자동 군집화를 위한 지능화된 데이터 마이닝 에이전트

박정은* · 전성해* · 오경환*

*서강대학교 컴퓨터학과

Intelligent Data Mining Agent for Automatic Clustering

Jung-Eun Park* · Sung-Hae Jun* · Kyung-Whan Oh*

*Dept. of Computer Science, Sogang University

요약

인터넷 환경에서 발생되는 수많은 데이터를 지능적으로 처리할 수 있는 자동화된 분석 시스템의 필요성이 제기된다. 이러한 시스템의 데이터 분석은 크게 지도 학습과 자율 학습으로 나뉜다. 본 논문에서는 특히 자율학습 군집화에 대한 자동화된 시스템으로서 지능화된 데이터 마이닝 에이전트를 제안한다. 군집화 과정에서는 데이터를 분석하는 분석가가 군집화의 방법과 결과 해석에 실시간으로 관여하기 어렵기 때문에 이러한 작업을 담당하는 지능화된 에이전트가 자동화된 군집화를 담당하면 효과적인 군집화 전략이 될 수 있다. 본 논문의 자동 군집화를 위한 지능화된 데이터 마이닝 에이전트 시스템은 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 다중 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행한다. UCI Machine Repository 데이터를 이용한 실험을 통해 제안 시스템의 성능 평가를 수행하였다.

Key words : Intelligent Data Mining Agent, Automatic Clustering, PCA, SOM, K-means, VC

1. 서론

인터넷 환경의 e-Business 시스템에서는 에이전트 기반 모델에 대한 필요성이 제기된다. 특히 대용량의 데이터로부터 유용한 패턴을 찾아내는 데이터 마이닝 기법은 에이전트 기반의 e-Business 모델에서 매우 유용하게 사용된다. 데이터 마이닝 기법 중에서 군집화는 사용자, 웹 문서 등 인터넷 거래 행위의 주체 및 개체들을 서로 유사한 것들끼리 묶어 주는 역할을 담당한다. 대부분의 이벤트들이 실시간에 발생되고 처리되어야 하는 인터넷 환경에서는 분석가가 군집화의 방법과 결과 해석에 계속적으로 관여하기 어렵기 때문에 이러한 분석가의 업무의 상당 부분을 담당하는 지능화된 에이전트가 필요하며 이러한 에이전트가 자동화된 군집화를 담당하게 된다. 본 논문의 자동 군집화를 위한 지능화된 데이터 마이닝 에이전트 시스템은 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 다중 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행한다. 본 논문에서 제안하는 자동 군집화를 위한 지능화된 데이터 마이닝 에이전트는 이러한 알고리즘으로 이루어진 두 개의 에이전트가 최적의 군집화를 위하여 작동하게 되는 멀티 에이전트 시스템이다. 제안 시스템에 대한 성능 평가를 위한 실험은 IRIS 데이터와 GLASS identification 데이터를 이용하였다. 이 논문의 2절에서는 제안 시스템에 대한 관련 연구를 알아보고, 다음으로 제안 시스템에 대한 통합설계와 프로세스는 3절에서 나타내고 있다. 4절에서는 UCI Machine Repository 데이터를 이용한 실험을 통해 제안 시스템의 성능 평가를 수행하였고 마지막으로 5절에서 결론 및 향후 연구과제에 대하여 논의하였다.

2. 관련 연구

2.1 지능형 데이터 마이닝

분산되어 있는 대용량의 운영계 데이터베이스로부터 거대한 데이터 웨어하우스를 구축하여 마이너(miner)가 주가 되어 주어진 데이터에 대한 모형을 구축해 가는 기존의 오프라인 데이터 마이닝 프로세스에 비해 인터넷 환경의 웹 마이닝 프로세스는 실시간에 발생되는 트랜젝션의 분석과 처리에

있어서 오프라인에서 마이너가 담당했던 많은 작업들을 에이전트가 담당해야 할 필요성이 많아지게 되었다. 즉, 사람에 의한 실시간 데이터의 분석 및 처리는 어렵게 되고 상당 부분의 마이닝 작업은 기계에 의해 자동적으로 처리되어 져야 하기 때문이다.[2]

2.2 주성분 분석

2.2.1 주성분의 개념

여러개($p \geq 2$)의 반응변수에 대하여 얻어진 다변량 자료를 분석의 대상으로 하는 주성분 분석은 다차원적인 변수들을 축소하는 차원의 단순화와 서로 상관되어있는 변수들 상호간의 복잡한 구조를 분석한다. 즉, 반응변수들을 선형 변환시켜, 주성분이라는 서로 독립적인 새로운 변수들을 유도한다. 이때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도의 순서를 생각할 수 있는데, 그들 중 상위 m개($m << p$)의 주성분에 의해 원래자료에 내재하는 전체변이 중 가능한 한 많은 부분이 보유되도록 변환시킴으로서 정보의 손실을 최소화하는 차원의 축소(dimension reduction)가 이루어진다.

서로 상관되어 있는 $p(\geq 2)$개의 확률변수 $X_1, X_2, \ldots, X_p$를 원소로 하는 확률벡터 X가 평균벡터 μ와 공분산행렬 Σ를 다음과 같이 표현한다.

$$X = \begin{matrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_i \\ \cdot \\ \cdot \\ X_p \end{matrix} \quad \mu = \begin{matrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_i \\ \cdot \\ \cdot \\ \mu_p \end{matrix} \quad \Sigma = \begin{matrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \sigma_{1j} & \cdot & \cdot & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdot & \cdot & \sigma_{1j} & \cdot & \cdot & \sigma_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{i1} & \sigma_{i2} & \cdot & \cdot & \sigma_{ij} & \cdot & \cdot & \sigma_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdot & \cdot & \sigma_{pj} & \cdot & \cdot & \sigma_{pp} \end{matrix}$$

여기서 $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ji}$는 X_i와 X_j의 공분산이다. 그리고 $\mu_i = E[X_i]$는 X_i의 평균이다. 또한 p는 변수의 개수를 나타낸다. σ_{ii}는 i번째 변수의 분산이고 σ_{ij}는 i번째 변수와 j번째 변수의 공분산을 나타낸다. 주성분분석은 입력 변수벡터 X를 선형변환시켜 정보의 손실을 최소화하면서 p보다 매우 작은 m개의 새로운 인공변수를 생성함으로서, p차원 변이를 m차원으로 축소하여 전체 데이터의 특성을 요약하여 전체 변수들간의 복잡한 구조를 파악하고자 하는 것이다. 이 분석은 X의 원소들 간의 상관구조관계를 나타내는 Σ를 분석대상으로 한다. Σ의 p개의 고유값(eigen value), δ_j들을 크기순으로 배열하고 각각의 고유값에 대응되는 고유벡터(eigen vector), r_j의 짝들을 $(\delta_1, r_1), (\delta_1, r_2), (\delta_1, r_3), (\delta_1, r_4), \cdots, (\delta_p, r_p)$ 라 하고, δ_j들을 크기순으로 배열하면 다음 식과 같이 표현된다.

$$\sum r_j = \delta_j r_j, \; j = 1, 2, \ldots, p \qquad \text{<식 2-1>}$$

$\delta_1 \geq \delta_2 \geq \delta_3 \cdots \geq \delta_p$의 관계를 갖게 된다. 이 값에 의해 p개의 입력변수를 m개의 주성분으로 차원을 축소하여 데이터의 특성을 파악하게 된다. 이때 주성분의 개수는 다음 절에서와 같이 몇 가지 판정 기준을 통해 결정하게 된다.

2.2.2 보유주성분의 수에 관한 판정기준

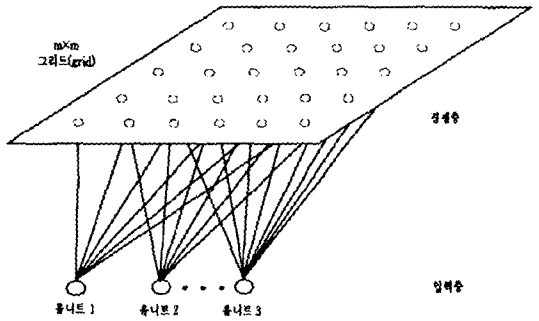
우선 전체 변이에 대한 공헌 정도로 결정할 수 있다. 즉, 보유 주성분들이 전체 분산에 대하여 주어진 일정 비율(예를 들어 80~90%) 이상을 설명할 수 있기 위해 필요한 최소 개수의 주성분을 보유하는 방법이다. 다음으로 고유값의 크기를 이용할 수 있다. 이 기준은 주성분으로 보유되기 위해서 대응되는 고유값은 적어도 1이상이 되어야 한다는 것으로서, 이는 Kaiser의 규칙(Kaiser, 1960)으로 알려져 있다.[11]

2.3 자기조직화 지도

2.3.1 코호넨 네트워크

여러 가지 신경망 모형 중에서 인간의 뇌 구조를 가장 잘 모형화 한 자기 조직화 지도(Self Organizing Maps ; SOM)는 1980년대 초에 kohonen에 제안된 신경망 모형이다. 자기 조직화 지도는 신경망 중에서도 학습 자료에 대한 결과값을 모르고 학습이 수행되는 자율 학습(unsupervised learning) 구조를 가지고 있다. 음성 인식, 문자 인식, 구문 분석 등 다양한 분야에 응용되는 자기 조직화 지도는 입력층과 출력층으로 구성된 순방향 단층 신경망 구조를 갖는다. 자기 조직화 지도의 연결강도인 가중치는 정규화된 입력 벡터에 대응되는 출력 노드의 중심값과 같은 역할을 하며 학습 동안에 입력 벡터와 가장 가까운 측도(유클리디안 거리)를 갖는 출력 노드가 승자(winner)가 되고, 이 승자 노드와 이웃하는 것들의 가중치만이 갱신한다. 특히 자기 조직화 지도는 다층 신경망과 같은 지도 학습(supervised learning) 모형에 비해서 매우 단순한 2개의 층(layer)으로 이루어지면서 다차원의 자료를 2차원의 형상 지도(feature maps)로 투영(projection)시켜 스스로 경쟁 학습(competitive learning)을 할 수 있도록 한다. 2개의 층은 입력벡터를 갖는 입력층(input layer)과 형상 지도를 갖는 출력층(output layer)이다. 출력층으로서의 경쟁층(competitive layer)은 일반적으로 2차원 지도(maps)로 되어 있다. 자기 조직화 지도 모형은 오류 역전파(backpropagation) 모형과는 달리 여러 단계의 피드백을 거치지 않고 오직 한 번의 전방 전달(feedforward flow)만을 사용하며 입력층에서 출력층으로는 모두가 연결(fully connected)되어 있는 구조를 갖고 있다. Kohonen network를 만들 때 다른 신경망들에서는 일반적으로 필요하지 않는 두 가지 일을 해야 한다. 하나는 층내의 뉴런의 연결강도 백터가 임의값을 가지면서 적합하게 초기화되어야 한다. 다른 하나는 연결강도 백터와 입력백터가 통상 0에서 1사이의 정규화된(normalized) 값을 사용한다. 이런 두 가지 요인은 Kohonen network에 있어서 매우 중요하다.

[그림 2.1] 코호넨 네트워크



2.3.2 경쟁학습(Competitive Learning)

Kohonen의 학습에서 각 뉴런은 연결강도 벡터와 입력벡터가 얼마나 가까운가를 계산한다. 그리고 각 뉴런들은 학습할 수 있는 특권을 부여받으려고 서로 경쟁하려는데 거리가 가장 가까운 뉴런이 승리하게 된다. 이 승자 뉴런이 출력신호를 보낼 수 있는 유일한 뉴런이다. 또한 이 뉴런과 이와 인접한 이웃 뉴런들만이 제시된 입력벡터에 대하여 학습이 허용된다. 이것은 학습에 있어서 전혀 새로운 접근 방식이다. 이 모델이 있기 이전에는 network에 있는 모든 뉴런들이 반복되는 훈련 과정에서 연결강도를 조정한다. Kohonen network의 학습 철학은 '승자 독점(winner take all)'이다. 승자만이 출력을 낼 수 있으며, 승자와 그의 이웃들만이 그들의 연결강도를 조정할 수 있다. 승자 뉴런을 결정하고 난 후에는 Kohonen의 학습 규칙에 따라 뉴런의 연결강도를 조정해야 한다. 이 규칙은 다음식으로 표현된다.

$$W_{new} = W_{old} + \alpha (X - W_{old}) \qquad \text{<식 2-2>}$$

여기서 W_{old}는 조정되기 이전의 연결 강도 백터이며, W_{new}는 조정된 후의 새로운 연결강도 백터이고, X는 입력패턴 백터이며, α는 학습상수이다.

2.3.3 자기조직화 형상지도 알고리즘

Step 1 : initialization :
Choose random values for the initial weights
Step 2 : winner finding :
Find the winner neuron j^* at time k, using the minimum-distance criterion :
$j^* = argmin_j \| x(k) - w_j \|, \quad j = 1, ..., N^2$
where, $x(k)$ represents the kth input pattern and $\| . \|$ indicated the euclidean norm.
Step 3 : weight updating :
Adjust the weights of the winner and its neighbors, using the following rule

$$w_j(k+1) = \begin{cases} w_j(k) + \eta(k)(x(k) - w_j(k)) & \text{if } j \in N_{j^*}(k) \\ w_j(k) & o.w. \end{cases}$$

where, $\eta(k)$ is a positive constant and $N_{j^*}(k)$ is the neighborhoodset of the winner neuron j^* at time k.
Repeat until given conditions satisfication.

2.3.4 형상 지도의 차원 결정 문제

Kohonen network는 여러 가지 장점들을 가지고 있지만 또한 문제점도 있다. 그 중 하나가 형상 지도(feature maps)의 차원을 주관적으로 결정해야 한다는 것이다. 차원이 크면 군집수가 그에 따라 증가하며, 반면에 차원이 작으면 군집수가 감소하게 된다. 이러한 Kohonen network의 문제점을 해결하기 위한 방안이 주성분분석을 이용한 객관적인 형상지도의 차원결정이다.[1][5][12]

3. 제안 시스템의 통합 설계

3.1 지능형 데이터 마이닝 에이전트를 이용한 자동 군집화

군집화는 사용자, 웹 문서 등 인터넷 거래 행위의 주체 및 개체들을 서로 유사한 것들끼리 묶어주는 역할을 담당한다. 대부분의 이벤트들이 실시간에 발생되고 처리되어야 하는 인터넷 환경에서는 분석가가 군집화의 방법과 결과 해석에 계속적으로 관여하기 어렵기 때문에 이러한 분석가의 업무중 일정 부분을 담당하는 지능화된 에이전트가 필요하며 이러한 에이전트가 자동화된 군집화를 담당하게 된다. 본 논문의 자동 군집화를 위한 지능화된 데이터 마이닝 에이전트 시스템은 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 다중 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행한다.

3.1.1 군집화 수행 에이전트

자동적으로 군집화를 수행하는 군집화 수행 에이전트에서는 군집화 알고리즘으로서 코호넨이 제안한 자기 조직화 형상지도를 이용한다. 이 기법을 사용한 이유는 자기 조직화 형상지도가 매우 빠른 군집화를 가능하게 하여 실시간 웹 데이터의 군집화에 적합하기 때문이다. 자기 조직화 형상지도에서 형상 지도의 차원이 커지면 최종 군집 수가 커지는 문제점을 해결하기 위해 본 논문에서는 다변량 통계 기법중의 하나인 주성분 분석을 이용해 최적의 차원을 결정하였다.

3.1.2 군집화 성능 평가 에이전트

군집화 성능 평가 에이전트는 군집화 수행 에이전트로부터의 결과에 대한 성능 평가를 담당한다.

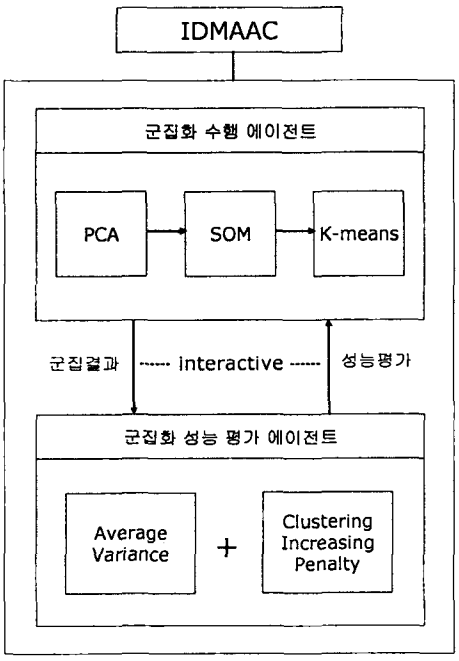
만약 성능이 좋지 않은 결과를 얻게 되면 군집화 수행 에이전트에게 더 나은 군집 분석을 요구한다. 군집 결과의 성능 평가는 본 논문에서 제안하는 Variance Criterion(VC)을 통해 수행한다. 이 기준은 군집화에 사용된 군집 변수 중 연속형 변수에 대해서 분산(variance)을 적용하여 이 값들이 작은 것을 좋은 결과로 결정하였다. 이러한 척도를 적용한 근거는 군집화의 개념이 같은 군집내의 개체들은 서로 동질성이 크고 서로 다른 군집간의 개체들간에는 이질성이 크도록 하기 때문이다. 또한 VC 척도에서는 군집 수에 대한 페널티를 포함하는데, 동일 데이터에 대한 군집 수가 증가하면 군집 결과의 각 군집에 대한 동질성은 증가하지만 군집의 수가 너무 많아지면 분석의 의미가 없어지므로 군집수의 증가는 결과의 성능 평가에서 페널티로 작용하게 하였다. 즉, VC는 연속형 군집 변수의 분산과 군집수 증가에 따른 페널티로 이루어져 있다. 본 논문에서 제안하는 VC 측도는 <식 3-1>과 같이 정의하였다.

$$VC_M = \sum_{i=1}^{M} v_i/M + 0.1 * M \qquad \text{<식 3-1>}$$

여기서 M는 군집의 수이다. 그리고 v_i는 i번째 군집의 평균 분산이 된다. 두 번째 항에 있는 $0.1 * C$는 군집수에 따른 패널티이다. 즉 <식 3-1>의 값이 작을 수록 좋은 군집 결과이다.

3.1.3 IDMAAC 시스템

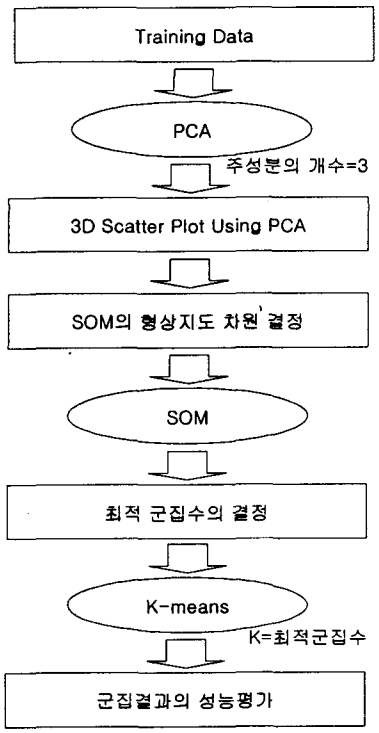
[그림 3-1] IDMAAC의 시스템 구조



군집화 성능 평가 에이전트는 군집화 수행 에이전트를 포함하는 자동 군집화를 위한 지능화된 데이터 마이닝 에이전트(Intelligent Data Mining Agent for Automatic Clustering: IDMAAC)는 [그림 3-1]과 같은 시스템 구조를 가지고 있다.

[그림 3-1]처럼 IDMAAC는 군집화 수행 에이전트와 군집화 성능 평가 에이전트가 서로 대화식의 구조를 띠고 있다. 즉 한쪽은 군집화만 수행하고 다른 한 쪽은 군집 결과에 대한 성능 평가만을 수행한다.

[그림 3-2] IDMAAC를 이용한 자동 군집화 절차



3.2 자동 군집화의 필요성

본 논문에서는 [그림 3-2]에서 보여 지는 것과 같이 자동 군집화를 수행한다. 우선 학습 데이터에 대하여 주성분 분석을 통하여 3개의 주성분을 이용하여 전체 데이터에 대한 산점도를 그린다. 보유 주성분의 개수를 3개로 하여 3차원 산점도를 그린 이유는 시각적으로 관찰할 수 있는 산점도가 최대 3차원이기 때문이다[3]. 이 산점도를 이용하여 전반적인 데이터의 군집 구조를 파악한다. 이 산점도를 통하여 SOM의 형상 지도의 차원을 결정한다. 본 논문에서는 형상지도의 차원은 산점도에 의한 (군집수*군집수)로 결정하였다. 이는 여러 차례의 실험을 통하여 휴리스틱하게 결정되었다. 주성분 분석을 통하여 형상지도의 차원이 결정된 SOM을 이용하여 최적 군집수 결정을 위한 자율 학습(unsupervised learning)을 수행한다. 이 결과를 이용

하여 최적의 군집수를 결정하고 이 값을 K-평균 군집 분석의 초기 군집수로 결정하여 최종적인 군집화를 수행한다. 이 과정까지를 군집화 수행 에이전트가 담당한다. 다음은 군집화 수행 에이전트가 VC 판단기준을 이용하여 군집화 성능 평가를 수행한다.

4. 실험 및 결과

4.1 Iris 데이터를 이용한 실험 및 결과

Iris Plants Database는 150개의 학습 데이터로 이루어져 있다. 4개의 입력변수, x_1(sepal length in cm), x_2(sepal width in cm), x_3(petal length in cm), x_4(petal width in cm)이 있고 이 변수들의 붓꽃의 종류를 결정해 준다. <표 4-1>은 이 데이터의 간단한 특성을 보여주고 있다.[8]

<표 4-1> Iris 데이터의 요약 정보

	Min	Max	Mean	SD
x_1	4.30	7.90	5.84	0.83
x_2	2.00	4.40	3.05	0.43
x_3	1.00	6.90	3.76	1.76
x_4	0.10	2.50	1.20	0.76

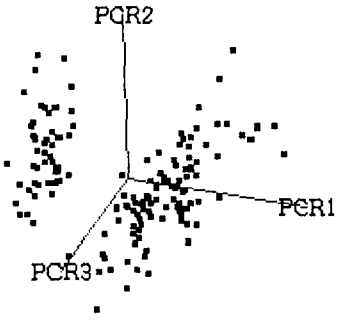
우선 4개의 입력 변수를 가지고 있는 Iris 데이터에 대한 주성분 분석을 수행한 결과가 <표 4-2>에 나타나 있다. 보유 주성분 3개가 전체 데이터의 99.48%를 설명하고 있다. 따라서 차원 축소에 따른 정보 손실은 거의 없다고 볼수 있다.

<표 4-2> Iris 데이터의 주성분 분석 결과

성분	고유값	누적(%)
1	2.910818	72.77
2	0.921221	95.80
3	0.47353	99.48

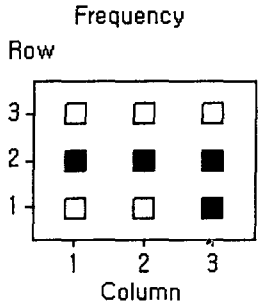
3개의 주성분을 이용하여 전체 데이터에 대한 3차원 산점도를 그린 결과가 <그림 4-1>과 같이 나타났다. 전체적으로 3개의 군집수를 관찰할 수 있다. 이 결과로부터 SOM의 형상 지도의 차원을 (3*3)로 결정할수 있다.

[그림 4-1] Iris 데이터의
3개 주성분의 3D 산점도



다음으로 (3*3)의 형상 지도 차원을 갖는 SOM의 군집 결과가 [그림 4-2]와 같이 나타났다.

[그림 4-2] Iris 데이터의
(3*3) SOM의 학습 결과



[그림 4-2]의 9개의 노드 중에서 진한 색을 나타내고 있는 노드가 군집이 형성된 노드이다. 그림에 의하면 3개의 군집이 형성됨을 알수 있다. <표 4-3>은 Iris 데이터를 이용한 K-평균 군집화 결과이다.

<표 4-3> Iris 데이터의 k-means(k=3) 결과

Cluster	x_1	x_2	x_3	x_4	Avg.
1	0.50	0.29	0.46	0.24	0.37
2	0.35	0.38	0.17	0.11	0.25
3	0.48	0.30	0.54	0.30	0.41

<표 4-3>은 각 군집에 대한 각 입력 변수의 분산 값을 나타내고 있고 마지막 열(Avg.)은 각 군집의 평균 분산을 나타내고 있다. 이 표의 값으로부터 3개의 군집에 대한 VC 값을 다음과 같이 구할 수 있다.

$$VC_3 = (0.37 + 0.25 + 0.41)/3 + 0.1 * 3 = 0.64$$

<표 4-4> Iris 데이터의 k-means(k=2) 결과

Cluster	x_1	x_2	x_3	x_4	Avg.
1	0.61	0.31	0.73	0.40	0.51
2	0.36	0.50	0.77	0.33	0.49

<표 4-5> Iris 데이터의 k-means(k=4) 결과

Cluster	x_1	x_2	x_3	x_4	Avg.
1	0.24	0.28	0.15	0.12	0.20
2	0.50	0.29	0.46	0.24	0.37
3	0.21	0.25	0.19	0.06	0.18
4	0.48	0.30	0.54	0.30	0.41

<표 4-6> Iris 데이터의 k-means(k=5) 결과

Cluster	x_1	x_2	x_3	x_4	Avg.
1	0.25	0.29	0.17	0.11	0.21
2	0.32	0.27	0.39	0.19	0.29
3	0.37	0.36	0.39	0.26	0.35
4	0.21	0.25	0.16	0.07	0.17
5	0.33	0.28	0.43	0.35	0.35

<표 4-4>부터 <표 4-6>은 각각 군집수가 다른 K-평균 군집화 결과이다. 이 표들로 부터의 결과를 이용한 각각의 군집 결과에 대한 VC 값들이 <표 4-7>에 나타났다.

<표 4-7> 각 군집 결과 VC값

군집수	VC값
2	0.7
3	0.64
4	0.69
5	0.07

<표 4-7>로부터 제안 시스템으로부터 결정된 3개의 군집 결과의 VC 값이 가장 작음을 알수 있다. 따라서 Iris 데이터는 3개의 군집 결과를 얻게 된다.

4.2 Glass identification 데이터를 이용한 실험 및 결과

Glass Identification 데이터는 214개의 학습 데이터로 이루어져 있다. 9개의 입력 변수는 다음과 같다. x_1은 RI(refractive index), x_2는 Na(Sodium), x_3는 Mg(Magnesium), x_4는 Al(Aluminum), x_5는 Si(Silicon), x_6는 K(Potassium), x_7는 Ca(Calcium), x_8는 Ba(Barium), 그리고, x_9는 Fe(Iron)이다. <표 4-8>은 이 데이터의 간단한 요약 정보를 나타내고 있다.[8]

<표 4-8> Glass Identification 데이터의 요약 정보

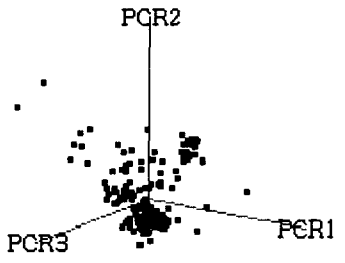
	Min	Max	Mean	SD
x_1	1.5112	1.5339	1.5184	0.0030
x_2	10.73	17.38	13.4079	0.8166
x_3	0	4.49	2.6845	1.4424
x_4	0.29	3.5	1.4449	0.4993
x_5	69.81	75.41	72.6509	0.7745
x_6	0	6.21	0.4971	0.6522
x_7	5.43	16.19	8.9570	1.4232
x_8	0	3.15	0.1750	0.4972
x_9	0	0.51	0.0570	0.0974

<표 4-9> Glass Identification 데이터의 주성분 분석 결과

성분	고유값	누적
1	2.511164	0.2790
2	2.050072	0.5068
3	1.404844	0.6629

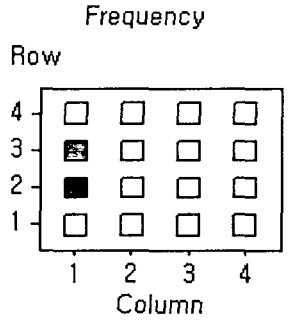
우선 9개의 입력 변수를 가지고 있는 Glass Identification 데이터에 대한 주성분 분석을 수행한 결과가 <표 4-9>에 나타나 있으며, Glass Identification 데이터를 통한 주성분 분석의 결과를 이용한 3개의 보유 주성분을 이용한 3차원 산점도가 [그림 4-3]에 나타나 있다.

[그림 4-3] Glass Identification 데이터 3개 주성분의 3D 산점도



[그림 4-3]의 결과를 이용하여 SOM의 형상 지도를 (4*4)로 결정하였다. [그림 4-4]는 (4*4)의 형상 지도 차원을 갖는 SOM의 군집화 수행 결과이다.

[그림 4-4] Glass Identification
데이터의 (4*4) SOM의 학습 결과



<표 4-10> Glass 데이터의 k-means(k=2) 결과

	x_1	x_2	x_3	x_4	x_5
1	0.002	0.74	1.26	0.50	0.71
2	0.004	1.20	0.85	0.50	1.20

	x_6	x_7	x_8	x_9	Avg.
1	0.68	0.74	0.48	0.10	0.58
2	0.28	1.61	0.67	0.11	0.71

<표 4-10>로 부터의 VC 값은 0.845가 된다.이와 같이 학습 데이터의 군집화를 수행하여 최적의 군집 결과를 얻게 된다.

5. 결론

제안된 자동 군집화를 위한 데이터 마이닝 에이전트는 최적의 초기 군집수를 주성분 분석과 자기조직화 형상지도에 의해 결정하고, 최종적으로 K-평균 군집화 알고리즘을 사용하였다. 그리고 군집 결과의 성능 평가는 VC를 제안 적용하였다. 실험에서 제안하는 최적 군집화 수행 절차에 의한 결과가 다른 방법에 비해 군집의 동질성을 높이는 결과로 나타났다. 본 논문에서 통합 설계된 자동 군집화를 위한 데이터 마이닝 에이전트는 웹 마이닝 프로세스에 동적으로 적용될 수 있게 된다. 향후 예측을 위한 데이터 마이닝 에이전트의 개발이 이루어지면 지도 학습과 자율 학습을 모두 포함하는 지능화된 데이터 마이닝 에이전트 시스템이 구축될 수 있을 것이다.

Acknowledgments

본 연구는 과학 기술부 주관 뇌신경정보학 사업에 의해 지원되었음.

References

[1] T.Kohonen, "Self-Organizing Maps", Springer, 1995
[2] Cabena, Hadjinian, Stadler, Verhees, Zanasi, "Discovering Data Mining From Concept to Implementation", Prentice-Hall, 41~102쪽, 1997
[3] FRIEDMAN, J. H. On bias, variance, 0/1-loss and the curseof -dimensionality. Tech. rep., Department of Statistics and Stanford Linear Accelerator Center, Stanford University, 1996.
[4] Sung-Hae Jun · Jihoon Yang · Kyung- Whan Oh, "Automatic Determination of Cluster Size Using Machine Learning Algorithms", Italy, July 29-August 4, SSGRR 2002s, 2002.
[5] Jaakko Hollmén, "Process Modeling Using the Self-Organizing Map", 1996
[6] Sheldon M. Ross, "Introductory Statistics", Mc Graw Hill, 244~248쪽, 1996
[7] Juha Vesanto · Johan Himberg, "SOM Toolbox For Matlab 5", Espoo 2000, 2000
[8] http://www.ics.uci.edu/~mlearn/MLRepository.html
[9] 강현철 · 한상태 · 최종후 · 김은석 · 김미경, "SAS Enterprise Miner 4.0을 이용한 데이터 마이닝 방법론 및 활용", 자유아카데미, 190~211쪽, 2001
[10] 구자홍 · 김진경 · 박헌진 · 이재준 · 전홍석 · 최지훈 · 황진수, "통계학", 자유아카데미, 61~67쪽, 1997
[11] 김기영 · 전명식, "SAS 주성분 분석", 자유아카데미, 4~7쪽 · 55~64쪽, 1992
[12] 김대수, "신경망 이론과 응용(Ⅰ)", 하이테크정보, 169~183쪽, 1992