논문 99-24-2B-16 한국통신학회논문지 '99-2 Vol.24 No.2B

개선된 Kohonen 신경망 SOM을 이용한 데이터 탐색에 관한 연구

정회원 진 영 근*, 김 태 균*

Study on the Data Exploration with Modified Self-organizing maps

Young-Goun Jin*, Tae-Kyun Kim* *Regular Members*

요 약

데이터 탐색은 구조 또는 특성이 파악되지 않는 대용량의 데이터를 효율적으로 분류하고 그 데이터에 감추어져 있는 특성을 발견하여 분석을 가능하게 하는 분야이다. 본 연구에서는 개선된 Kohonen 신경망 SOM을 이용하여 온라인 데이터 탐색을 수행하는 방법을 제시한다. SOM은 비교사 학습 방법으로 데이터 클러스터링 기능이 우수한 특징을 가지고 있으므로 이러한 특징을 유지하면서 온라인에 적용하도록 개선시켰다. 제시된 방법의 효율성을 보여주기 위하여 기존의 온라인 데이터 탐색에 사용되는 on-line K-means알고리듬과 성능을 비교 평가하여 보인다.

ABSTRACT

Data exploration is new field which explores unknown hugh data set and find data structures or characteristics hidden on the data and make it possible to analyze or estimate data behavior. In this paper we suggest on-line data exploration method based on modified Kohonen Self-Organizing Map. SOM which uses unsupervised learning shows good characteristic in data clustring. We modified SOM to apply to on-line data clustering keeping its good characteristic. Simulation result showed the effectiveness of suggested methods compare with on-line K-means methods.

I. 서 론

데이터 탐색(Data Exploration)은 구조 또는 특성 등이 파악되지 않는 대용량의 데이터를 효율적으로 분류하여 그 데이터에 숨겨져 있는 구조 또는 특성을 발견하여 데이터를 체계적으로 해석 또는 파악할 수 있도록 하는 것이다. 데이터 탐색 방법은 적용 방법에 있어 크게 심볼릭 방법과 통계 수치 방법으로 나눈다. 심볼릭 방법은 데이터 분류 및 해석을 자연어나 심볼릭하게 처리하는 것으로 대규모 문서 또는 문장에서 의미있는 중요한 단어, 어절등을 추출하는데 주로 사용되며 그 응용으로는 문장, 문서의 자동 요약문 작성, 그리고 대량의 문서를 유사 주제어 또는 개념으로 묶어 분류하는 데 사용되고 있다. 통계수치 방법은 대용량의 데이터에서 그 데이터가 가진 특성을 유지하도록 통계 수치적 방법을 이용하여 클러스터링 또는 분류를 수행하여 데이터의 속성 또는 특성을 보여주거나 의미없은 데이터를 제거하므로 데이터량을 줄여 저장하거나 또는 인간이 쉽게 알 수 있도록 표현할 수 있는 방법이다. 예로서 미국의 지구 관측위성 MODIS 두대가 하루에 지상으로 보내오는 총 영상 데이터량은 1,902GB 정도로 이들을 저장하고 분석하기는 매우 힘들고 시간과 인력이 많이 필요로 하는 일이다. 단

* 충남대학교 컴퓨터공학과(ygjin@chongyang.ac.kr)
** 논문번호 : 98180-0423, 접수일자 : 1998년 4월 23일

순히 데이터량을 줄이는 것이 목적이면 현재 개발된 각종 압축 알고리듬을 이용하여 그 데이터량을 수십분의 일 이내로 줄이는 것이 가능하다. 그렇지만 보다 인간이 그러한 대용량의 데이터를 수집하는 근본적인 목표는 데이터가 가지고 있는 유용한 정보를 파악하여 과학 발전에 도움이 되고자 하는 것이다. 그러나 이러한 대용량의 데이터를 사람이 자세히 살펴서 데이터 구조를 파악하는 것은 쉽지 않는 일이다. 다시 말해 사람이 다루기는 너무 많은 데이터로 인력 낭비가 매우 심해진다. 그러므로 대용량의 데이터를 자동으로 검색하여 데이터 구조 및 특성을 인간이 쉽게 이해할 수 있는 형태로 만드는 처리과정이 필요하며 이러한 과정을 데이터 탐색 또는 데이터 채굴이라고 한다. 분석해야할 대용량의 데이터는 과학실험분야 뿐만 아니라 실생활 속에서도 발생한다. 예를 들면 하루에도 수천 내지 수억번 일어나는 은행의 온라인 트랜잭션처리, 주식 매매 등의 전산거래 등에 대한 데이터를 효율적이고 실시간 분석하여 정부와 기업이 금융자금의 흐름 및 투자 방향을 설정하는 데 결정적인 도움을 줄 수 있다. 정보화 사회가 진행될수록 이러한 대용량의 데이터 발생은 늘어나는 추세이며 이를 분석할 수 있는 사람의 능력은 한정되어 있어 빠른 판단과 필요한 특성을 즉시에 제시하기 위한 데이터 탐색 기술 도입은 필수적이다. 그러므로 데이터량을 줄이면서 그 데이터가 가지고 있는 특성을 분석 파악할 수 있게 하는 효율적인 데이터 탐색 방법이 필요하다. 데이터 탐색은 현재 인문, 사회과학 및 정보해석과 대용량 데이터베이스 응용 분야에서 확산되고 있는 데이터 채굴(Data Mining)[1]의 한 방법으로 채굴 방식이 bottom-up인 경우에 해당한다. 데이터 탐색 방법은 데이터의 발생 방법에 따라 off-line, on-line 두가지 방식으로 나눌 수 있다. 데이터가 실시간으로 발생된다면 on-line 방식으로 탐색을 해야 한다. On-line데이터 탐색 방법은 off-line에 사용되는 방법에 비해 데이터 반복 입력이 불가능하므로 일회성 정보를 다룰 수 있어야 하며 수렴성이 좋아야 한다. 데이터 탐색에 사용되는 방법은 탐색의 최종 목표에 따라 대부분 패턴인식[2], 기계학습과 multivariate 해석에 사용되는 방법등을 차용하여 사용한다. 그 중에서 데이터 클러스터링 방법[3]은 수치 데이터 탐색에 가장 많이 사용되며 다른 탐색 방법을 적용하기 앞서 전단 처리 방법으로도 널리 사용된다. 최근에는 신경망을 이용한 데이터 탐색 방법[4][5][6]이 연구되고 있다. 여러 가지 신경망을 사용할 수 있지만 데이터 클러스터링 기능이 우수한 Kohonen의 SOM (self-organizing map)[7]을 이용하는 경우가 많다. SOM이 다른 탐색 방법들에 비하여 갖는 특징으로는 데이터의 분포특성을 사전에 알지 못하는 경우에도 사용될 수 있고, 비교사 학습 기능을 가지고 있으며 또 통계수치 데이터들을 자연적으로 쉽게 분석할 수 있다는 것이다. 본 논문에서는 개선된 SOM을 이용하여 on-line 데이터 탐색을 효율적으로 수행하는 방법에 대하여 설명한다. SOM이 성능을 발휘하기 위해서는 학습용 샘플 자료를 통한 반복 학습이 선행되어야 하므로 이러한 방법은 on-line에서는 만족하기 어렵다. 현재 많이 사용되는 또 다른 클러스터링 방법인 K-means[8][9]는 위성 원격탐사를 분류하기 위하여 많이 사용되는 방법[10][11][12][13]으로 데이터 클러스터링 개수와 초기 중심이 미리 정해져야만 수행할 수 있다. 이러한 두 방법의 단점을 보완할 수 있는 개선된 SOM이 on-line 데이터 탐색을 효율적으로 수행할 수 있음을 보인다. 제 Ⅱ장에서는 데이터 탐색에 사용되는 클러스터링 알고리듬들 중에 K-means, on-line K-means에 대하여 간략하게 설명하며 제 Ⅲ장에서는 Kohonen SOM과 개선된 Kohonen SOM에 대하여 설명하고 Ⅳ장에서는 시뮬레이션 실험 방법 및 비교 결과에 대하여 설명하고 제 Ⅴ장에서는 결론을 제시한다.

Ⅱ. 데이터 탐색에 사용하는 K-means 클러스터링 방법 고찰

현재 연구되는 데이터 탐색을 위한 대표적인 클러스터링 알고리듬은 Kohonen SOM과 on-line K-means가 있다. On-line K-means는 K-menas 알고리듬을 on-line에 적용한 것이다. K-means는 또 C-means라고도 칭해지며 알고리듬의 간결성으로 인하여 영상 및 음성처리 분야등 다양한 분야에서 응용되어 왔다. 그러나 알고리듬의 원활한 수행을 위하여 초기에 클러스터링해야 할 개수를 미리 정해야하고 또 클러스터의 초기값에 따라 클러스터링된 결과의 수렴성이 달라지는 단점이 있다. K-means알고리듬은 m개의 데이터를 K개로 나눌 때 그 목적함수가 최소가 되도록 나누는 방법이다. Euclidean 거리 척도를 사용한다면 목적함수 E는

다음과 같다

$$E = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \| x_{ij} - z_i \|^2 ,$$

$$\text{여기서 } m = \sum_{i=1}^{K} n_i \tag{1}$$

여기서 K는 나누어야 할 클러스터의 개수를, m은 전체 데이터 개수를, x_{ij}는 i번째 클러스터의 j번째 데이터를, z_i는 i번째 클러스터의 중심을, n_i는 i번째 클러스터에 속한 데이터의 개수를 나타낸다. 목적함수 E를 최소화시키는 z_i를 구하기 위해 K * m 크기의 소속 행렬 U를 정의한다.

$$u_{ij} = \begin{cases} 1 \text{ if } \|x_j - z_i\|^2 \le \|x_j - z_l\|, \text{ for each } l \neq i \\ 0 \text{ otherwise} \end{cases}$$

$$\sum_{i=1}^{K} u_{ij} = 1, \ \forall j = 1, \ldots m \text{ and } \sum_{i=1}^{K} \sum_{j=1}^{n_i} u_{ij} = m \tag{2}$$

여기서 x_j는 m개의 데이터중 j번째 데이터를 나타낸다. u_{ij}가 1이면 데이터 x_j는 i번째 클러스터에 소속된다. 소속 행렬 U가 정해지면 클러스터의 중심들 z_i는 다음 식으로 구한다.

$$z_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \tag{3}$$

실제 알고리듬은 다음 순서로 진행된다.

1단계 : 클러스터의 개수 K와 중심들을 초기화한다.(통상 데이터 범위내의 임의 값을 지정하거나 샘플링함)

2단계 : (2)식을 사용하여 소속행렬 u_{ij}를 구한다.

3단계 : (1)식을 이용하여 목적함수를 구한다. 목적함수의 값이 오차 허용치보다 작거나 그 변화치의 개선도가 지정 문턱치보다 낮으면 종료한다.

4단계 : (3)식을 이용하여 클러스터의 중심들을 다시 구한다. 그리고 2단계로 가서 반복한다

기술된 K-means알고리듬은 반복 수행을 거치기 때문에 전체 데이터를 가지고 있어야 하므로 off-line데이터만 클러스터링 할 수 있다. 그러므로 실시간 즉 on-line으로 들어오는 데이터를 클러스터링 하기는 어렵다. 그러나 승자 클러스터를 얻는 과정을 변경하여 on-line 데이터를 클러스터링 할 수 있도록 할 수 있다. On-line K-means 알고리듬의 수행순서는 다음과 같다.

1단계 : 클러스터의 개수 K와 중심들을 초기화한다.

2단계 : j번째 on-line 데이터 x_j 와 가장 가까운 z_i를 구한다.

3단계 : 승자 클러스터의 중심 zi를 개선시킨다. 개선치는 $\triangle z_i = \varepsilon(x_j - z_i)$으로 계산한다. 여기서 ε는 학습 비율로서 학습이 진행될수록 값이 적어진다.

4단계 : 더 이상의 입력이 없을 때까지 2단계로 계속 넘어간다.

이 on-line알고리듬에서의 학습 방법은 신경망에서 사용하는 학습 방법과 동일하나 학습 범위가 승자노드에 국한된다. 그리고 적절한 ε를 선택함에 따라 수렴 또는 발산하게 되므로 이 값은 많은 실험을 통하여 선택되어야 한다. 또 of-line K-means처럼 반복하여 수렴시키지 않으므로 초기 클러스터의 중심 값 선택도 수렴에 매우 중요하다. 그러므로 입력되는 데이터 집합의 발생 확률 분포에 따른 초기값 선택을 할 필요가 있다. 임의 값을 초기값으로 할 경우에는 데이터 샘플수가 충분하지 않으면 대부분의 경우 국지최소에 빠져 만족할 수 있는 결과를 얻기가 어렵다.

Ⅲ. Kohonen SOM과 개선된 Kohonen SOM을 이용한 클러스터링 방법

SOM은 Kohonen에 의해 제시된 비교사 학습 신경망으로 다차원의 입력 데이터를 유사한 특성을 가진 데이터들이 인접하도록 전사(mapping)시키는 기능이 뛰어난 알고리듬이다. Kohonen은 SOM을 이용하여 음소 타자기의 가능성을 보여주어 신경망 이론이 실제로 유용할 수 있음을 알렸다. 뛰어난 클러스터링 기능으로 인하여 타 신경망에 비하여 가장 많이 실생활에 적용되어 왔다. SOM은 출력노드와 입력사이의 거리가 최소되는 출력노드를 승자로 선정한다는 방식에 있어서는 K-means와 매우 유사한 관계가 있다. 그러나 K-means에서는 승자 클러스터만을 다시 학습시키는 반면에 SOM은 그 주변까지도 학습을 수행한다는 차이점이 있다. 또 K-means는 초기 입력으로 클러스터의 개수 및 초기값들이 필요하지만 SOM에서는 클러스터의 개수에 대한 엄격한 규제는 존재하지 않는다. SOM의 경우도 효율적인 분류를 위해서는 출력 노드의 개수가 클러스터의 개수보다 커야 하고 입력 데이터

의 차원이 커지면 분류해야 할 데이터 특성에 따라 weight벡터 초기값을 설정할 필요가 있으며 학습비율 상수를 적절히 선택해야 한다. Kohonen의 SOM 알고리듬의 수행 순서는 다음과 같다.

1단계 : 입력 벡터 x 와 모든 weight 벡터 w_i 중에 가장 유사한 승자 노드를 구한다.

$$||x - w_c|| = \min||x - w_i|| \quad \text{for} \forall i \tag{4}$$

여기서 w_c는 승자노드의 weight벡터이다.

2단계 : 승자노드와 그 주변노드의 weight벡터를 학습시킨다.

학습 방법으로는 승자 주변 범위 NBc내의 weight벡터들을 아래 식으로 개선시킨다. 여기서 η는 학습 비율 상수로 학습이 진행될수록 그 값이 줄어드는 양의 실수이다. 또 NBc도 학습을 진행할수록 범위가 줄어든다.

$$\triangle w_i = \eta(x - w_i), \quad i \in NB_c \tag{5}$$

영향을 받는 주변 범위는 주변학습 함수로 표현될 수 있으며 통상 가우시안 함수를 사용한다. weight벡터 학습식은 (식 7)을 사용한다. 주변학습 함수에서 p_i는 i번째 노드의 위치를 ,p_c는 승자노드의 위치를 나타내고 σ는 영향을 미치는 범위를 설정한다.

주변학습함수

$$\Omega_c(i) = \exp\left(\frac{-||p_i - p_c||^2}{2\sigma^2}\right) \tag{6}$$

학습치

$$\triangle w_i = \eta\Omega_c(i)(x - w_i) \tag{7}$$

그러므로 SOM을 사용하기 위해서는 학습 진행율을 조정하는 상수 η와 NBc 또는 η와 σ를 학습 횟수마다 적절히 줄여 주어야 하며 선택치의 적절도에 따라 학습 진행시 안정된 영역으로 수렴하거나 국지 안정 영역에 빠질 수 있다. 또 SOM은 반복 학습을 시켜야 하므로 실시간 on-line으로 들어오는 데이터를 학습, 탐색하는데 적용하기가 어렵다. 이러한 단점을 보완하여 on-line에도 적용 가능하도록 SOM의 학습 성능을 개선시켰다. 개선된 SOM은 Hard C-means와 유사한 학습을 하게된다. 즉 기본적인 골격은 SOM을 유지하며 학습방법 즉 학습진행율 조정 상수를 변경하여 SOM이 Hard C-means와 유사한 기능의 클러스트링툴로 작용하도록 한다. 그러기 위해서는 SOM의 출력노드들은 각각의 자신의 우승회수 즉 hit ratio를 기억할 수 있는 메모리 H_i를 가진다.

1단계 : SOM의 1단계와 동일하다.

2단계 : 학습방법은 승자노드와 그 주변 노드를(주변학습함수가 T_h보다 큰 경우) 다음 식으로 개선시킨다.

$$\text{if } \Omega_c(i) > T_h$$

$$w_i = (w_i * H_i + x)/(H_i + 1) \tag{8}$$

(식 8)은 승자노드와 정해진 범위내의 노드들이 현재 자신이 기억하고있는 우승회수와 현재의 weight에 따라 다음 weigth를 계산하며, 그 계산 방식은 K-means와 유사하게 평균을 취하는 개념으로 수행하게 된다. On-line학습에 적용하기 위한 개선으로 인하여 SOM의 학습회수에 따라 변경되는 학습비율 상수 η를 제거하므로 이 상수의 선택값에 따라 국지 안정 또는 진동하는 불안정된 상태로 가는 것을 줄일 수 있으며 Hard C-means[16]와 유사한 방법으로 학습하므로 수렴성과 응용성이 우수하나 주변 노드 학습 영역을 결정하는 고정된 상수 T_h 설정이 필요하다.

Ⅳ. 실험 방법 및 결과

On-line 데이터 탐색 성능을 검증하기 위하여 시뮬레이션을 수행하였다. 시뮬레이션에 사용된 입력 데이터로는 우선 시스템의 특성을 파악하기 위하여 유니폼 분포를 갖는 2차원 데이터 샘플에 대하여 알고리듬을 적용하여 보았다. 그림1은 원 데이터 분포를 보여준다. 원데이터를 클러스터링하기위한 방법으로 클러스터의 개수가 5개인 경우에 반복하여 클러스터링하는 K-means와 반복 학습시킨 SOM 및 on-line즉 1회 수행시킨 on-line K-means와 개선된 SOM을 사용하였다. 그림 2는 K-means를 적용시킨 결과이며 초기 클러스터의 값은 데이터 분포 확률 구조를 파악하여 선정하였으며 반복 횟수는 평균 30회다. 초기 값을 임의로 한 경우에는 수렴을 위한 반복 횟수가 늘어나고 그리고 수렴된 성능에서도 차이를 보이므로 초기값 선정이 중요하다는 것을 알 수 있다. 그림 3은 SOM을 적용시킨 결과를 보

여주며 초기 1회 반복 수행한 결과이다. 그림 4는 15회 학습후의 결과로 학습이 진행됨에 따라 분류 성능은 좋아지는 것을 보여 준다. 출력 유닛은 이차원으로 8*8 크기이고, 초기 weight값은 임의 값, 초기 학습상수 값은 0.5, 초기 주변학습 범위는 크기의 1/3 이다. 그림 5는 on-line K-means를 적용시킨 결과를 보여 준다. 초기 클러스터의 중심값을 임의로 정할 경우에 1회 반복으로 수렴은 거의 불가능하나 데이터 분포를 고려한 초기값 설정에서는 비교적 잘 수렴하며 적절한 학습치에 따라 분류 성능이 좌우된다. 초기 학습상수는 0.5이다. 그림 6은 본 논문에서 제시한 개선된 SOM을 적용시킨 결과를 보여 준다. 반복 횟수는 1회로 사전 학습 없이 데이터가 입력되면 학습하면서 바로 분류된다. 상수 T_h는 0.5로 하였고 초기 weight벡터는 임의 값으로 하였다. 30회 반복 학습하여 최종 분류한 K-means 보다 약간의 데이터 분류 오류가 있지만 입력 데이터 차원이 낮을 때는 on-line K-means와 분류 성능이 유사한 것을 볼 수 있다. 두 번째 실험에서는 대부분의 클러스터링 또는 패턴인식 교재에서 실험 데이터로 사용하는 아이리스 식물 데이터를 on-line으로 분류시키는 데 적용시켜 보았다. 이 실험 데이터는 미리 전문가에 의해 3개의 클러스터로 분류되어있는 데이터로 새로 고안되거나 기존의 클러스터링 알고리듬들의 성능을 평가하기 위한 데이터로 사용된다. 총 데이터 개수는 150개이고 데이터 차원은 4이며 각각의 데이터는 0부터 2까지 3개의 클래스로 기분류된 값을 가지고 있다. 표 1은 k-NN을 이용하여 분류시켰을 경우의 혼돈 행렬(Confusion matrix)로 iris 데이터에 첨부되어 있는 값으로 ESPRIT(European Strategic Program for Research and development in Information Technology)의 기초연구인 ELENA(Enhanced Learning for Evoluative Neural Architecture)보고서[14][15]에 나온 값이다. 표 2는 on-line K-means로 분류한 경우, 표 3은 본 논문에서 제시한 개선된 SOM을 이용하여 on-line으로 분류시킨 경우의 혼돈 행렬값을 나타낸다. 출력노드의 크기는 5*5이며 상수 T_h는 0.5로 하였다. 혼돈 행렬은 분류 알고리듬의 평가 방법 중에 하나로 분류할 클래스의 크기와 같은 행과 열을 가진다. 행렬의 주축 성분들은 그자신의 클래스에 속하는 확률을 백분율%로 표시한 것이고 주축 외의 성분들은 다른 클래스로 오분류될 확률을 백분율로 표시한 것이다. iris 데이터 분류의 경우 데이터 차원이 낮아 on-line 수행에서도 비교적 좋은 결과를 보여주며 특히 on-line K-means 보다 본 논문에서 제시한 개선된 SOM을 사용한 방법이 더 뛰어남을 알 수 있다. 세 번째는 인공위성에서 수신된 다색채널의 영상 데이터를 적용시켜 보았다. 이 데이터 역시 클러스터링 또는 데이터 분류기의 성능을 평가하기 위해 사용되는 데이터로 원 영상은 호주 지역을 찍은 LandSat위성의 다색채널 영상이다. 실험에 사용되는 영상의 크기는 82*100이고 채널의 수는 가시영역 2채널 및 적외선 영역 2채널로 총 4개의 채널이 있다. 이 영상에서 7가지의 토지 형질을 구분하는 것으로 전문가의 현지 답사에 의해 7가지 형질로 분류되었으나 6번째 클러스터는 분류 확인의 불확실성으로 인해 제외되고 실제는 6개의 클러스터로 분류하는 것이다. 하나의 데이터는 3*3 픽셀 영상 및 4색 채널로 구성되므로 총 36개 차원이 된다. 전체 분류해야 할 데이터 개수는 6,435개로 그 중 4,435개가 학습용으로 2,000개가 분류용으로 사용되나 본 실험에서는 6,435개를 on-line으로 분류시켰다. 전체 데이터를 랜덤하게 편집하여 분배하였으므로 실험 데이터에서 82*100크기의 원 영상으로 다시 복원하는 것은 불가능하여 부득이 원 영상 데이터를 복원하지 못하였다. 표 4는 역시 ELENA 보고서에서 인용된 데이터에 첨부되어 있는 값으로 k-NN으로 분류된 경우의 혼돈 행렬을 나타내고 표 5는 on-line K-means로 분류한 경우로 초기 값은 6,435개의 데이터를 각각의 클래스 즉 토질형태에 따라 평균을 취한 값으로 하였다. 이러한 초기 값에도 불구하고 클래스 상호간의 간섭이 심한 것을 볼 수 있다. 표 6은 본 논문에서 제시된 방법에 따른 혼돈 행렬을 나타낸다. 출력노드의 크기는 13*13이고 상수 T_h는 0.5이다. 개선된 SOM역시 데이터의 차원이 커지면 on-line 분류시 오차가 크지는 것을 알 수 있다. 그러나 본 논문에서 제시한 방법이 on-line K-means보다 성능이 뛰어남을 알 수 있었다.

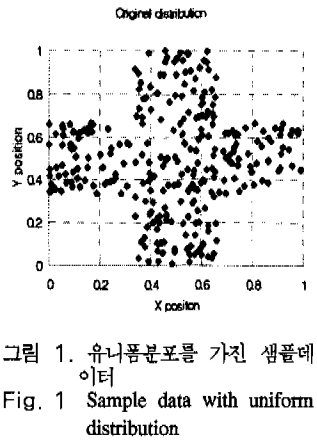

그림 1. 유니폼분포를 가진 샘플데이터
Fig. 1 Sample data with uniform distribution

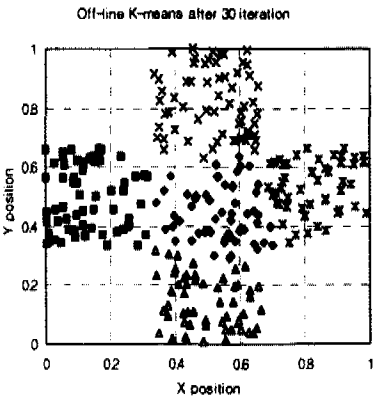

그림 2. K-means로 분류한 결과(30회 반복)
Fig. 2 Classification by K-means (30 iteration)

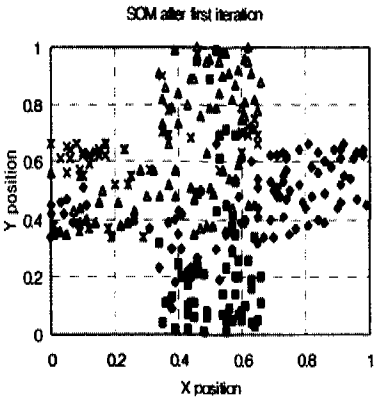

그림 3. SOM을 사용한 분류 (1회째)
Fig. 3 Classification by SOM (after first iteration)

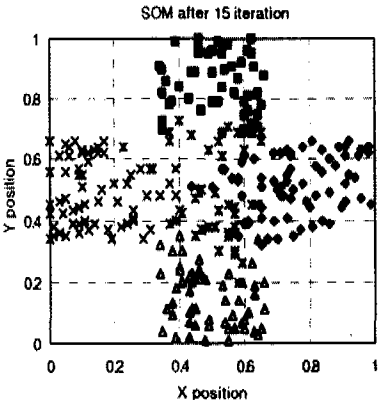

그림 4. SOM을 사용한 분류 (15회 반복)
Fig. 4 Classification by SOM (after 15 iterations)

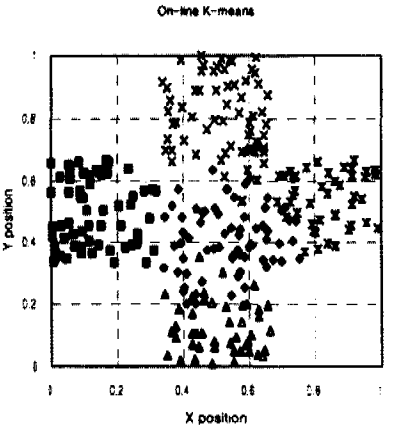

그림 5. On-line K-means에 의한 분류
Fig. 5 Classification by on-line K-means

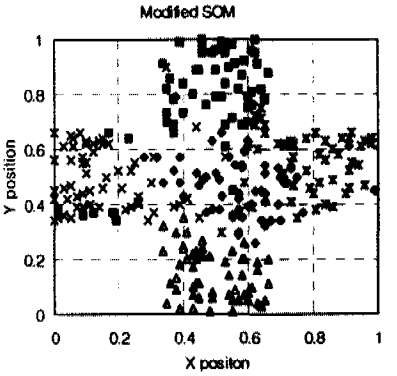

그림 6. 개선된 SOM에 의한 분류
Fig. 6 Classification by modified SOM

V. 결론 및 향후 연구방향

과학 기술이 발전함에 따라 발생되는 정보량은 증대하여 실시간으로 인간이 조사하여 분석하는 일은 점점 불가능해진다. 본 논문에서 제시한 개선된 SOM을 이용한 데이터 탐색 방법은 이러한 분야에 응용될 수 있는 알고리듬으로 SOM의 뛰어난 분류 기능을 유지하면서 학습과정이 필요 없고 on-line에 적용할 수 있도록 하였다 기존에 사용되는 on-line K-means방법에 비해 효율적으로 실시간 데이터에 적용시킬 수 있으며 시뮬레이션 결과 수렴 성능도 뛰어남을 알 수 있었다. 그러나 입력 데이터 차원에 따른 적절한 출력유닛 크기 선정 및 학습범위 상수가 기존의 SOM 경우처럼 수렴 및 분류의 최적화에 영향을 미치므로 이들을 자동적으로 설정할 수 있는 연구가 필요하다.

표 1. k-NN을 사용한 iris 데이터 분류 혼돈 행렬
Table 1. Confusion matrix of iris by k-NN classification.

Class	0	1	2	개수
0	100.0	0.0	0.0	50
1	0.0	96.0	4.0	50
2	0.0	6.0	94.0	50

표 2. On-line K-means를 사용한 iris 데이터 분류 혼돈 행렬
Table 2. Confusion matrix of iris data by on-line K-means classification

Class	0	1	2
0	98.0	2.0	0.0
1	2.0	90.0	8.0
2	2.0	10.0	88.0

표 3. 개선된 SOM을 사용한 iris 데이터 분류 혼돈 행렬
Table 3. Confusion matrix of iris data by modified SOM classification

Class	0	1	2
0	100.0	0.0	0.0
1	2.0	90.0	8.0
2	0.0	2.0	98.0

표 4. k-NN을 사용한 satimage 데이터 분류 혼돈 행렬
Table 4. Confusion matrix of satimage by k-NN classification

Class	1	2	3	4	5	7	개수
1	98.1	0.2	1.1	0.1	0.5	0.0	1,553
2	0.0	96.5	0.1	0.7	2.0	0.7	703
3	0.5	0.1	93.4	4.6	0.0	1.4	1,358
4	0.0	0.8	13.7	70.6	0.8	14.1	626
5	3.1	0.8	0.1	0.8	89.7	5.5	707
7	0.0	0.1	1.9	7.3	2.0	88.7	1,508

표 5. On-line K-means를 사용한 satimage 데이터 분류 혼돈 행렬
Table 5. Confusion matrix of satimage by on-line K-means classification

Class	1	2	3	4	5	7
1	36.3	0.0	54.3	5.4	4.0	0.0
2	0.0	97.6	0.0	1.3	1.1	0.0
3	0.4	0.0	84.6	14.7	0.3	0.0
4	3.2	0.0	0.0	88.3	8.5	0.0
5	24.3	0.1	3.5	0.0	72.1	0.0
7	0.4	1.1	28.6	0.0	0.0	69.9

표 6. 개선된 SOM을 이용한 on-line satimage 데이터 분류 혼돈 행렬
Table 6. Confusion matrix of satimage data by modified SOM

Class	1	2	3	4	5	7
1	93.8	0.3	2.5	0.0	3.0	0.3
2	2.8	89.9	0.0	0.1	6.3	0.9
3	1.9	0.0	92.9	3.3	0.2	1.7
4	2.7	0.0	23.5	42.0	2.1	29.7
5	11.9	4.4	0.7	0.3	64.5	18.2
7	1.1	0.3	3.7	8.0	6.2	80.7

참 고 문 헌

[1] Fayyad, U.M. Data mining and knowledge discovery:making sense out of data. IEEE Expert, pp20-25. Oct. 1996.

[2] Fukunaga,K. Introduction to Statistical Pattern Recognition. Academic Press, Inc., 1250 Sixth Avenue, San Diego, CA 92101, 2nd edition, 1990.

[3] Anil,K. and Richard C. Dubes. Algorithm for Clustering Data, Prentice Hall. 1988.

[4] Honkela, T., Kaski, S.,Lagus,K., and Kohonen,T. Exploration of full-text database with selt-organizing maps. Proceeding of ICNN'96, IEEE International Conference on Neural Networks, Vol1, pp56-61. IEEE Service Center, Piscataway, NJ.

[5] Kaski,S. and Kohonen Kohonen, T. Structures of Welfare and poverty in the world discovered by the self-organizing map. Technical Report A24, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland

[6] Zhang and Li,Y. Self-organizing map as a new method for clustering and data analysis. Proceedings of IJCNN'93, pp2448-2451. IEEE Service Center, Piscataway, NJ 1993.

[7] Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani. Neuro-Fuzzy and Soft Computing. Prentice-Hall International, Inc., USA, pp423-432, 1997

[8] Lloyd,S.P. Least squares quantization in PCM. IEEE Transactions on Information Theory,

IT-28, pp129-137. 1982.

[9] Moody,J. and Darken,C. Fast learning in networks of locally-tunned processing units. Neural Computation, pp281-294. 1989.

[10] P.M. Kelly and J.M. White. Preprocessing remotely-sensed data for efficient analysis and classification. SPIE Application of Artificial Intelligence 1993: Knowledge-Based Systems in Aerospace and Industry. pp24-30, 1993.

[11] Gong P. and P.J. Howarth,. Frequency-based contextual classification and grey-level vector reduction for land-use identification. Photogrametric Engineering and Remote Sensing, Vol.58(4) pp421-437 1992.

[12] Gong P. and P.J. Howarth,. Impreciseness in land-cover classification: its determination, representation and application. The International Geoscience and Remote Sensing Symposium, IGARSS '90, pp. 929-932. 1990.

[13] Mark J. Carlotto. Spectral Shape Classification System for Landsat Thematic Mapper. Proceedings SPIE,, Vol. 2758, Orlando, Florida, 1996.

[14] Jutten,C. Deliverable R3-B4-P Task B4 : Benchmarks, Tech. Rep., ElenaNervesII. Enhanced Learning for Evolutive Neural Architecture. ESPRIT-Basic Research Project No. 6891 June 1995.

[15] Jutten,C. Deliverable R3-B1-P Task B1 : Databases, Tech. Rep., ElenaNervesII. Enhanced Learning for Evolutive Neural Architecture. ESPRIT-Basic Research Project No. 6891 June 1995.

[16] Young Won Lim, Sang Uk Lee, "ON THE COLOR IMAGE SEGMENTATION ALGORITHM BASED ON THE THRESHOLDING AND THE FUZZY C-MEANS TECHNIQUE", Pattern Recognition, Vol.23, No.9, pp.935-952, 1990.

진 영 근(Young-Goun Jin) 정회원

1961년 8월 20일생

1984년 2월 : 한양대 전자공학과 학사

1986년 2월 : 한국과학기술원 전기 및 전자공학과 석사

1992년 3월~현재 : 충남대학교 컴퓨터공학과 박사 과정

1986년 4월~1989년 5월 : 국방과학연구소 연구원

1990년 11월~현재 : 항공우주연구소 근무

<주관심 분야> 패턴인식, 칼라영상처리, 인공지능

김 태 균(Tae-Kyun Kim) 정회원

1944년 1월 25일생

1976년 2월 : 서울대 공업교육학과 학사

1980년 2월 : 동경공업대학 물리정보공학과 석사

1984년 2월 : 동경공업대학 물리정보공학과 박사

1978년~현재 : 충남대학교 컴퓨터공학과 정교수

<주관심 분야> 패턴인식, 영상처리, 멀티미디어