하이브리드 SOM을 이용한 효율적인 지식 베이스 관리

윤 경 배[†] · 최 준 혁[††] · 왕 창 종[†††]

요 약

정보 기술 분야의 지능화 요구는 매우 빠르게 증가하고 있다. 특히 대량의 데이터로부터 지식을 찾아내어 최적의 의사결정을 해야하는 KDD(Knowledge Discovery in Database)분야에서는 그 요구가 더욱 더 크게 된다. 자능화된 의사결정을 위해서는 대용량 지식 베이스(Knowledge Base)의 효율적인 관리가 무엇보다도 중요하다. 본 논문에서는 이러한 지식 베이스로부터 의사결정 관리에 필요한 지식을 얻기 위해 효율적으로 지식 베이스를 검색하고 갱신하는 관리 방법을 위해 자율학습 신경망인 자기조직화 지도에 확률적 분포 이론을 결합한 하이브리드(Hybrid) SOM을 제안한다. 제안 방법을 이용한 효율적 지식 베이스의 관리를 시뮬레이션 실험을 통하여 수행하였다. 실험을 통해 본 논문에서 제안하는 Hybrid SOM이 지식 베이스 관리에 효율적인 성능을 나타냄이 증명되었다.

An Efficient Knowledge Base Management Using Hybrid SOM

Kyungbae Yoon[†] · Junhyeog Choi[††] · Changjong Wang[†††]

ABSTRACT

There is a rapidly growing demand for the intellectualization of information technology. Especially, in the area of KDD (Knowledge Discovery in Database) which should make an optimal decision of finding knowledge from a large amount of data, the demand is enormous. A large volume of Knowledge Base should be efficiently managed for a more intellectual choice. This study is proposing a Hybrid SOM for an efficient search and renewal of knowledge base, which combines a self-study nerve network, Self-Organization Map with a probable distribution theory in order to get knowledge needed for decision-making management from the Knowledge Base. The efficient knowledge base management through this proposed method is carried out by a stimulation test. This test confirmed that the proposed Hybrid SOM can manage with efficiency Knowledge Base.

키워드 : SOM, 하이브리드 SOM(Hybrid SOM), 데이터 마이닝(Data Mining), 지식 베이스 관리(knowledge base management), GTM(Generative Topographic), 이산분포(Statistical Discrete Distribution)

1. 서 론

지능형 지식 관리 시스템은 1990년대 이후 IT 시스템 분야에서 해결해야 할 가장 시급한 문제중의 하나이며, 특히 이러한 지능화 전략은 네트워크, 운영체제 등을 비롯한 대부분의 정보기술 분야에 적용되어져야 한다는 필요성이 대두되고 있다[4]. 본 논문에서는 이러한 지식 베이스(knowledge base)의 효율적 관리를 위하여 Kohonen이 제안한 자기조직화 지도(SOM ; Self Organizing Maps)를 확률적 모델(probabilistic model)로 변환시킨 Hybrid SOM을 제안하고, 이를 바탕으로 지식 베이스의 효율적 관리 방안을 연구한다. 특히 크고 작은 프로젝트의 관리를 위하여 프로젝트 사례에 기반한 지식 베이스를 구축하고 이를 효율적으로 관리하는 방안에 대해서 연구한다. 프로젝트의 효율적 관리는 프로젝트의 과정을 계획하고, 향후 발생할 수 있는 문제점을 예측하여 예상되는 해결책에 대해 과거의 경험을 얼마나 활용하느냐에 달려있다. 그런데 보통의 경우 새로운 프로젝트가 발생하면 기존의 유사한 프로젝트에 대한 산출물들을 추출하는데 많은 제약이 따른다. 따라서 새로운 프로젝트를 수행하는데 있어 과거의 경험을 토대로 현재의 프로젝트를 관리, 예측하는 데에는 한계가 있다. 기존의 프로젝트 관리 시스템의 문제점은 계획 수립에 많은 시간과 비용이 투자되지만 정확도가 떨어져 대부분의 프로젝트에서 일정과 비용이 초과되고 이전의 경험이 축적되지 못하며 공통 모듈의 재사용률이 떨어진다[12]. 따라서 프로젝트에 문제가 발생하였을때 이를 해결하기 위한 유사 사례가 없어 이를 해결하는데 많은 어려움을 겪게된다. 본 논문에서는 과거 프로젝트의 경험 데이터로부터 프로젝트 계획 및 스

※ 이 논문은 2002학년도 김포대학의 연구비 지원에 의하여 연구되었음.
† 준 회 원 : 김포대학 컴퓨터계열 교수
†† 종신회원 : 김포대학 컴퓨터계열 교수
††† 정 회 원 : 인하대학교 컴퓨터공학부 교수
논문접수 : 2002년 7월 3일, 심사완료 : 2002년 8월 31일

케줄링에 필요한 기준자료(통계자료, 템플릿 등), 규모예측, 위험 관리방안 등에 대한 데이터를 추출하고 활용하여 보다 효율적인 프로젝트 관리를 위하여 Hybrid SOM을 이용한 지식베이스의 구축 및 활용에 대해 설계하고 구현하였다.

2. 하이브리드 SOM

2.1 SOM과 지식 베이스 구축

과거의 프로젝트 수행에 있어 성공한 사례 또는 실패한 사례들은 모두 새로운 프로젝트를 위한 좋은 경험치가 된다. 따라서 이전의 프로젝트 경험을 효율적으로 저장하여 새롭게 발생되는 각 프로젝트의 관리를 위한 빠르고 정확한 정보의 저장 및 추출 시스템이 필요하다. 이러한 지식 베이스는 현재의 프로젝트 수행을 위하여 과거의 유사 사례를 이용하고, 과거 프로젝트의 성공 및 실패 결과는 지식 베이스의 갱신을 위한 사례로 적용되어, 향후에 발생되는 프로젝트에 대한 정확한 수행 및 관리가 더욱 지능화 될 수 있도록 한다.

본 논문에서는 이처럼 기존에 수행된 프로젝트 결과들에 대한 지식 베이스를 구축, 갱신하여 새로이 발생하는 프로젝트에 대한 예측관리를 위하여 자율적 데이터 마이닝 기법(unsupervised data mining technique)으로 분류할 수 있는 하이브리드 SOM을 제안하고 적용한다.

인간의 뇌 구조를 가장 잘 모형화한 자기조직화 지도(Self-Organizing feature Maps : SOM)는 1980년대 초에 Kohonen에 의해 제안된 신경망 모델이다[6-8]. SOM은 신경망 중에서도 학습 자료에 대한 결과값을 모르고 학습이 수행되는 자율 학습 구조를 가지고 있으며, 음성인식, 문자인식, 구문분석 등 다양한 분야에 응용되고 있다.

SOM의 연결 강도인 가중치는 정규화된 입력 벡터에 대응되는 출력 노드의 중심값과 같은 역할을 하며 학습 동안에 입력 벡터와 가장 가까운 유클리디안 거리를 갖는 출력 노드가 승자(winner)가 되고, 이 승자 노드와 이웃하는 노드들의 가중치만이 갱신한다. 특히, SOM은 다층 신경망(MLP)과 같은 지도 학습 모형에 비해서 매우 단순한 2개의 층으로 이루어지면서 다차원의 자료를 2차원의 형상 지도(feature maps)로 투영시켜 스스로 경쟁 학습을 할수 있도록 한다. 이때, 2개의 층은 입력 벡터를 갖는 입력층과 형상지도를 갖는 출력층이다. SOM은 오류 역전파의 퍼셉트론(perceptron) 모형과는 달리 여러 단계의 피드백을 거치지 않고 오직 한 번의 전방 전달만을 사용하며, 입력층에서 출력층으로는 모두가 연결되어 있는 구조를 갖고 있다. 자기 조직화 지도에서 초기 뉴런의 연결강도는 0에서 1사이의 균일 분포(uniform distribution)에서 생성된 임의의 난수 값으로 초기화되며, 이때 입력 벡터의 값들은 0에서 1사이의 값을 갖도록 정규화된다. 초기화 이후의 학습은 출력층의 각 노드의 가중치와 대응되는 입력 벡터의 각 값들 간의 유사도를 측정하기 위한 측도로서 거리(distance)를 사용한다.

본 논문에서는 거리 측도로서 유클리디안 거리를 사용하며, 최종적으로 주어진 입력 벡터와의 거리가 가장 작은 출력 노드가 승리하게 되고 이 노드만이 출력을 하게 된다. 가중치 갱신 학습은 승자 노드와 이에 이웃하는 노드들도 가능하다. SOM의 학습 규칙은 승자 독점(winner take all)으로서 모든 출력과 갱신이 승자 노드를 중심으로 이루어진다. 승자 노드의 이웃 영역은 아주 가까운 노드로만 제한하지 않고 학습의 초기에는 영역의 범위(neighborhood size)를 층내의 모든 뉴런으로 확장한 후 학습이 진행되는 동안 점차로 줄여 나가는 방법을 사용한다. 최종적으로는 단지 승자 노드의 가중치만이 갱신된다. 승자 노드를 중심으로 한 가중치의 갱신은 식 (1)의 규칙을 따른다[6].

$$W_j(k+1) = W_j(k) + \eta(k)\,(X(k) - W_j(k)) \qquad (1)$$

식 (1)은 k + 1단계에서 새롭게 갱신되는 가중치는 현재의 k단계의 가중치에 학습율과 현재의 입력 자료와의 차이를 반영하는 식에 의해 나타남을 보여 주고 있다.

2.2 하이브리드 Self Organizing Maps 알고리즘

Kohonen이 제안한 SOM 모형은 다층 신경망과 같은 다른 신경망 모형과 마찬가지로 블랙박스와 같은 가중치 갱신 결과를 얻기 때문에 모형에 대한 설명력이 부족하고 학습이 반복되는 동안 전역적 최적값(gobal optima)으로 수렴하지 못하고 지역적 최적값(local optima)으로 빠지는 경우가 종종 있다[1]. 따라서 본 논문에서는 이러한 SOM의 문제점을 해결하여 효율적인 지식 베이스의 관리에 적용하기 위하여 Hybrid SOM 모형을 제안한다. 이 모형은 기존의 SOM 모형의 초기 가중치 결정과 가중치 갱신에 확률 분포(probability distribution)를 적용한다. 확률 분포란 확률 변수가 취할수 있는 값의 확률을 의미하는데 본 논문에서는 신경망의 가중치를 확률 변수로 정의하고 가중치에 대한 확률 분포를 구하게 된다. 이러한 방법은 Bishop이 제안한 GTM (Generative Topographic Mapping)과 비교할때, SOM의 확률 버전(version)이라는 유사성이 있지만 GTM은 SOM의 통계적 접근 모형을 만들어 확률 분포와 결합하였지만[1] Hybrid SOM은 Kohonen이 제안한 SOM 모형에 직접 확률적 분포를 결합하여 가중치의 갱신에 확률 규칙을 적용하여 결과적으로 신경망 모형에 대한 설명력을 부여하였다.

Step 1* ; *Initialization

1-1. Initialize the weight vector, $w_j(0)$ to have probabilistic

distribution, $N(0,1)$.

1-2. Initialize the learning rate $\alpha(0)$, $\alpha(t) \propto t^{-a}$; $0 < a. < 1$

1-3. Initialize the neighborhood function, $K(j, j^*)$, K decreases as to increase $|j-j^*|$.
where, $K(j, j^*)$: Neighborhood function.

Step 2 ; Determine the winner node

2-1. Normalization of input vector, Gaussian distribution with mean 0, variance 1.

2-2. Choose the distribution of weights, $w \sim f(\theta)$.

2-3. Choose the winner node $j^* = \arg\max y_j$ using Euclidean criteria.

Step 3 ; Update of weights

3-1. $w_j^{New} = w_j^{Old} + \alpha(j)K(j, j^*)(X - w_j^{Old})$,
where, w_j^{New} : updated weight
w_j^{Old} : current weight

3-2. Replace old distribution by current.

Repeat Step 2, Step 3 Until given criteria satisfaction.

(알고리즘) Hybrid SOM algorithm

2.3 하이브리드 SOM을 이용한 지식 베이스 관리

과거의 프로젝트 경험을 통하여 새로운 프로젝트에 대한 적용 규칙을 찾아내기 위해서는 프로젝트 사례들에 대한 지식 베이스를 구축하여 사례 기반 추론(Case Base Reasoning : CBR)을 해야 한다[3, 5].

새로운 프로젝트에 대한 문제를 해결하기 위해서는 과거의 프로젝트와 유사한 사례들을 탐색하여 이에 대한 데이터를 적용하게 된다. 즉, 새로운 프로젝트를 위하여 구축된 지식 베이스로부터 유사한 프로젝트 사례를 검색(retrieve)하여 새로운 프로젝트에 대한 문제를 해결하고(reuse), 해결된 프로젝트는 지식 베이스에 추가되어 기존의 지식 베이스를 갱신하며(revise), 마지막으로 전체 지식 베이스를 관리하여(retain) 새로운 프로젝트의 출현에 적절하게 대처하게 한다[2].

성능이 우수한 지식 베이스를 구축하기 위해서는 프로젝트 사례들을 데이터 베이스에 구축하는 적절한 방안과 유사한 프로젝트 사례들을 데이터 베이스에서 검색하는 방안, 새로운 프로젝트에 과거의 유사한 사례들을 적용하는 방안, 최초의 프로젝트 사례를 얻는 방안 등을 고려해야 한다.

3. 시뮬레이션 지식 베이스

프로젝트 관리란 프로젝트를 수행하는데 있어 이해관계자(또는 발주자)의 요구나 기대에 부응하기 위해 행하는 프로젝트 활동에 대한 지식기술, 도구, 테크닉의 응용을 의미한다.

프로젝트 관리를 정의하기 위하여 프로젝트는 일정한 단위 기간 동안 주어진 목표를 수행하기 위한 작업들의 모임이라고 정의할 수 있다. 이는 뚜렷한 목적물, 한정된 기간, 최소의 비용, 각종 제한된 자원(인력, 장비, 자재 등)들을 동원하여 완수하고자 하는 일련의 행위 집합을 의미한다[11]. 즉, 프로젝트 관리는 프로젝트를 성공적으로 완성하기 위해 필요한 인원, 자원, 시스템, 기술의 총합으로, 프로젝트 관리의 목적은 예정된 완료일 이전에 예산 범위 내에서 기존의 자원을 효율적으로 활용하여 프로젝트를 완료하는 것이다.

만약, 프로젝트 관리가 잘못된다면 일정지연, 비용초과, 인력부족, 노력의 낭비, 기능 명세의 비효율적 이용 등의 현상이 발생하며 이는 소프트웨어 위기의 원인이 될수 있다. 본 논문에서는 이러한 문제들을 해결하기 위하여 선행 프로젝트 수행 사례를 통한 관리 요소를 찾아내고자 하였다. 프로젝트를 수행해본 정보기술 분야의 종사자들을 대상으로 설문 조사한 결과 프로젝트의 특성을 27개의 변수로 분류할 수 있었으며, 이들 변수들을 이용하여 기존의 프로젝트 사례들을 유형별로 군집화한 후, 이를 바탕으로 지식 베이스를 구축한다.

일반적인 지식 베이스에서는 과거 개개의 프로젝트를 모두 검색하여 유사한 사례들을 모아 이것들을 종합하여 새로운 프로젝트에 적용하게 되지만, 본 논문에서는 자율적 데이터마이닝 기법인 SOM에 의해 과거 프로젝트 사례들을 서로 유사한 것들끼리 군집화 하여 새로운 프로젝트에 대한 검색을 적용한다. 이러한 프로젝트 관리는 빠르고 정확하게 현재 발생된 프로젝트의 문제를 해결할 수 있다. 이를 위한 실험에서 초기의 지식 베이스의 구축을 위한 사례수집은 통계적 확률분포를 이용한 시뮬레이션을 이용하여 계산하였다[10].

프로젝트 관리를 위한 지식 베이스의 구축, 검색 및 유지보수를 위한 자율적 데이터마이닝 모델의 입력 요인들에서 프로젝트 군집화를 위한 SOM모형에 사용되는 입력 변수는 <표 1>과 같다.

실제 지식 베이스의 관리를 위하여 사용되는 <표 1>의 프로젝트 분류 변수는 사전에 정규화(normalization)를 수행시키며, 이 정규화된 변수가 SOM의 입력 변수로 사용된다.

새로운 프로젝트에 대한 최적의 설계를 위해, 기존프로젝트 결과에 의해 구축된 지식 베이스를 활용한다. 이때 초기의 지식 베이스를 구축하는 것은 중요하다. 하지만 처음으로 지식 베이스를 구축하기 위해서는 많은 사례들을 수집해야 하는데 본 논문에서는 이러한 지식 베이스 구축을 위한 사례들을 모의 실험을 통하여 만들어 내고, 이를 통해 구축된 지식 베이스를 이용하여 최적의 유사 프로젝트 유형을 찾아낸다. 이러한 작업을 수행하기 위하여 자율적 데이터마

이닝 모형을 사용하며 모형의 각 입력 변수에 대한 시뮬레이션 데이터의 생성은 통계적 이산분포(statistical discrete distribution)를 이용하며 각 범주에 대한 확률 분포를 할당하여 수행하였다. 총 27개의 변수에 대해 1,010개의 프로젝트 사례를 통계적 모의 실험을 통하여 얻었다.

〈표 1〉 프로젝트 관리 요소에 대한 입력변수 분류 및 성격

No	입력변수(분류)	세부항목(가능한 성격)
1	사업분야	공공, 제조, 금융
2	제품타입	H/W, S/W 개발, 패키지, N/W
3	업무명	인사, 회계, 급여, 생산, 구매
4	금 액	1억이하, 1억~5억, 5억~10억, 10억이상
5	투입인력	5명이하, 5명~10명, 10명~50명, 50명이상
6	기 간	1개월이하, 1개월~3개월, 3개월~6개월, 6개월 이상
7	서버 OS 수	1개, 2개, 3개 이상
8	서버 OS 종류	오라클, 사이베이스, 솔라리스
9	Client OS 수	1개, 2개, 3개 이상
10	Client OS 종류	Windows XP, Windows 2000, UNIX
11	DBMS 수	1개, 2개, 3개 이상
12	DBMS 종류	Oracle, SQL, Sybase
13	CASE Tool 수	1개, 2개, 3개 이상
14	CASE Tool 종류	SA, Designer 2000, ICONIX
15	개발도구 수	1개, 2개, 3개 이상
16	개발도구 종류	PB, VB, COBOL
17	방법론	Method/1, TRANSFORM, INOVATOR
18	프로젝트 전체 원가 진척율(투입원가/계획원가)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
19	프로젝트 단계별 원가 진척율(투입원가/계획원가)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
20	프로젝트 단계별 원가 구성비(단계별원가/전체원가)	요구사항정리, 분석, 설계, 개발, 시험, 구현
21	프로젝트 전체 일정 진척율(투입일정/계획일정)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
22	프로젝트 단계별 일정 진척율(투입일정/계획일정)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
23	프로젝트 단계별 일정 구성비(단계별원가/전체원가)	요구사항정리, 분석, 설계, 개발, 시험, 구현
24	프로젝트 전체 인력 달성율(투입인력/계획인력)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
25	프로젝트단계별 인력 달성율(투입인력/계획인력)	50% 이하, 50%~90%, 90%~110%, 110%~150%, 150% 이상
26	프로젝트단계별 인력 구성비(단계별인력/전체인력)	요구사항정의, 분석, 설계, 개발, 시험, 구현
27	직급별 인력 구성비	사원, 대리, 과장, 차/부장

4. 실험 및 결과

4.1 프로젝트 관리를 위한 지식베이스 구축 실험

27개의 프로젝트 사례 분류 변수를 기준으로 자율적 데이터마이닝 기법인 SOM 알고리즘과 확률 분포를 사용하여 프로젝트 유형의 패턴화를 수행하였다. 형상 지도의 초기 가중치는 구간[0, 1]을 갖는 균등 분포(uniform distribution)로부터 난수를 생성하여 부여하였다. 초기 학습률은 0.2로 하였으며, 학습이 진행되는 동안 학습 진행 회수에 비례하여 감소시켰다. 이는 모형이 학습되어감에 따라 갱신의 폭을 안정화시키기 위함이다. 또한 승리 노드의 이웃 반경(neighborhood size)의 크기는 이웃 함수(neighborhood function)에 의해 초기에는 전체 형상지도내의 노드를 모두 포함하지만 학습이 진행되는 동안 점차 그 크기가 줄어드는 학습 전략을 취했다. 이는 초기에는 가중치의 갱신이 전역적이고 큰 폭으로 이루어지게 학습을 유도하다가 허용 범위내의 가중치로 접근하여 갈때부터 갱신의 변경 폭을 줄여서 모형이 안정화되게 하였다.

형상 지도의 차원의 크기는 4×4, 5×5, 6×6의 세 가지로 나누어 학습을 진행시켰다.

(그림 1)은 4×4, (그림 2)는 5×5, 그리고 (그림 3)은 6×6의 형상지도를 갖는 SOM의 군집 결과를 나타낸다. 여기서, 형상지도의 차원이 커지면 학습의 결과로서 형성되는 군집의 수도 증가함을 알수 있다. 그림에서 진한 부분은 데이터들이 많이 몰려있는 군집을 의미하고, 이 정보를 이용하여 최종 군집의 개수를 결정하며 군집화를 수행한다. 3개의 그림을 종합한 최종 군집화 결과는 <표 2>와 같이 나타낼 수 있다.

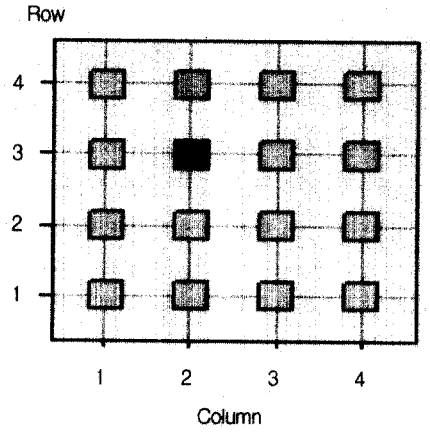
〈표 2〉 형상지도의 차원의 크기에 따른 군집화 결과

Feature maps / Clustering	Dimension of 4×4	Dimension of 5×5	Dimension of 6×6
Cluster 개수	2개	4개	8개
최종 군집화 노드	(3, 2), (4, 2)	(3, 1), (1, 2) (3, 4), (2, 5)	(3, 1), (2, 1) (1, 1), (1, 2) (1, 6), (2, 6) (3, 6), (4, 6)

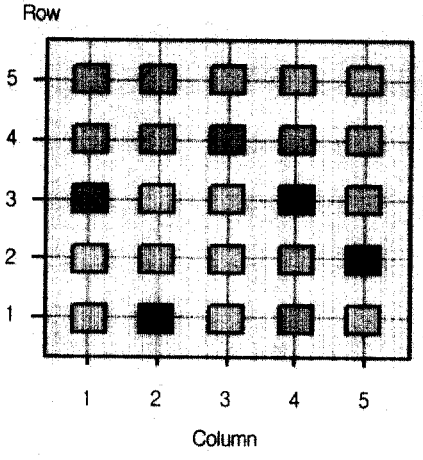
(그림 1)에서 (그림 3)까지의 결과에서 각 노드의 명암이 진할수록 노드에 포함된 사례수가 많음을 의미한다. 이 결과에 의하면 형상지도의 크기에 따라 최종 군집 수가 증가하는 것으로 나타났다.

형상 지도의 크기가 4×4인 경우에는 최종적으로 2개의 군집이 형성되었다. 또한 형상 지도의 차원의 크기가 5×5인 경우에는 최종 군집의 수가 4개로 나타났으며, 형상지도

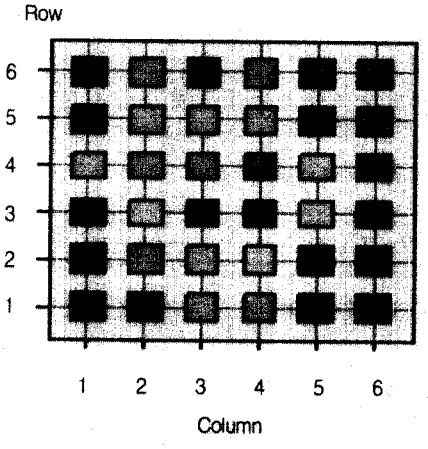
의 크기가 6×6인 경우에는 최종 군집의 수가 8개가 되었다. 이러한 군집의 개수는 프로젝트 지식 베이스에서 유형으로 구축하는 프로젝트 군집의 수가 된다. 형상 지도 차원의 크기에 따른 여러 군집 수 중에서 실제 적용하려는 분야의 전문가가 자신의 업무영역에 맞게 결정하게 된다.



(그림 1) 4×4 군집 결과



(그림 2) 5×5 군집결과



(그림 3) 6×6 군집결과

〈표 3〉 형상 지도의 차원의 크기에 따른 각 프로젝트의 할당

Feature-maps / Items	Dimension of 4×4		Dimension of 5×5		Dimension of 6×6	
	distance	winner node	distance	winner node	distance	winner node
1	1.3735	(1,4)	1.3133	(1,4)	1.2952	(2,5)
2	1.3666	(3,4)	1.3802	(2,2)	1.2659	(6,6)
3	1.2492	(4,3)	1.2026	(5,4)	1.1931	(6,4)
4	1.3722	(4,3)	1.3262	(5,4)	1.4275	(1,6)
5	1.2514	(4,2)	1.2856	(5,3)	1.2771	(3,4)
...	...	...	...	...	...	
1006	1.2774	(1,1)	1.3014	(1,1)	1.2087	(2,2)
1007	1.4027	(4,1)	1.2944	(4,2)	1.3048	(5,2)
1008	1.2354	(1,4)	1.2459	(1,4)	1.2791	(1,6)
1009	1.4294	(1,1)	1.5162	(5,5)	1.4762	(5,1)
1010	1.4666	(1,3)	1.3197	(1,4)	1.4279	(2,2)

이러한 군집 결과를 본 논문에서의 프로젝트 패턴 지식 베이스의 구축을 위한 전략으로 판단할때, 2개의 군집 수는 패턴화되는 규칙의 수가 너무 작아 프로젝트의 유형별 관리를 위한 적절한 구분이 아닌 것 같다. 또한, 프로젝트 군집 수를 8개로 결정하는 것은 패턴화되는 프로젝트의 유형이 너무 세분화되어 효율적인 지식 베이스의 관리에 어려움이 있을 것으로 판단된다. 이를 위해 실세계에서 관리와 규칙을 위한 적용에는 최종적으로 4개의 군집패턴을 사용하였으며, 형상지도의 차원을 5×5로 결정하였다. 최종적으로 각 프로젝트는 4개의 군집에 할당되었으며, 군집결과는 <표 4>와 같다.

〈표 4〉 최종 군집의 프로젝트 할당 수

Cluster	Frequency	Percentage (%)
1	280	27.72
2	238	23.56
3	258	25.54
4	234	23.18
Total	1010	100.00

각 군집에 대한 정의를 내리기 위해서는 변수에 대한 각 군집의 평균 <표 5>와 변수 자체의 표준편차 <표 6>을 이용한다.

새로운 프로젝트가 발생하면 $X_1, \cdots, X_{27}$의 27개의 프로젝트 분류 변수를 이용하여 각 변수의 평균과 분산을 기초로, 가장 유사한 사례들을 찾아내어 프로젝트 수행 지식으로 활용하기 위해 찾아낸 사례들을 종합적으로 고려하여 프로젝트 관리를 수행한다. 이 프로젝트 수행에 대한 결과의 성공 및 실패의 사례들 모두가 다시 지식 베이스의 갱신에 활용된다. 이렇게 구축된 지식 베이스를 지속적으로 관리하

여 향후 발생하는 프로젝트와 가장 유사한 기존의 프로젝트 군집(본 논문에서는 4개의 군집)의 특성을 이용하여 새로운 프로젝트의 수행과 관리를 반복적으로 수행한다.

〈표 5〉 Cluster Means

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	2.28	2.81	3.00	2.91	2.68	2.75	1.31	1.46	1.17
2	2.29	2.71	2.93	2.77	2.73	2.82	1.31	1.65	1.30
3	2.28	2.5	2.79	2.76	2.71	2.90	1.34	1.43	1.29
4	2.32	2.79	2.85	2.80	2.74	2.65	1.23	1.44	1.29
Cluster	X10	X11	X12	X13	X14	X15	X16	X17	X18
1	1.42	1.45	1.32	1.43	1.56	1.64	1.56	1.59	3.31
2	1.40	1.57	1.30	1.49	1.60	1.63	1.62	1.64	3.35
3	1.40	1.39	1.30	1.58	1.60	1.56	1.64	1.58	3.33
4	1.36	1.53	1.35	1.47	1.60	1.55	1.62	1.67	3.38
Cluster	X19	X20	X21	X22	X23	X24	X25	X26	X27
1	3.25	4.56	3.36	3.22	4.66	3.44	3.23	2.81	2.35
2	3.18	3.26	3.29	3.39	1.55	3.32	3.18	4.79	2.34
3	3.40	1.43	3.29	3.10	4.67	3.51	3.34	3.00	2.47
4	3.19	3.29	3.25	3.27	1.15	3.48	3.24	1.38	2.59

〈표 6〉 Total Standard Deviation(STD) of each variable

Variable	STD	Variable	STD	Variable	STD
X1	0.7935	X10	0.6527	X19	1.2081
X2	1.3030	X11	0.6738	X20	1.7097
X3	1.1884	X12	0.6590	X21	1.0387
X4	0.9776	X13	0.6741	X22	1.2965
X5	0.7825	X14	0.6548	X23	1.8783
X6	0.9206	X15	0.7985	X24	1.0462
X7	0.6468	X16	0.6787	X25	1.1928
X8	0.7904	X17	0.6534	X26	1.7733
X9	0.5471	X18	1.1926	X27	0.9602

4.2 하이브리드 SOM의 성능평가 및 결과

앞 절에서 구축한 지식베이스를 이용하여 새로운 프로젝트 관리에 대한 기존 방법과 제안 방법의 성능 평가를 위한 실험은 <표 6>의 프로젝트 형태를 이용하였다. <표 6>을 이용하여 총 27개의 프로젝트 분류 항목을 (1)부터 (27)까지의 입력 변수로 정의하였다.

우선 새로운 프로젝트에 대해서 4.1절에서 구축한 기존의 프로젝트 사례들의 지식베이스로부터 가장 유사한 프로젝트를 찾아내는 실험을 수행하였다.

지식베이스의 전체 사례들을 검색하는 일반적인 방법과 본 논문에서 제안하는 자율적 데이터마이닝 알고리즘에 의한 방법으로 수행한 결과에 얻어지는 과거의 유사 프로젝트 사례들에 대한 정확도 관점에서 성능비교를 하였다. 구체적인 성능 비교 방법으로 검색된 사례들이 얼마나 새로운 프로젝트 관리에 필요한 정보를 안정적으로 제공할 수 있는지에 대한 동질성 검증(homogeneity testing)을 수행하였다. 왜냐하면 하나의 새로운 프로젝트 관리를 위한 정보를 얻기 위해서 구축된 지식베이스로부터 검색한 결과로서의 사례들이 서로 이질적인 것들로 섞여 있다면 어느 것을 현재의 프로젝트에 적용해야 할지 혼란에 빠지게 되며 이것들로부터 얻게 되는 정보를 새로운 프로젝트 수행을 위한 일반화 규칙으로 사용하기는 어렵다. 하지만 검색된 사례들이 서로 동질적인 것들로 구성되면 이것들로부터 얻어지는 정보는 새로운 프로젝트 관리를 위한 효율적인 규칙으로 사용이 가능하게 된다. 이러한 성능평가의 도구(tool)로서 χ^2- test [10]를 실시하였다. 또 다른 성능평가는 최종적으로 얻어진 사례들의 전체 산포의 측도인 표준편차를 이용하였다. 즉, 표준편차가 크면 그 만큼 이질적인 것들이 많이 내재해 있다는 것으로 해석할 수 있기 때문이다. 전자는 빈도에 의한 성능평가이고, 후자는 양에 의한 성능 평가이다. 기존의 방법과 제안 방법을 적용하여 각각 상위 20개의 유사 사례를 검색하였다.

〈표 7〉 실험을 위한 새로운 프로젝트

입력변수	세부항목	입력변수	세부항목
(1)	공 공	(15)	2개
(2)	N/W	(16)	VB
(3)	회 계	(17)	Method/1
(4)	1억~5억	(18)	50~90%
(5)	10명~50명	(19)	90~110%
(6)	1~3개월	(20)	설 계
(7)	2개	(21)	50~90%
(8)	오라클	(22)	50~90%
(9)	3개이상	(23)	분 석
(10)	UNIX	(24)	150%이상
(11)	2개	(25)	50%이하
(12)	SQL	(26)	구 현
(13)	1개	(27)	대 리
(14)	SA		

기존 방법과 제안 방법에 의해 검색된 상위 20개의 사례들에 대한 동질성 검증을 수행하였다. 이 검증을 위한 가설은 다음과 같이 나타낼 수 있다.

$$H_0 : C_1 = C_2 = ,\cdots, = C_{20} \quad vs \quad H_1 : /H_0$$

귀무 가설(null hypothesis) H_0는 모든 사례들이 서로 유사하여 동질성이 크다는 것으로 해석되고, 대립 가설(alternative hypothesis) H_1은 이중에 이질적인 것들이 포함되어 있다는 것으로 해석된다. 검색된 상위 20개의 프로젝트 사례들에 대한 카이제곱 동질성 검정 결과는 <표 8>과 같다.

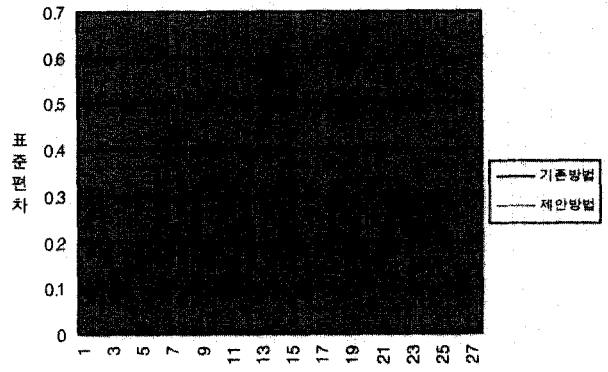
〈표 8〉 카이제곱 검정을 통한 성능비교

	χ^2	p-value
기존 방법	33.45	0.02
제안 방법	25.15	0.16

결과에 의하면 기존 방법에 의해 검색된 사례들의 동질성 검증은 95%의 신뢰수준에서 유의확률(p-value) 0.02로서 귀무 가설(H_0)을 기각한다. 즉 서로 이질적인 프로젝트 사례들이 섞여있어 안정적인 일반화 프로젝트 관리에 필요한 정보를 얻기가 어렵다.

본 논문에서 제안하는 방법을 이용하여 검색한 상위 20개의 프로젝트 사례들에 대한 결과는 유의 확률이 0.16으로서 검색된 사례들이 모두 유사하다는 귀무 가설을 기각할 수 없기 때문에 검색된 사례들에 대한 동질성이 유지된다고 볼 수 있다. 또한 서로 동질성이 유지되는 프로젝트 사례들을 이용하여 새로운 프로젝트의 관리를 위한 객관적이고 안정적인 정보를 통한 적용 규칙을 찾아낼 수 있다.

(그림 4)는 기존 방법과 제안 방법에 의해 검색된 상위 20개의 프로젝트 사례들에 대해 각 입력 변수에 대한 표준편차를 비교한 결과이다. 여기서, X축은 각 입력변수를 나타내고, Y축은 표준편차를 나타낸다. (그림 4)에 따라 본 논문에서 제안한 알고리즘의 표준편차가 기존의 방법에 비해 작음을 알수 있다.



(그림 4) 표준 편차를 통한 비교

5. 결론 및 향후 과제

본 논문에서는 지식 베이스의 효율적 관리를 위하여 자율학습 신경망과 확률적 분포를 결합한 Hybrid SOM 알고리즘을 제안하였다. 이 방법은 기존의 SOM 모형의 장점은 유지하고 단점을 통계적 분포이론으로 보완한 방법이다. 프로젝트 사례 기반의 지식 베이스 관리에 대한 실험 결과에서 나타나듯이 지식 베이스 관리에 본 논문에서 제안하는 방법이 우수한 성능을 발휘하고 있다. 향후, 보다 빠르고 지능화된 프로젝트 관리를 위한 지식 베이스의 구축 및 유지 보수를 위하여 본 논문에서 사용한 자율학습 신경망 이외에도 유전자 알고리즘(genetic algorithm), SVM(support vector machine)등의 기계학습 알고리즘[7]과 베이지안 학습(Bayesian learning)을 포함한 여러 다양한 통계적 학습 이론(statistical learning theory)을 결합하는 지식 베이스 관리 알고리즘들이 개발되면 좀 더 우수한 성능 결과가 기대되어 진다.

참 고 문 헌

[1] C. M. Bishop, M. Svensen, C. K. I. Williams, GTM ; The Generative Topographic Mapping, Neural Computation. 10. 1. pp.215-235, 1996.

[2] R. B. Chase & N. J. Aquilano, Production and Operations Management, Irwin, pp.502-504, 1989.

[3] Giarratano & Riley, Expert System, PWS, 1998.

[4] C. Guilfoyle, Ventors of agent technology, in Proc. UNICOM Seminar Intell. Agents and Their Business Applicat., London, U. K., pp.135-142, 1995.

[5] Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[6] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics, 43, pp.59-69, 1982.

[7] T. Kohonen, Self-Organizing and Associative Memory, Springer-Verlag, Berlin, 1984.

[8] T. Kohonen, Self Organizing Maps, Springer, 1997.

[9] Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[10] William J. Kennedy, Jr James E. Gentle, Statistical Computing, Marcel Dekker, INC., 1980.

[11] 김제국외, "종합전산화 추진체계의 고찰", 전력기술, 제31호, pp.6-17, 1997.

[12] 이순용, "생산관리론", 법문사, 서울, p.562, 1989.

윤 경 배

e-mail : kbyoon@kimpo.ac.kr

1986년 인하대학교 수학과(이학사)

1994년 인하대학교 산업대학원 정보공학과 (공학석사)

1998년 서강대학교 경제대학원 정보기술 경제학 (경제학석사)

1999년~현재 인하대학교 대학원 전자계산공학과(박사과정 수료)

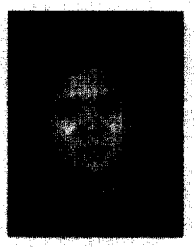
1986년~1987년 대우자동차(주) MIS 근무

1988년~1991년 LG-EDS(주) 기술연구소 근무

1992년~1997년 동부정보기술(주) 연구소 근무

1998년~현재 김포대학 컴퓨터계열 조교수

관심분야 : 소프트웨어 공학, 지식기반 데이터베이스, 데이터마이닝, CRM, 인공지능 등



최 준 혁

e-mail : jhchoi@kimpo.ac.kr

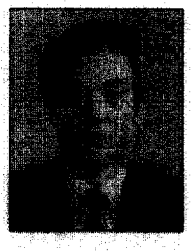
1990년 경기대학교 전자계산학과 졸업 (이학사)

1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사)

2000년 인하대학교 대학원 전자계산공학과 졸업(공학박사)

1997년~현재 김포대학 컴퓨터계열 교수

관심분야 : 정보검색, 데이터마이닝, 신경망, 유전자 알고리듬 등



왕 창 종

e-mail : cjwangse@inha.ac.kr

1979년~현재 인하대학교 전자계산 공학과 교수

1997년~2001년 인하대학교 사회 교육원장

1992년~1994년 한국정보과학회 부회장

관심분야 : 소프트웨어공학, 분산객체 컴퓨팅, 지능형웹기반교육시스템 등