입력 속성의 개수이다). <그림 5.2.1>에서 보듯이 하나의 은닉 층(Hidden Layer)을 가지는 완전히 연결된 네트워크(fully

#### Summary of SOM Algorithm

1. *Initialization.* 연결 강도 벡터 (weight vector)를 초기화 한다.
  2. *Sampling.* 확률을 가지는 입력 분포로부터 샘플  $x$ 를 뽑는다
  3. *Similarity Matching.* 시점  $n$ 에서 유플리디언 거리가 최소인 획득(winning) 뉴런을 찾는다
- $$i(x) = \arg \min_j \|x(n) - w_j\|, \quad j = 1, 2, \dots, N$$
4. *Updating.* 모든 뉴런의 연결 강도 벡터를 갱신 한다.
- $$w_j(n+1) = \begin{cases} w_j(n) + \eta(n)[x(n) - w_j(n)], & j \in \Lambda_{i(x)}(n) \\ w_j(n), & \text{otherwise} \end{cases}$$
5. *Continuation.* 특성 지도 (feature map)에서 주목 할 만한 변화가 관찰되지 않을 때까지 단계 2로 계속 한다.

<그림 5.1.1> SOM 알고리즘의 요약



&lt;그림 5.2.1&gt; 하나의 은닉 층을 가지는 완전히 연결된 네트워크의 예

connected network)를 고려해 보자.

이 네트워크는  $n$ 개의 입력,  $M$ 개의 은닉 유닛과  $m$ 개의 출력 유닛을 가진다. 설명의 복잡성을 막기 위해서 네트워크의 형태는 출력 값이 하나이고 단지 하나만의 은닉 층을 가지고 완전히 연결된 네트워크 구조를 가진다고 가정한다. 은닉 층에서  $j$ 번째 유닛의 입력은 먼저  $n$ 개의 입력 값들의 선형 가중 조합을 형성하고 거기에 한계 값(threshold value)을 추가함으로써 얻어진다. 그식은 다음과 같다.

$$o_j^{in} = \sum_{i=0}^n w_{ji}^{(1)} x_i \quad (1)$$

$$y^{out} = \sum_{j=0}^M w_j^{(2)} o_j^{out} \quad (2)$$

여기서  $w_{ji}^{(1)}$ 은 입력  $i$ 에서 은닉 유닛  $j$ 로 가는 첫 번째 계층(layer)의 가중치를 나타내고,  $w_j^{(2)}$ 는 두 번째 계층을 각각 나타낸다.

<그림 5.2.1>에서 나타나듯이, 은닉과 출력 유닛에 대한 한계 값들에 대해서는 추가적인 입력

변수  $x_0 = 1$  와 추가적인 은닉 변수  $o_0 = 1$ 를 포함시킴으로써 처리된다. 은닉 유닛  $o_j$  와 출력 유닛  $y$ 의 활성화(activation)는 활성화 함수를 사용하여 식 (1)과 (2)에서 선형 합계를 변형함으로써 얻어진다. 일반적인 활성화 함수(activation function)는 시그모이드 함수 ( $y = 1 / (1+e^{-x})$ )이다. 이때, <그림 5.2.1>에서 출력 함수에 대한 표현을 다음의 형태로 얻을 수 있다.

$$\begin{aligned} o_j^{out} &= \sigma(o_j^{in}) \\ y^{out} &= \phi\left(\sum_{j=0}^M w_j^{(2)} o_j^{out}\right) \end{aligned} \quad (3)$$

식(3)에서 출력 유닛의 활성화 함수를  $\psi(\cdot)$ 의 형태로 나타낸 것은 은닉 유닛에서 사용한 함수와 같은 함수를 사용할 필요는 없다는 것을 강조하기 위해서이다(예, hyperbolic tangent function).

일반적으로 강하게 연결된 가중치를 가지는 입력노드가 중요한 변수로 간주된다. 가중치 집합에서 유용한 어떤 것을 찾아내는 연구에 대한 아이디어는 처음에는 불필요한 가중치를 없애는

것을 목적으로 하는 가중치 절삭(weight pruning)과 가중치 소멸(weight decaying)로부터 나왔다. 가중치 절삭은 네트워크의 예측 능력을 유지하면서 학습된 네트워크를 단순화 시키는 작업이다. 신경 회로망을 사용하여 가중치 절삭이나 특성치 선택을 하는 방법에 관한 자세한 기술은 (Reed 1993)와 (Setiono 1997)에서 발견된다. 본 연구에서는 특성치에 대한 가중치 결정 메커니즘으로 다음 네가지를 활용할 것을 제안한다: Sensitivity, Activity, Relevance and Saliency.

#### A. 민감도 (Sensitivity)

입력 노드의 민감도는 학습된 네트워크로부터 입력노드를 제거하면서 계산된다. 입력노드에 연결된 모든 가중치가 0이 된다면 그 입력노드는 제거될 수 있다. 입력 변수들의 민감도의 측정은 어떤 특성 가중치(feature weight)가 제거되었을 때의 오차와 계속 남아 있을 때의 오차간에 차이이다. 입력 변수의 민감도  $S_i$ 는 다음과 같다.

$$S_i = \text{Max}\left(\frac{E(0) - E(w^f)}{E(w^f)}, 0\right) \quad (4)$$

여기서  $w^f$ 는 학습 후 입력 노드  $i$ 에 연결된 가중치의 최종 값이고,  $E(w^f)$ 는 그때의 오차 값이고  $E(0)$ 는 그 노드가 제거되었을 때의 오차 값이다. 음수 값의 sensitivity는 입력 속성이 결과 예측에 전혀 기여를 못한다는 것을 의미하기 때문에 식 (4)에서 최대값 함수가 필요하다. 민감도를 측정하기 위하여 사용되는 오차는 다음과 같다.

$$E = \sum_L |t_p - o_p| \quad (5)$$

여기서  $L$ 은 학습자료의 집합이고,  $t_p$ 와  $o_p$ 는 각

각 학습자료의 타겟 값과 네트워크의 출력 값이다. 식(5)은 오차가 작을 때 적절성(relevance)의 좋은 예측치를 제공해 준다(Reed 1993). 마지막으로 입력  $i$ 에 관한 가중치 값은 다음과 같이 계산된다.

여기서  $n$ 은 입력 속성의 개수이다. 다시 각 가중치 값은 표준화된 sensitivity의 형태로 나타내어 진다(즉,  $\|w\| = 1$ ).

#### B. 활동성 (Activity)

한 노드의 활동성은 학습 자료에 대한 활성화 정도의 분산으로 측정된다. 노드의 활동성 값이 입력 값에 따라서 크게 변화할 때, 노드의 활동성이 높다. 반대로 노드의 활동성 값이 학습 자료에 대해 일정하게 유지되면, 그 노드의 활동성은 0이다. 그러므로 은닉 노드  $j$ 의 활동성은 수식 (7)로 나타내지고, 입력노드  $i$ 의 활동성은 수식 (8)로 나타내어 진다.

여기서  $\text{var}(\cdot)$ 은 분산 함수이다. 가중치는 각각의 입력들의 활동성을 표준화함으로써 계산된다.

$$A_j^{\text{hidden}} = (w_j^{(2)})^2 \cdot \text{var}(\sigma(\sum w_{ji}^{(1)} x_i)) \quad (7)$$

$$A_i^{\text{input}} = \text{var}(x_i) \cdot \sum_{j=1}^M ((w_{ji}^{(1)})^2 \cdot A_j^{\text{hidden}}) \quad (8)$$

#### C. 적절성 (Relevance)

가중치의 분산이 노드의 적절성에 좋은 예측치가 되고 또 노드의 가장 큰 가중치가 삭제되었을 때 노드의 적절성이 예측된 RMS (root mean square) 오차의 증가에 대한 좋은 추정치가 된다는 연구결과가 있다(Segeee and Carter 1991). 이

러한 개념을 따라서 은닉 노드  $j$ 의 적절성은 수식 (9)의 형태가 되고 입력 노드  $i$ 에 대한 전체적인 적절성은 수식 (10)의 형태로 나타난다.

$$R_j^{hidden} = (w_j^{(2)})^2 \cdot \text{var}(w_{ji}^{(1)}) \quad (9)$$

$$R_i^{input} = \sum_{j=1}^M ((w_{ji}^{(1)})^2 \cdot R_j^{hidden}) \quad (10)$$

과거의 경우와 같이 가중치는 각각의 입력노드에 대한 표준화된 적절성에 의해 계산된다.

#### D. 돌출성 (Saliency)

L. Cun et al. (1990)은 가중치의 돌출성을 가중치와 관계된 오차의 이차 도함수를 추정하여 측정하였다. 그들은 네트워크의 연결을 반복적으로 절삭하는데 돌출성 측정을 사용하였다: 즉, 적당한 오차범위 내까지 학습을 시키고서, 돌출성을 계산하고 낮은 돌출성 가중치를 삭제하고 학습을 계속시킨다. 가중치에 대한 돌출성 측정 값은 가중치의 제곱에 비례한다. 우리는 그들의 결과를 변형하여 입력노드의 돌출성을 측정하였다. 입력노드의 돌출성 계산식을 수식 (11)과 같다.

$$Saliency_i = \sum_{j=1}^M ((w_{ji}^{(1)})^2 \cdot (w_j^{(2)})^2) \quad (11)$$

### 6. 참고문헌

1. Aha, D.W., Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, *International Journal of Man-Machine Studies*, Vol. 36 (1992), 267-287.
2. Alonso, G., Fiedler, U., Hagen, C., Lazzcano, A., Schultdt, H., Weiler, N., "WISE: Business-to-Business E-Commerce", *IEEE SMC '99 Conference Proceedings*, Vol.3 (1999), 1054-1059.
3. Benitez, J. M., Castro, J. L., Requena, I., "Are Artificial Neural Networks Black Boxes?", *IEEE Transactions On Neural Networks*, Vol. 8 (1997), 1156-1164.
4. Craven, M. W., Shavlik, J. W., "Using Neural Networks for Data Mining", *Submitted to the Future Generation Computer Systems special Issue on Data Mining*, (1997), 101-112.
5. Cun, Y. L., Denker, J. S., Solla, S. A., *Optimal brain damage*, In *Advances in Neural Information Processing(2)*, D.S.Touretzky, Ed., 1989.
6. Dussart, C., "Transformative Power of e-Business Over Consumer Brands", *European management journal*, Vol. 19 No. 6 (2001), 629-637
7. Irani, K. B., Cheng, J., Fayyad, U. M., Quan, Z., "Applying Machine Learning to Semiconductor Manufacturing", *IEEE Expert*, Vol.8 No. 1 (1993), 41-47.
8. Kang, B. S., Lee, J. H., Shin, C. K., Yu, S. J., Park, S. C., "Hybrid Machine Learning System For Integrated Yield Management in Semiconductor Manufacturing", *Expert Systems With Applications*, Vol.15 (1998), 123-132.
9. Kohonen, T., "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, Vol. 43 (1982), 59-69.
10. Lu, H., Setiono, R., "Effective Data Mining Using Neural Networks", *IEEE Transactions On Knowledge and Data Engineering*, Vol. 8 (1996), 957-961.
11. Reed, R., "Pruning Algorithms - A Survey", *IEEE Transaction On Neural Networks*, Vol. 4 (1993), 740-747.
12. Segee, B. E., Carter, M. J., "Fault tolerance of pruned multilayer networks", *Proceedings of International Joint Conference on Neural*

- Networks*, Vol. II (1991), 447-452.
- 13. Setiono, R., Liu, H., "Neural-Network Feature Selector", *IEEE Transaction on Neural Networks*, Vol.8 (1997), 654-662.
  - 14. Shin, C. K., Park, S. C., "Memory and Neural Network Based Expert System", *Expert System with Applications*, Vol. 16 (1999), 145-155.
  - 15. Towell, G., Shavlik, W., *Refining Symbolic Knowledge Using Neural Networks, Machine Learning: A Multistrategy Approach*, Morgan Kaufmann, CA., 1994.
  - 16. Yu. S. J., Lee. J. H., Park. S. C., "Forecasting in A Complex Environment Using Feature Manipulating Technique Added In Traditional Forecasting System", *IEEE IEMC '98*, Puerto Rico (1998), 291-294.
  - 17. Yu. S. J., Park. S. C., "Applying Machine Learning to the Analysis Of a Quality Survey in a Mobile Telecommunication Industry", *KAIST Working Paper* (1999).
  - 18. Zhong Tian, Liu, L.Y., Jing Li, Jen-Yao Chung, "Business-to-Business e-Commerce with Open Buying on the Internet", *International Conference on WECWIS* (1999), 56-62.

Abstract

## Web Mining for successful e-Business based on Artificial Intelligence Techniques

Jang Hee Lee\* · Sung Jin Yu\*\* · Sang Chan Park\*\*

Web mining is an emerging science of applying modern data mining technologies to the problem of extracting valid, comprehensible, and actionable information from large databases of web in e-Business environment and of using it to make crucial e-Business decisions. In this paper, we present the noble framework of data visualization system based on web mining for analyzing the characteristics of on-line customers in e-Business. We also propose the framework of forecasting system for providing the forecasting information of sales/purchase through the use of web mining based on artificial intelligence techniques such as back-propagation network, memory-based reasoning, and self-organizing map.

**Key words:** e-Commerce, e-Business, web mining, artificial intelligence, Data Visualization System

---

\* School of Industrial management, Korea University of Technology and Education  
\*\* Department of Industrial Engineering, KAIST