한국지능정보시스템학회논문지 제10권 제2호 2004년 11월(pp.15~37)

시공간 데이터를 위한 클러스터링 기법 성능 비교

강나영
삼성전자
(nayoung.kang@samsung.com)

강주영
이화여자대학교 컴퓨터학과 박사과정
(jykang@kepri.re.kr)

용환승
이화여자대학교 컴퓨터학과 부교수
(hsyong@ewha.ac.kr)

최근 데이터 양이 급증하면서 데이터 마이닝에 대한 연구가 활발하게 진행되고 있으며 특히 GPS 시스템, 감시 시스템, 기상 관측 시스템과 같은 다양한 응용 시스템으로부터 수집된 데이터를 분석하고자 하는 시공간 데이터 마이닝 연구에 대한 관심이 더욱 높아지고 있다. 기존의 시공간 데이터 마이닝 연구들에서는 비시공간 데이터 기반의 일반적인 클러스터링 기법들을 그대로 적용하고 있으나 데이터의 속성이 다른 시공간 데이터 마이닝에서 기존의 알고리즘들이 어느 정도의 성능을 보장하는지, 데이터의 시공간 속성에 따라 적절한 마이닝 알고리즘을 선택하기 위한 기준이 무엇인지 등에 대한 연구는 미흡한 실정이다.

본 논문에서는 기존의 시공간 데이터 마이닝 연구에서 일반적으로 많이 사용되어 온 알고리즘인 SOM(Self-Organizing Map)을 기반으로 시공간 데이터 마이닝 모듈을 개발하고, 개발된 클러스터링 모듈의 성능을 K-means과 두 가지 응집 계층(Hierarchical Agglomerative) 알고리즘들과 균질도, 분리도, 반면영상 너비, 정확도의 네 가지 평가 기준을 기반으로 비교하였다. 또한 입력 데이터의 특성 가시화 및 클러스터링 결과의 정확한 분석을 위해 시공간 데이터 클러스터링을 위한 가시화 모듈을 개발하였다.

논문접수일 : 2004년 2월 게재확정일 : 2004년 10월 교신저자 : 용환승

1. 서론

최근 위성 시스템, 자연 과학 관측 시스템(natural science observation systems), 교통 시스템이나 모니터링 시스템 등과 같은 다양한 과학 기술 응용 도메인으로부터 수집된 방대한 양의 시공간 데이터를 효율적으로 분석하고자 하는 시공간 데이터 마이닝에 대한 관심이 높아지고 있다. 기존의 데이터 마이닝 연구는 문자나 숫자 데이터(alphanumeric data)들을 기반으로 하고 있는 반면, 이러한 시공간 데이터 마이닝의 분석 대상이 되는 데이터는 시간과 공간의 속성을 동시에 지니고 있기 때문에 데이터 분석 시 이와 같은 속성들을 적절하게 고려해 주어야만 한다[1]. 또한 기존의 데이터 마이닝과는 달리 시공간 데이터 마이닝의 경우 지식 탐사의 절차와는 상관없이 입력의 형태나 속성, 도출된 결과의 의미에 대한 적절한 해석을 더욱 중요하게 다루어야 한다는 특징이 있다[2]. 이러한 점들로 미루어볼 때 시공간 데이터의 특성을 고려하지 않은 채 기존의 문자와 숫자 기반의 데이터 마이닝 기법들을 그대로 시공간 데이터 마이닝에 적용하는 것은 그 성능과 결과의 정확성 면에 있어서 한계가 있다고 할 수 있다[1]. 현재까지는 주로 K-means, SOM, 응집 계층 알고리즘과 같은 기본적인 클러스터링 마이닝 알고리즘들을 기반으로 한 시공간

데이터 마이닝 연구가 일반적인데[3,4], 기존의 비 시공간 데이터 대상의 마이닝과 비교해 볼 때 실제적으로 이러한 알고리즘들이 어느 정도의 성능을 보장하는지, 혹은 알고리즘 별로 입력 데이터의 시공간 속성에 따라 그 수행 능력은 어떻게 변화하는지, 입력 데이터의 시공간 속성에 따른 적절한 마이닝 알고리즘의 선택 기준은 무엇인지 등에 대한 연구는 미흡한 실정이다. 그러므로 시공간 데이터 마이닝이 응용 별로 더욱 적합한 마이닝 결과를 도출하기 위해서는 시공간 데이터 마이닝을 위한 적절한 알고리즘, 모델링 기법, 인덱스 및 저장 기법 등의 연구가 필요할 뿐 아니라, 기존의 기법들을 확장하여 적용함에 있어서 시공간 데이터의 특성 및 응용의 속성에 따라 적절한 마이닝 기법을 선택적으로 적용할 수 있도록 객관적인 기준을 제시할 필요가 있다.

본 논문에서는 시공간 데이터 마이닝에 대한 선행 연구들에서 일반적으로 사용되어 온 알고리즘들 중 패턴 인식과 클러스터링 능력이 뛰어나다고 알려진 SOM에 대해 분석하고, 이를 수정하여 시공간 데이터를 기반의 클러스터링을 수행하는 마이닝 모듈을 개발하였다. 데이터베이스의 벤치마킹 통합 시스템인 GSTD(Generate Spatio-Temporal Data) 툴[5] 에서 제공하는 시공간 데이터 생성 모듈을 이용하여 실험 데이터를 생성하고 이 데이터를 기반으로 개발된 SOM 모듈과 다른 세 가지 클러스터링 알고리즘에 대한 성능 평가 및 비교 작업을 수행하였다. 시공간 데이터 마이닝에서 SOM의 성능에 대해 실제적이고 객관적인 기준을 제시하기 위하여 본 논문에서 개발한 SOM 모듈과 K-means, 평균연결 법(Average Linkage Method), Ward 방법의 세 가지 알고리즘들의 클러스터링 결과를 균질도(homogeneity), 분리도(separation), 반면영상 너비(silhouette width), 정확도(accuracy)의 네 가지 평가 기준을 기반으로 비교하였다. 비교 대상이 된 세 알고리즘은 Insightful사의 통계 분석 및 마이닝 프로그램인 S-PLUS[6]의 마이닝 모듈을 사용하였다. 일반적인 문자 및 숫자 데이터의 클러스터링 결과의 분석을 위해서는 위와 같은 통계치를 기준으로 하여 평가하는 것이 일반적이나 시공간 데이터의 경우 입력 데이터의 시공간 속성에 따라 평가 기준 수치가 클러스터링 결과의 정확성 및 성능을 적절히 나타내지 못하는 경우가 있다. 이러한 점을 고려하여 본 연구에서는 시공간 데이터 마이닝을 위한 가시화 모듈을 개발하고 이를 통해 클러스터링 결과를 좀 더 정확하게 비교 분석 하였다.

2. 관련연구

현재까지 여러 가지 시공간 데이터 마이닝 기법이 연구되어 왔으며 또한 기존의 데이터 마이닝 기법의 시공간 데이터에 대한 적용에 대한 연구가 활발히 진행되어 왔으나 각 기법의 객관적인 성능을 나타내어 줄 만한 결과를 제시하는 연구는 미흡한 실정이다. 시공간 속성을 가지는 데이터에 대한 마이닝 기법의 성능평가 연구로써 고차원 데이터에 대해 MST, CLARANS, CURE 등의 알고리즘의 성능을 비교하는 연구[7]가 이루어진 바 있으나 이동성이 없는 단순한 위치 값만을 가진 객체 데이터를 대상으로 하고 있다. 대부분의 클러스터링 기법 비교 연구는 유전자 서열을 분석하는 분야에서 주로 이루어져 왔으며[8,9] 이동 객체 데이터를 기반으로 하는 시공간 데이터 마이닝 기법으로써의 클러스터링 기법의 성능 평가에 대한 연구는 부족한 실정이다. 본 절에서

는 시공간 데이터를 분석하기 위한 기법으로 현재까지 제안되고 적용되어 온 기법들에 대해 살펴보고 그 중 클러스터링 기법의 성능 평가에 대한 관련 연구들을 살펴본다.

2.1 시공간 데이터 마이닝

시공간 데이터는 시간에 따라 변화하는 기하학 객체들에 대한 데이터로 이산적이거나 연속적인 위치 정보를 모두를 포함할 수 있다. 객체 공간에서의 위치 자체만을 고려하는 경우 데이터는 이동 점으로 나타내어질 수 있는 반면 이동 지역으로 나타낼 경우 데이터는 증가하거나 줄어드는 객체의 크기에 관한 정보도 포함할 수 있다. 이러한 데이터들은 데이터를 공간과 시간의 흐름상에 위치시킬 수 있는 거리 속성 및 시간 속성을 갖는다는 점에서 비시공간 데이터와 다르다[10]. 지금까지 오랜 기간 동안 자연 과학 관측 시스템을 통해 이러한 시공간 데이터가 수집되어 왔음에도 불구하고 이를 효율적이고 심도 있게 분석할만한 기술이 부족했기 때문에 최근 데이터 마이닝 기법을 적용한 여러 가지 분석 방법들이 자연 과학 분야에서 활발하게 연구되고 있다. CONQUEST (CONTENT-based Querying in Space and Time)[11], TSA-Ttree[12], Quakefinder[13]와 같은 시스템들은 데이터 마이닝 기법을 기반으로 지구 기상 및 지질학 데이터로부터 사이클론, 지진, 기상 경향, 화산 분출과 같은 활동을 예측해내는 시스템으로 자연 과학 분야의 대표적인 시공간 데이터 마이닝 시스템들이다. 이러한 분야 이외에 이동 객체의 움직임을 분석, 추적하는 감시 시스템(surveillance system)이나 모니터링 시스템, GPS 시스템 혹은 교통 시스템 등에서 생성되는 위치 혹은 궤적(trajectory) 데이터를 데이터 마이닝 기법을 기반으로 분석하고자 하는 연구가 활발히 진행되고 있다. 이러한 기존 연구들에서는 GPS 데이터의 클러스터링 결과를 전자 지도의 개선(refinement) 작업에 이용하거나[14], 객체 추적 혹은 모니터링 데이터를 이용한 궤적 분류나 궤적 예측에 적용 하는 연구를 수행하였다[3,4,15, 16,17].

위와 같은 기존의 시공간 데이터 마이닝에 대한 연구는 접근 방법에 따라 크게 두 가지로 나누어 볼 수 있다. 첫 번째는 K-means나 신경망 등의 클러스터링 기법을 기반으로 하는 방법이고, 두 번째는 기존의 시계열 분석 방법에서 사용되어왔던 연관 규칙 기반의 패턴 탐사(pattern discovery) 기법을 확장한 방법이다. 현재까지는 첫 번째 방법인 클러스터링 기반의 시공간 마이닝 기법에 대한 연구가 더 활발히 이루어지고 있으며 본 논문에서 사용한 K-means, SOM 그리고 응집 계층(agglomerative hierarchical) 알고리즘이 일반적으로 가장 많이 사용되어 왔다. 실시간 추적 시스템을 통해 얻어진 데이터를 분석하여 객체 움직임의 패턴을 찾아내는 작업이나[16], 가상 현실 시스템에서 사용자의 수신호(手信號)를 입력하기 위해 착용하는 햅틱(haptic) 장갑과 같은 입력 장치로부터 수집된 데이터에 대해 K-means 알고리즘을 이용하여 클러스터링 작업을 수행하여 사용자의 수신호를 정확히 인식하고자 하는 연구가 수행된 바 있다[18]. Neil Johnson과 David Hogg [3], Jonathan Owens와 Andrew Hunter[4]는 고차원 벡터 데이터를 2차원 위상 구조로 대응시키는 SOM의 특징을 기반으로 하여 감시 시스템을 통해 수집된 이동 객체 데이터의 다차원 벡터 값을 클러스터링 함으로써 객체의 경로 분류나 예측이 가능한 시공간 데이터 마이

닝 시스템을 개발한 바 있다[3]. 또한 다양한 통계 기법이나 베이지안 네트워크, K-NN등과 같은 학습 기법들을 기반으로 한 궤적 클러스터링이나 자동차 주행 시간 예측 시스템 등에 대한 연구가 진행되어 왔다[17,19,20].

패턴 탐사 방법은 예전부터 시계열 데이터 및 시퀀스 데이터의 분석에 자주 이용되어 왔다. 환자의 관절 운동을 물리적으로 지원하기 위한 아이소키네틱(isokinetic) 기계로부터 수집된 데이터를 분석하여 환자의 부상 여부를 판단하고 근육 상태 진단하며, 재활을 돕고, 상해(傷害)를 방지하고, 환자의 물리 치료에 대한 평가와 치료 계획 수립하기 위한 진단 시스템 개발을 위해 패턴 탐사 데이터 마이닝 방법이 사용되었으며[21], 이외에도 이동 객체의 궤적 데이터를 일종의 확장된 시계열 데이터로 간주하고 패턴 탐사 방법을 적용하여 분석한 연구가 수행된 바 있다.

2.2 클러스터링의 성능 평가

클러스터링은 서로 간에 높은 유사도를 가지는 객체들을 같은 클러스터로 그룹화 하는 작업으로, 크게 분할 방법(partitioning method), 계층적 방법(hierarchical method), 밀도 기반 방법(density-based method), 그리드 기반 방법(grid-based method), 그리고 모델 기반 방법(model-based method)의 다섯 가지 카테고리로 나눌 수 있다[22]. 분할 방법으로는 K-means와 PAM, 계층적 방법으로는 BIRCH나 CURE, 밀도 기반 알고리즘으로는 DBSCAN등이 일반적으로 가장 널리 사용된다[22]. 또한 이러한 클러스터링 알고리즘들 이외에도 벡터 양자화를 통해 클러스터링을 작업을 수행하는 SOM(Self-Organizing Map)과 같은 신경망이나 베이지안 네트워크(bayesian network)와 같은 인공 지능의 학습 기법들도 클러스터링 알고리즘의 하나로 사용된다.

이러한 클러스터링 알고리즘들의 성능 평가에 대한 연구는 대부분 통계적인 분석을 기반으로 진행되어 왔다. 특히 클러스터링 결과에 대한 해석의 정확성이 다른 어느 응용 분야에서보다 중요하다고 할 수 있는 유전자 발현 분석 연구 분야에서 주목을 받아 왔는데, [8]에서는 유전자 데이터를 클러스터링 하는 데 있어서 네 가지의 클러스터링 알고리즘의 성능을 균질도(homogeneity), 분리도(separation), 반면영상 너비(silhouette Width), 정확도(accuracy), 중복성 점수(redundant Score), WAPD(데이터의 교란에 대한 클러스터링 결과의 강건성), 클러스터의 크기와 일관성 등과 같은 항목들을 기준으로 비교하였다. 실제 클러스터링 결과의 정확성, 즉 예상 클러스터와 결과 클러스터의 일치도 또한 위의 기준들과 더불어 클러스터링 결과의 평가 기준으로 사용되기도 한다[9].

3. SOM 기반 시공간 데이터 클러스터링 시스템

본 논문에서는 시공간 데이터 마이닝의 선행 연구에서 가장 일반적으로 사용되어 온 SOM의 성능을 객관적으로 평가하기 위해 SOM 기반의 시공간 데이터 클러스터링 시스템을 구현하고 이의 성능을 상용 마이닝 프로그램에서 제공하는 K-means, 평균연결법, Ward 방법의 세가지 알고리즘의 성능과 비교 분석하였다. 본 논문에서 사

용한 4가지 클러스터링 기법의 수행 방법과 차이점은 다음과 같다.

3.1 시공간 데이터 클러스터링 기법

■ K-means

K-means 방법은 객체를 K개의 그룹으로 분할하는 방법으로 수행 과정은 다음과 같다.

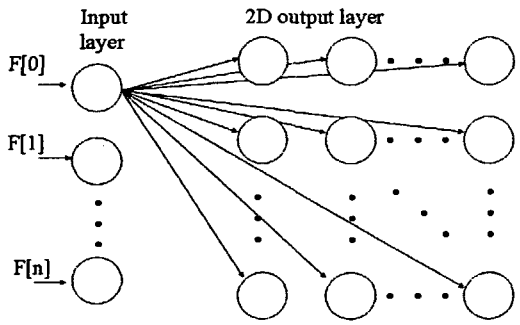
1) 군집의 수 K를 결정한다.
2) 초기 K개 군집의 중심을 선택한다.
3) 주어진 중심점을 기준으로 각 객체를 가장 가까운 군집에 할당한다. 이 때 중심점과 객체간의 거리는 유클리드 거리(Euclidean distance)를 사용해 계산한다.
4) 새로 할당된 객체를 중심으로 각 군집의 새로운 중심점을 계산한다.
5) 만약 기존 중심점과 새로운 중심점간에 차이가 없으면 수행을 중지하고, 그렇지 않으면 2번으로 돌아가 다시 수행한다.

■ 응집 계층 알고리즘

응집 계층 알고리즘은 계층 알고리즘 중 bottom-up 방식의 일종으로써 모든 n개의 객체가 n개의 서로 다른 그룹이라 가정한 후에 그룹간의 유사도에 따라 가장 유사한 두 개의 그룹을 합병해 그룹 수를 줄여 가는 과정을 반복하여 전체 그룹 수가 사전에 지정된 k개가 될 때까지 반복함으로써 k개의 그룹을 찾아내는 방법이다. 응집 계층 알고리즘은 크게 연결법(Linkage Method)과 워드법(Ward Method)으로 분류된다. 연결법에는 단일 연결법(single linkage method), 완전 연결법(complete linkage method), 평균 연결법(average linkage method)과 같은 방법이 있다. 단일 연결법은 클러스터 사이의 거리를 각 클러스터 내에 포함된 객체들 사이의 거리 중 최소의 거리를 기준으로 결정하는 방법이고, 완전 연결법은 각 클러스터 내에 포함된 객체들 사이의 거리 중 최대 거리를 두 클러스터의 거리로 보는 방법이다. 평균연결법은 단일연결법과 완전연결법의 특징을 절충한 알고리즘으로, 각 클러스터 사이의 거리는 각 클러스터 내에 포함된 객체들의 중심 사이의 거리로 계산한다[15]. 일반적인 데이터 마이닝 응용에서는 연결법 중 평균연결법이 주로 사용되어 왔다. 워드법은 전체 군집 내 제곱합을 이용하여 군집의 수를 줄여 가는 방법으로, 군집 수를 줄이기 위해 현재 군집 내의 객체들에 대해 전체 제곱합을 구한 후 이 값이 가장 작은 군집끼리 병합하는 방법이다. 본 논문에서는 S-PLUS 제품의 평균연결법과 워드법을 사용하여 클러스터링을 수행하였다.

■ SOM(Self-Organizing Map)

SOM은 Kohonen이 제안한 신경망 기반의 자기조직화 알고리즘으로 해부학적인 이론에 근거하여 인간의 두뇌 구조를 모델링 한 방법이다. 즉, 인접한 출력노드들은 비슷한 기능을 수행할 것이라고 예측하여, 기존의 경쟁 학습(Competitive Learning)을 개선하여 입력노드와 가장 가까운 출력노드들뿐만 아니라 그 출력노드의 이웃노드들도 함께 학습시키는 알고리즘이다[23]. [그림 1]과 같이 SOM은 기본적으로 2개의 층으로 이루어져 있다. 첫 번째 층은 입력층(Input Layer)이고, 두 번째 층은 2차원의 출력층(Output Layer)이며, 모든 연결은 입력층에서 출력층의 방향으로 연결되어 있다.



[그림 1] SOM의 입출력 노드의 구성도

SOM은 패턴 인식과 클러스터링 능력이 뛰어나기 때문에 현재 진행되고 있는 시공간 데이터 마이닝 연구에 많이 응용되고 있다. SOM이 수행되는 과정은 다음과 같다.

1) K개의 출력노드를 위상공간 내에 배치한다.
2) 학습률 $\alpha(t)$을 초기화한다. 학습률은 0과 1사이의 값을 가지며, 시간이 지남에 따라 감소한다.
3) 새로운 $\chi(t)$입력벡터를 입력노드에 제시한다.
4) 입력벡터와 모든 출력노드들과의 거리를 계산하여 최소거리를 가지는 승자노드를 찾는다. 이 때 거리는 Euclidian 거리로 계산한다.

$$|x(t)-m_c(t)| = \min |x(t)-m_i(t)|$$

5) 승자노드와 이웃한 출력노드들의 가중치를 갱신한다. 이웃 노드의 가중치를 갱신하기 위해서 이웃 함수(Neighborhood Function)를 사용한다. 다음과 같이 이웃 함수 $\lambda_{ci}(t)$로 Gaussain 함수를 사용하여 이웃 노드를 계산한다.

$$m_i(t+1) = m_i(t) + \alpha(t)\lambda_{c_i}(t)[x(t)-m_i(t)]$$

$$\lambda_{ci}(t) = \exp(-\frac{|r_c - r_i|^2}{2\sigma^2(t)})$$

6) 2번으로 가서 반복한다.

본 논문에서는 시공간 데이터에 SOM을 적용하여 클러스터링 하기 위해 SOM 기반 시공간 마이닝 모듈을 개발하였다.

3.4 SOM 기반 시공간 클러스터링 시스템의 설계 및 구현

본 논문에서 개발한 SOM기반 시공간 클러스터링 시스템은 크게 전처리 모듈, 클러스터링 모듈 그리고 가시화 모듈의 세 부분으로 나누어진다. 시스템의 구현 환경은 <표 1>과 같다.

<표 1> SOM 기반 시공간 클러스터링 시스템의 구현 환경

DBMS	Oracle 9i
개발도구 및 언어	Oracle PRO*C Microsoft Visual C++ 6.0 Microsoft Visual Basic 6.0 ADO 2.6 Library

데이터 전처리 모듈과 SOM 모듈은 Oracle Pro*C를 이용하여 Oracle 9i에 저장되어 있는 시공간 데이터를 기반으로 작업을 수행하며 가시화 모듈은 VisualBasic 6.0과 ADO 2.6라이브러리를 사용하여 구현하였다.

3.4.1 전처리 모듈

SOM에 적용하기 위해서는 원시 데이터를 다차원 벡터로 벡터화하는 작업이 필요하다. 데이터베이스에 저장되어 있는 전처리 과정 전의 원시 데이터의 테이블 구조는 <표 2>와 같다.

<표 2> 원시 시공간 데이터 테이블 정의

원시 시공간 데이터 테이블		
필드명	데이터 타입	설명
VALID	VARCHAR(20)	데이터 위치의 유효성 여부
ID	NUMBER(4,0)	이동 객체의 ID
TIME	NUMBER(18,6)	이동 객체가 움직인 시각(타임스템프)
X	NUMBER(18,5)	이동 객체의 위치 - x좌표 값
Y	NUMBER(18,5)	이동 객체의 위치 - y좌표 값

n 프레임(frame)동안 움직인 이동 객체 i는 다음과 같이 n개의 흐름 벡터(Flow Vector)의 집합 Qi로 표현된다.

$$Qi = \{f1, f2, f3, \ldots, fn\}$$

하나의 흐름벡터는 다음과 같이4가지 요소로 이루어진다.

$$f = (x, y, dx, dy)$$

x와 y는 이동 객체의 특정 시간에서의 x 좌표값과 y 좌표값이고, dx와 dy는 객체가 특정 시간에 x축으로 움직인 순간 속도와 y축으로 움직인 순간 속도이다.

본 논문에서는 전처리 모듈을 통해 dx와 dy를 계산한다. 일반적으로 속도는 거리변화를 시간변화로 나누어서 계산하여 절대속도로 구한다. 이렇게 구해진 절대속도는 1보다 큰 값이 될 수도 있으며, 이 경우 SOM에 적용하면 0과 1사이의 값을 가지는 x, y에 비해 1보다 큰 값을 가질 수 있는 순간 속도 dx, dy에 더욱 민감하게 반응할 수 있다. 그러므로 본 논문에서는 dx와 dy를 절대속도가 아닌 상대속도로 계산한다. 즉 dx는 각 데이터에 대해 x축으로 움직인 절대속도를 구한 후 전체 데이터의 최대속도로 나누어서 계산한다. dy 역시 y축으로 움직인 데이터에 대해 동일한 방법으로 계산한다.

3.4.2 SOM 기반 클러스터링 모듈

전처리 모듈을 통해 계산된 데이터는 <표 3>과 같은 구조의 테이블에 저장되어 K-means와 SOM에 적용된다.

<표 3> 시공간 데이터 테이블 정의

시공간 데이터 테이블		
필드명	데이터 타입	설명
ID	NUMBER(4,0)	이동 객체의 ID
TIME	NUMBER(18,6)	이동 객체가 움직인 시각(타임스템프)
X	NUMBER(18,5)	이동 객체의 위치 - x좌표 값
Y	NUMBER(18,5)	이동 객체의 위치 - y좌표 값
DX	NUMBER(18,5)	이동 객체가x축으로 움직이는 순간 속도
DY	NUMBER(18,5)	이동 객체가y축으로 움직이는 순간 속도
CLUSTER	NUMBER(4,0)	결과 클러스터 번호

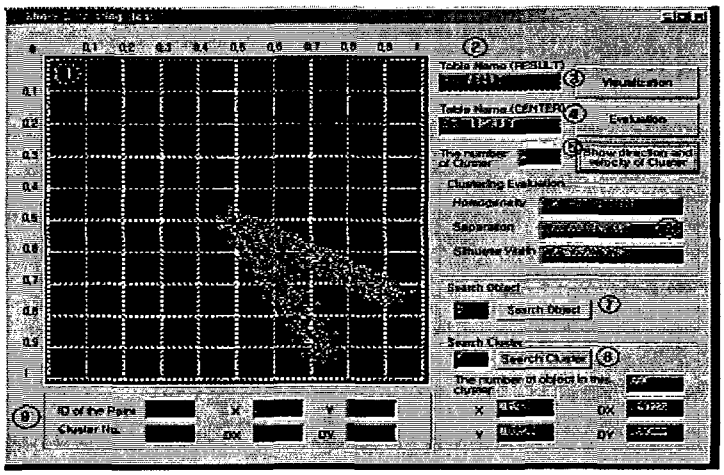
본 논문에서 구현한 SOM 모듈의 동작순서는 다음과 같다. 먼저 데이터베이스에 접속한 후 클러스터 수 K를 결정하고 초기 출력노드의 가중치를 설정하여 네트워크를 형성한다. 그리고 주어진 훈련횟수만큼 네트워크를 훈련시킨 후 출력노드의 최종 가중치의 값을 데이터베이스에 저장한다. 그리고 가시화 모듈에서 언제든지 클러스터링 된 결과를 보여주기 위해 각 데이터에 대한 결과 클러스터 번호 즉, 각 데이터에 대해 최종적으로 선택된 출력노드의 번호를 데이터베이스에 저장한다.

본 논문에서는 입력벡터들 중에서 K개를 임의로 선택하여 SOM의 초기 출력노드의 가중치로 한다. 훈련횟수는 일반적으로 신경망 알고리즘에서 수행되고 있는 훈련횟수인 10000번으로 정한다.

3.4.4 가시화 모듈

본 논문에서는 시공간 데이터의 특성을 고려하여 클러스터링 결과를 보다 쉽고 정확하게 보여주기 위한 가시화 모듈을 구현하였다. 입력 데이터는 2차원 좌표 상에 나타나며 결과 클러스터 별로 색상을 달리 표현하여 클러스터들간의 구분을 눈으로 쉽게 식별할 수 있도록 하였다. 또한 정확하게 특정 클러스터에 속하는 입력 데이터를 정확하게 조회할 수 있도록 검색 기능을 추가하고 객체 검색 기능을 통해 전체적인 클러스터 결과뿐 아니라 사용자가 지정하는 특정 데이터의 이동 궤적을 보여주는 기능을 추가하였다. 그리고 SOM 기반 알고리즘의 클러스터링 결과로 얻어지는 프로토타입 벡터(출력 노드)를 그 방향과 속도에 따라 화살표 모양을 이용해 가시화하여 클러스터의 특성을 쉽게 파악하도록 하고, 각 클러스터링 결과에 대해 본 논문에서 정의한 평가 기준들 중 균질도, 분리도, 반면영상너비의 값을 계산하여 결과 수치를 보여주도록 구현하였다.

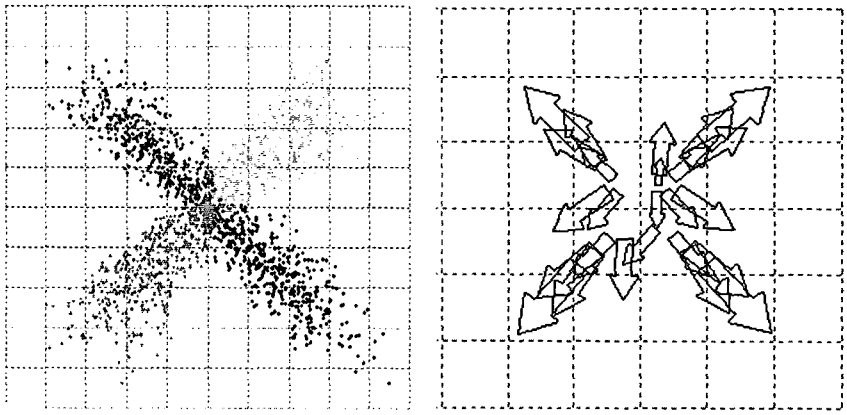
다음의 [그림 3]은 입력데이터와 그에 대해 SOM 클러스터링 결과의 출력 노드 가중치를 가시화 모듈을 통해 가시화해 본 결과이다. 그림에서 보는 바와 같이 마이닝 결과로 얻은 출력노드에 대한 가시화를 통해 클러스터의 개수와 각 클



[그림 2] 시공간 데이터 클러스터링 입력 데이터 및 결과 가시화 모듈

<표 4> 가시화 모듈의 상세 기능

번호	내 용
①	클러스터링 결과를 점 및 화살표 형태로 가시화
②	클러스터링 결과가 저장되어 있는 오라클 테이블의 이름
③	가시화 실행 버튼
④	성능 평가 실행 버튼
⑤	각 클러스터의 방향과 속도를 화살표로 가시화
⑥	균질도, 분리도, 반면영상너비의 성능 평가치
⑦	특정 객체의 ID를 입력하면 객체의 움직임을 시간 순서로 가시화
⑧	특정 클러스터의 입력 데이터를 표시하고 해당 클러스터 내 데이터 개수와 중심점 좌표를 나타냄
⑨	①번 화면상에서 특정 포인트를 선택하면 그 객체의 ID, 소속 클러스터 번호, 객체의 좌표 값을 보여줌



[그림 3] 입력 데이터와 그에 따른 SOM 결과 출력 노드의 가시화

러스터의 특성 즉, 각 클러스터 중심점의 위치와 방향, 속도의 크기를 파악 할 수 있다.

4. 성능 평가

4.1 실험 시공간 데이터 생성기

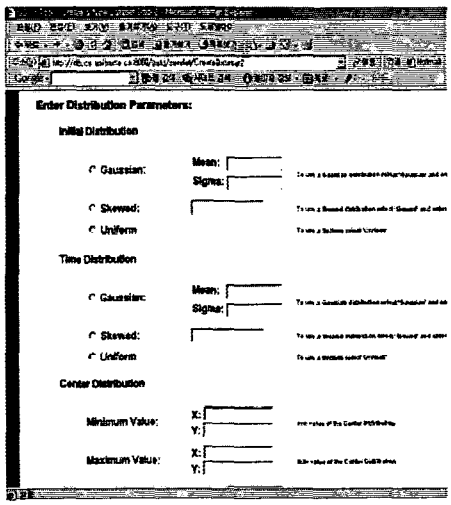
대부분의 시공간 데이터 마이닝 연구들은 기존에 축적되어 온 방대한 양의 자연 과학 데이터 혹은 이동 객체들을 추적하는 감시 혹은 모니터링 시스템으로부터 수집된 시공간 데이터들을 기반으로 하고 있다. 이러한 데이터들은 일반적으로 2차원의 위치 정보와 시간 정보를 포함한 3차원 혹은 그 이상의 표현 형식을 취하고 있는데, 이러한 원시 데이터를 기반으로 정확한 마이닝 결과를 얻기 위해서는 복잡한 사전 처리 작업이 필수적이라고 할 수 있다. 이러한 사전 처리 작업의 부담을 덜고 입력 데이터의 시공간 속성의 변화에 따른 성능 비교 작업을 용이하게 하기 위해 본 논문

에서는 실제 응용 시스템으로부터 얻어진 데이터가 아닌 GSTD라는 시공간 데이터 생성기를 통해 생성한 데이터를 사용하였다. GSTD는 캐나다의 Alberta 대학에서 개발한 시공간 데이터 생성기로써 [그림 4-a]와 같이 웹 사이트 http://db.cs.ualberta.ca:8080/gstd를 통해 서비스를 제공하고 있다. GSTD 생성기는 웹 상에서 사용자가 관련 파라메터를 적절히 조절하여 다양한 형태와 움직임의 시공간 데이터를 그림 4-b와 같이 XML 형태로 생성할 수 있도록 하며 자체적인 가시화 모듈을 통하여 생성된 데이터의 움직임을 확인할 수 있도록 해준다[5]. GSTD는 점 형태의 객체와 직사각형 형태의 데이터를 생성하도록 지원하고 있는데 본 논문에서는 점 형태의 객체를 이용하였다.

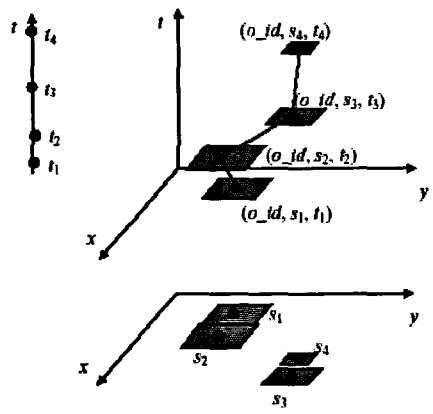
GSTD가 생성하는 데이터는 이동 객체의 ID

```
<?xml version="1.0" encoding="UTF-8"?>
<gstd type="p">
<instance valid="true" id="100" time="0.000000">
        <x>0.51562</x>
        <y>0.49392</y>
</instance>
<instance valid="true" id="101" time="0.000000">
        <x>0.49734</x>
        <y>0.50091</y>
</instance>
<instance valid="true" id="102" time="0.000000">
        <x>0.50638</x>
        <y>0.49850</y>
</instance>
<instance valid="true" id="103" time="0.000000">
        <x>0.48953</x>
        <y>0.49597</y>
</instance>
<instance valid="true" id="104" time="0.000000">
        <x>0.49592</x>
        <y>0.50286</y>
</instance>
<instance valid="true" id="105" time="0.000000">
        <x>0.50295</x>
        <y>0.50903</y>
</instance>
```

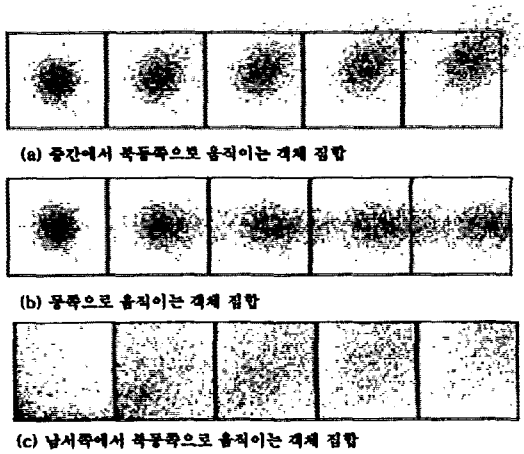
[그림 4-a] GSTD를 통해 생성한 XML 형태의 시공간 데이터



[그림 4-b] GSTD 사이트



[그림 5-a] 단일 객체의 위치 이동



[그림 5-b] GSTD로 생성한 예제 데이터

에 의해 식별되는 ()의 집합이다. 는 시간이 일 때의 객체 O의 위치이다. 이 때 와 는 모두 0과 1사이의 값을 가진다. [그림 5-a]는 이러한 형태의 객체들의 시간의 변화에 따른 위치를 직선으로 연결하여 객체 이동의 연속성을 나타낸 것으로 사용자는 이러한 객체들의 수뿐 아니라, 시작점 분포, 시간 분포, 중심점의 분포 등의 파라메터 값을 조절하여 다양한 시공간 속성을 가진 데이터를 생성할 수 있다. 또한 [그림 5-b]와 같이 생성한 데이터의 이동 모습을 가시화 모듈을 통해 확인할 있다[5].

4.2 실험 데이터

성능 평가를 위해 사용한 실험 데이터는 크게 세 가지 종류로 나누어진다. 첫 번째는 이동 객체 그룹간 방향에 차이를 준 데이터세트 D1~D6, 두 번째로 이동 객체 그룹간의 속도에 차이를 준 데이터세트 D7~D16, 마지막으로 임의의 시공간 속성을 지닌 이동 객체 그룹들의 데이터인 데이터세트 D17의 세 가지 종류, 총 17개의 데이터 세트에 대해 실험하였다. 이동 객체의 방향 차이에 대한 각 알고리즘의 정확도를 측정하기 위해 객체 그룹간의 이동 방향과 출발 지점을 조절한 데이터세트인 D1~D6를 사용하였고, 속도 차이에 따른 클러스터링 정확성을 평가하기 위해 두 개의 이동 객체 그룹간의 이동 방향을 통일하고 속도 차이만을 1.0/s에서 0.1/s로 감소시키며 조절한 데이터세트 D7~D16를 사용하였다. 마지막으로 임의의 시공간 속성을 지닌 이동 시공간 데이터에 대한 클러스터링 결과를 평가하기 위해 다양한 속도로 움직이면서 방향이 서로 다른 객체들을 포함하는 데이터인 데이터세트 D17에 대해 성능 평가를 수행하였다. 테스트 데이터 세트의 특징을 정리하면 다음과 같다.

[그림 6]의 (a)와 (b)는 데이터세트 D4와 D6(두

<표 5> 데이터세트 내의 객체 그룹 간 시공간 속성 차이

<table>
<tr><td rowspan="4">데이터 세트번호
(D1-D17)</td><td colspan="7">데이터세트 내의 객체 그룹 간 시공간 속성 차이</td></tr>
<tr><td colspan="5">방향 차이</td><td>속도 차이</td><td rowspan="2">방향 + 속도
(불규칙한 움직임)</td></tr>
<tr><td></td><td>90도</td><td>75도</td><td>60도</td><td>45도</td><td>30도</td><td>1.0/s~0.1/s</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>출발점 동일</td><td>D1</td><td>D2</td><td>D3</td><td>D4</td><td>D5</td><td rowspan="2">D7~D16</td><td rowspan="2">D17</td></tr>
<tr><td>출발점 비동일</td><td colspan="5">D6</td></tr>
</table>

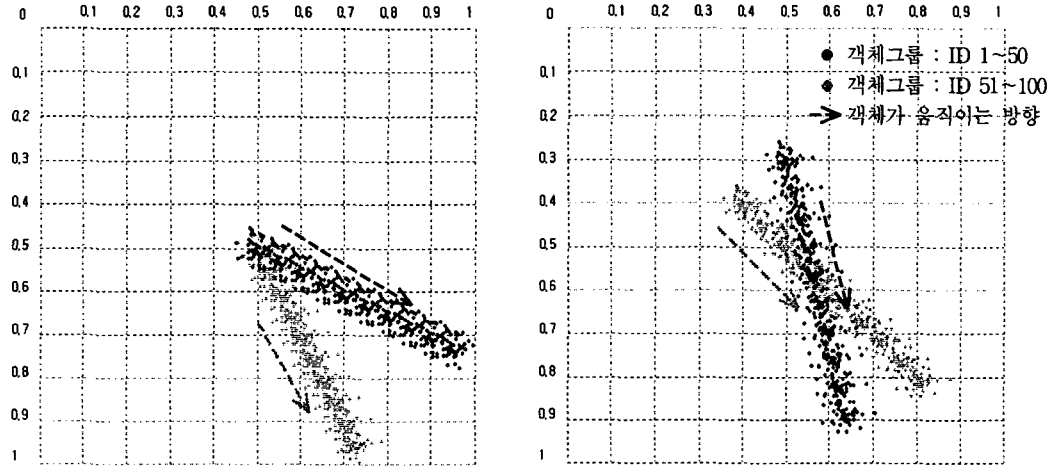
<표 6> 각 데이터세트의 객체 수 및 데이터 포인트 수

	데이터세트 D1~D10	데이터세트 D11~19	데이터세트 20
이동 객체 그룹 수	2	2	4
각 그룹 당 객체 수	50	50	50
총 이동 객체 수	100	100	200
각 그룹 당 포인트 수	500, 500	500, 347	530, 530, 530, 530

객체 그룹간의 방향차이가 45도인 경우)의 입력 데이터를 보여준다.

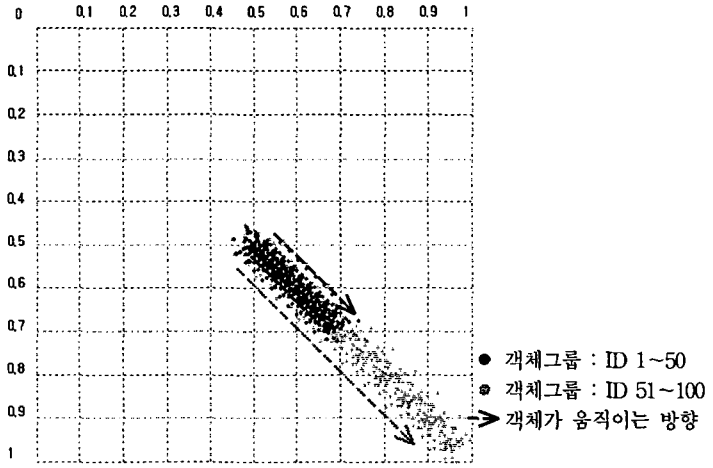
[그림 7]은 데이터세트 D11(두 객체그룹간의 속도차이가0.6/s인 경우)의 입력 데이터의 모습을 보여준다.

[그림 8]은 각 객체들이 객체 그룹 별로 서로 다른 방향을 향해 다양한 속도로 불규칙하게 움직이는 4개의 객체 그룹에 대한 데이터이다. 이 데이터세트 D17의 특징은 하나의 객체 그룹은 하나의 방향을 향하여 움직이는 객체들로 구성되지만 그 객체 각각은 다양한 속도로 불규칙하게 움직인다는 점이다.

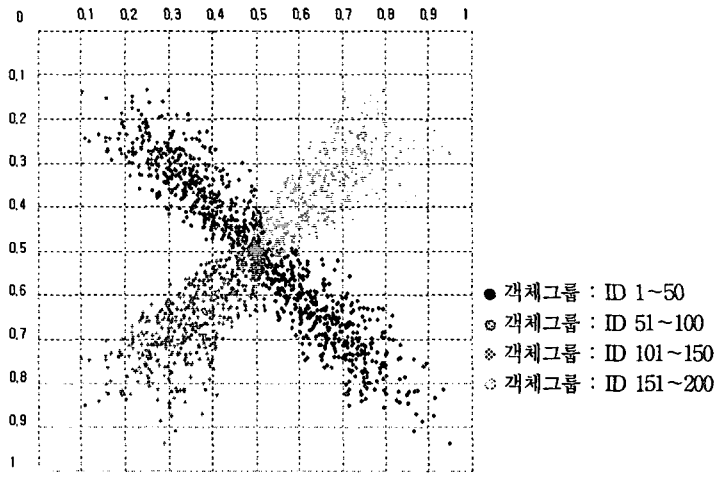


(a) 동일 출발점을 가지는 두 이동 객체 그룹 (b) 비동일 출발점을 가지는 두 이동객체 그룹

[그림 6] 객체의 이동 방향 속성에 차이를 준 시공간 데이터



[그림 7] 객체의 이동 속도 속성에 차이를 준 시공간 데이터



[그림 8] 객체의 시공간 속성이 불규칙한 시공간 데이터

4.3 성능 평가 기준

클러스터링의 성능 평가 즉 유효성 검사(validation)란 수치적, 그리고 객관적인 방식으로 클러스터 분석의 결과를 평가하는 작업이다. 보통 클러스터링의 결과 평가하는 방법은 크게 외적 기준 분석과 내적 기준 분석의 두 가지 방법으로 나뉜다. 외적 기준 평가는 클러스터링의 결과를 대상 데이터 객체를 분할하는 또 다른 최적 기준(Gold Standard)과 비교하는 작업이다. 보통 이러한 최적 기준은 대상 데이터와 독립적인 다른 프로세스를 통해 선택된다. 내적 기준 평가는 입력 데이터의 정보를 기반으로 입력 데이터 집합과 클러스터링 결과 사이의 적합성을 평가하는 방법이다[24]. 클러스터링 결과의 성능은 확장성, 클러스터 모양의 다양성, 분석 대상 데이터의 융통성, 잡음 데이터의 처리등과 같은 다양한 기준에 있어서 평가되어야 한다[22]. 본 연구에서는 이러한 기존의 클러스터링 평가 연구들을 참고하여 다음과 같은 네 가지 항목을 성능 평가 기준으로 선정하였다.

▣ **균질도(Homogeneity)**

균질도는 클러스터의 중심점과 그 클러스터에 속하는 포인트들간의 평균 거리로 계산한다. 균질도는 클러스터 내부의 분산을 나타내며 수치가 클수록 클러스터링이 잘되었다고 판단한다. 균질도 계산식은 다음과 같다.

$$H_{ave} = \frac{1}{N_{point}} \sum_{i} D(p_i, C(p_i))$$

□ D는 거리함수, p_i는 i번째 포인트, $C(p_i)$는 p_i가 속한 클러스터의 중심점, N_{point}는 전체 포인트 수이다.

▣ **분리도(Separation)**

분리도는 클러스터의 중심들간의 평균 거리로 계산한다. 분리도는 클러스터 사이의 분산을 나타내며 수치가 작을수록 클러스터간의 분할이 명확

하다고 판단한다. 분리도를 구하는 식은 다음과 같다.

$$S_{ave} = \frac{1}{\sum_{i \neq j} N_{ci} N_{cj}} \sum_{i \neq j} N_{ci} N_{cj} D(C_i, C_j)$$

C_i와 C_j는 i번째 클러스터와 j번째 클러스터의 중심점, N_{ci}와 N_{cj}는 i번째 클러스터와 j번째 클러스터에 속한 포인트의 수이다.

■ **반면영상 너비**(Silhouette Width)

반면영상 너비는 클러스터링 결과의 전체 질(quality)을 반영하며 클러스터들이 얼마나 컴팩트하며 분리가 잘 되었는가를 나타낸다. 반면영상 너비의 수치가 클수록 클러스터링이 잘되었다고 판단한다. 반면영상 너비를 구하는 식은 다음과 같다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$a(i)$는 i번째 포인트와 이 포인트와 같은 클러스터에 속한 다른 포인트들과의 평균 거리이고 $b(i)$는 i번째 포인트와 가장 근접해 있는 이웃 클러스터에 속한 포인트들과의 평균 거리이다.

■ **정확도**(Accuracy)

정확도는 예상 클러스터와 결과 클러스터를 비교하기 위한 기준이다. 각 포인트들이 얼마나 예상 클러스터에 정확하게 클러스터링 되었는지를 수치로 나타내어 클러스터링 결과의 정확성을 파악한다. 정확도의 수치가 클수록 클러스터링이 잘되었다고 판단하다. 정확도를 구하는 식은 다음과 같다.

$$A_{ave} = \frac{1}{N_{point}} \sum_{i} A_i$$

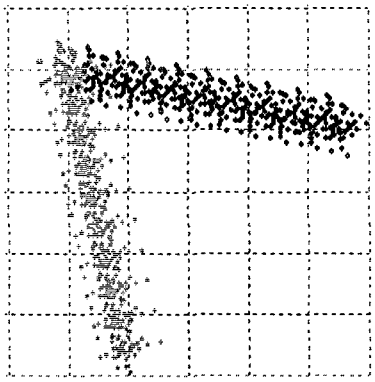
A_i는 i번째 클러스터에 정확하게 클러스터링된 포인트 수이다.

4.3 성능 평가 결과

본 절에서는 앞 절의 평가 기준에 대한 각 데이터세트들의 평가치를 구하고 이를 기반으로 각 알고리즘의 성능 평가 결과를 분석한다. 데이터 그룹간의 방향과 속도를 변화시킨 데이터세트 D1~D6과 데이터세트 D7~16에 대해 알고리즘 별로 정확한 클러스터링 결과를 제공하는 방향과 속도의 임계값(threshold)를 구하여 알고리즘 간의 성능을 비교한다. 임계값은 각 알고리즘의 클러스터링 결과의 정확도가 보장되는 임계 각도와 임계 속도를 나타낸다.

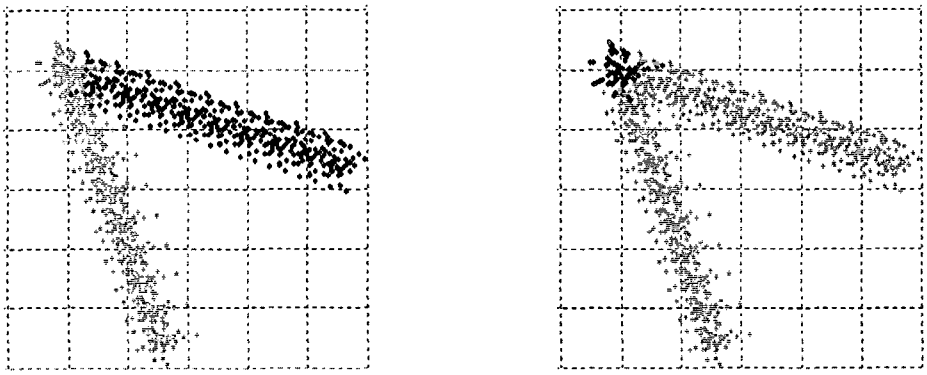
■ **데이터세트 D1~D6**

앞에서 정의한 바와 같이 데이터세트 D1~D5는 비슷한 속도로 서로 다른 방향으로 각도를 좁혀가며 움직이는 객체 집합들이다. [그림 9]에서 볼 수 있듯이 방향 차이가 75도 이상일 때는 모든 알고리즘의 클러스터링 결과가 동일하다. [그림 10, 11, 12]는 방향 차이가 각각 60, 45, 30도일 때 각 알고리즘들의 클러스터링 결과를 보여준다. 방향 차이가 60도와 45도인 경우 SOM과 Ward 방법은 비교적 클러스터링 결과가 정확하지만 K-means과 평균연결법은 두 객체 그룹을 제대로 클러스터링 하지 못한다는 것을 알 수 있다. 객체 그룹간의 방향 차이가 30도인 경우는 SOM만이 정확하게 클러스터링을 수행함을 확인할 수 있다.



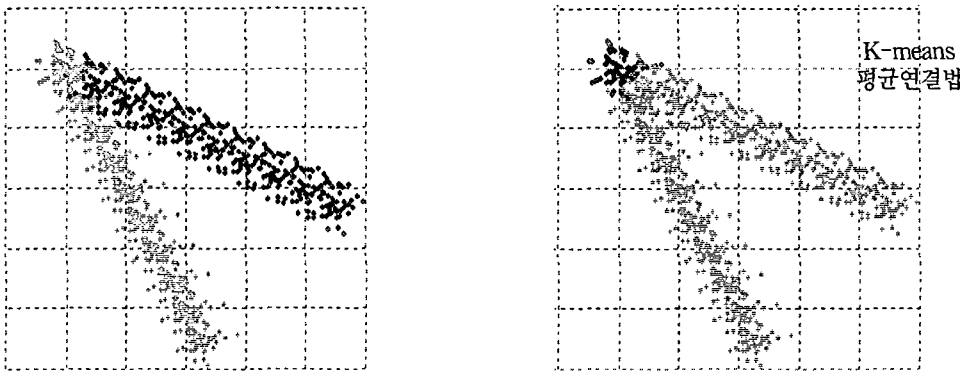
SOM, K-means, 평균연결법, Ward 방법

[그림 9] 방향 차이가 75 도일 때 클러스터링 결과의 가시화



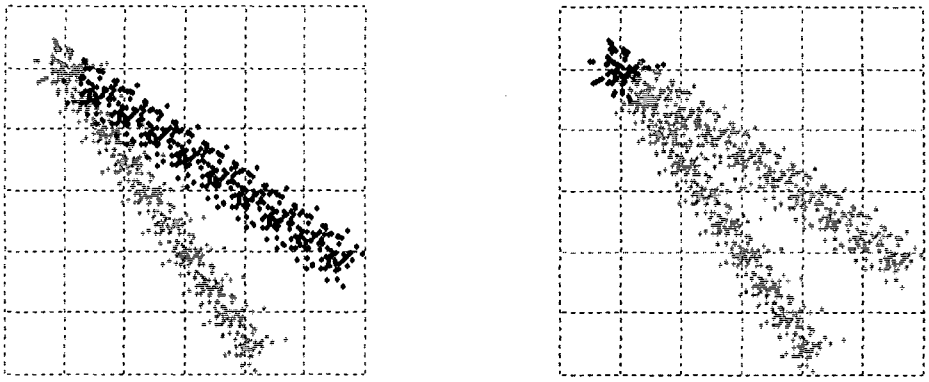
(a) SOM, Ward 방법 (b) K-means, 평균연결법

[그림 10] 방향 차이가 60 도일 때 클러스터링 결과의 가시화



(a) SOM, Ward 방법 (b) K-means, 평균연결법

[그림 11] 방향차이가 45 도일 때 클러스터링 결과의 가시화

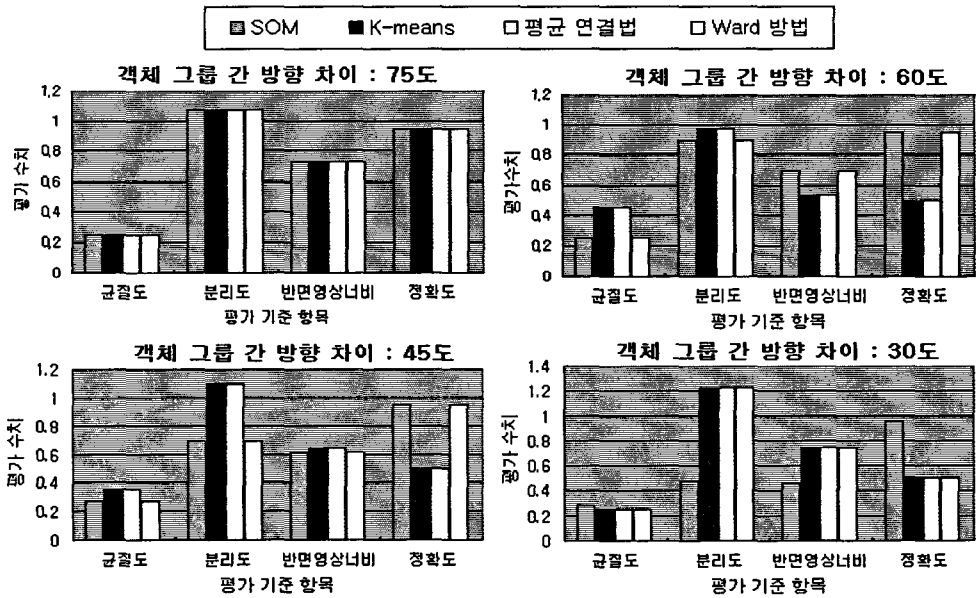


(a) SOM (b) K-means, 평균연결법, Ward 방법

[그림 12] 방향차이가 30 도일 때 클러스터링 결과의 가시화

이 실험을 통해 방향 차이에 따른 각 알고리즘의 방향 속성의 임계값을 구해 본 결과 SOM의 임계값은 30도, K-means와 평균연결법의 임계값은 75도, Ward 방법의 임계값은 45도로 측정되었으며, 즉 SOM의 경우 방향 차이가 작은 경우에도 다른 클러스터링 알고리즘보다 SOM의 클러스터링 성능이 우수하다는 것을 알 수 있었다.

[그림 13]은 데이터세트 D1~D5에 대한 각 클러스터링 알고리즘의 성능 평가 수치를 나타낸 그래프이다. 그림에서 볼 수 있듯이 균질도, 분리도, 반면영상너비와 같이 평가 기준은 입력 데이터의 특성을 고려하지 않고 결과로 얻어진 클러

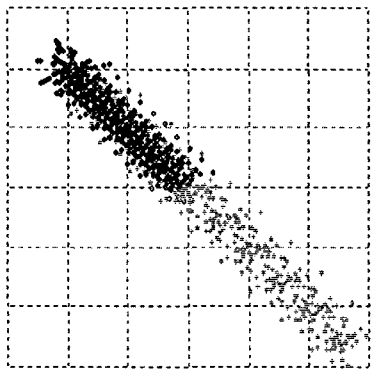


[그림 13] 데이터세트 1-5에 대한 각 클러스터링 알고리즘의 성능 평가 결과

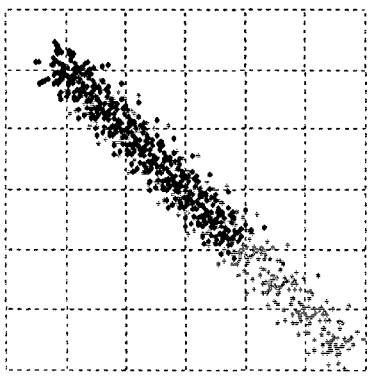
스터에 대한 통계치 만을 반영하는 외적 기준 수치는 알고리즘들 사이에 거의 차이점이 없으나, 실제 입력 데이터의 특성에 따라 클러스터링의 정확도를 평가한 결과 SOM이 다른 알고리즘보다 우수하였다. 이동 객체 그룹들 사이의 출발점이 서로 다른 데이터인 데이터세트 D6에 대한 결과도 이와 유사하게 측정되었다.

■ **데이터세트 D7~D16**

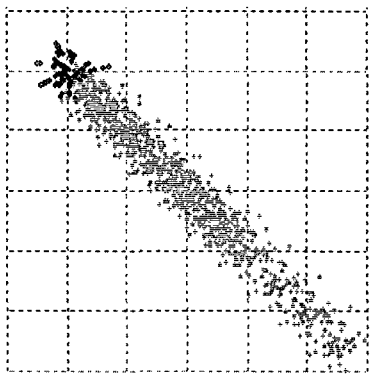
데이터세트 D7~D16은 동일한 방향으로 움직이되 객체 그룹간의 속도가 서로 다른 데이터로 서로 다른 속도로 움직이는 두 개의 객체 그룹에 대한 데이터이다. [그림 14]에서 볼 수 있듯이 객체 그룹간의 속도 차이가 0.6/s일 때 모든 알고리즘의 클러스터링 결과가 동일하게 정확함을 알 수 있다. [그림 15, 16]은 각각 객체 그룹간의 속도 차이가 0.4/s, 0.3/s일 때 각 알고리즘들의 클러스터링 결과를 보여준다. 속도 차이가 0.4/s인 경우 SOM과 Ward 방법은 비교적 결과가 우수하지만 K-means과 평균연결법은 두 객체 그룹을 정확하게 클러스터링 하지 못하며, 차이가 0.3/s인 경우 SOM만이 두 개의 그룹을 정확하게 클러스터링



[그림 14] 속도차이가 0.6/s일 때 클러스터링 결과의 가시화

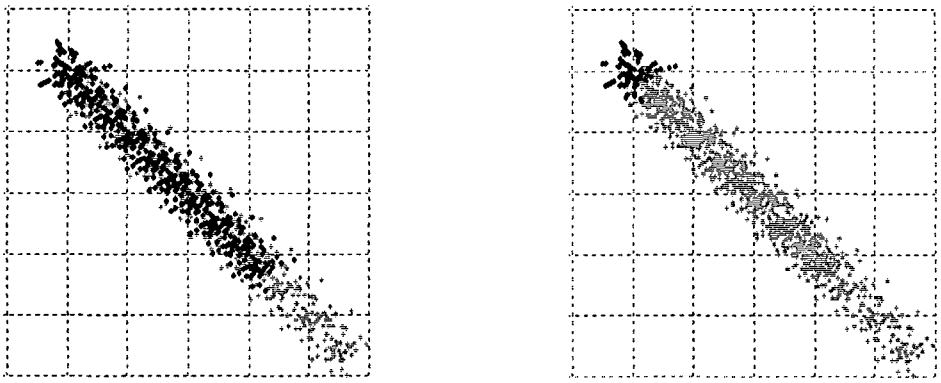


(a) SOM, Ward 방법



(b) K-means, 평균연결법

[그림 15] 속도차이가 0.4/s일 때 클러스터링 결과의 가시화



(a) SOM (b) K-means, 평균연결법, Ward 방법

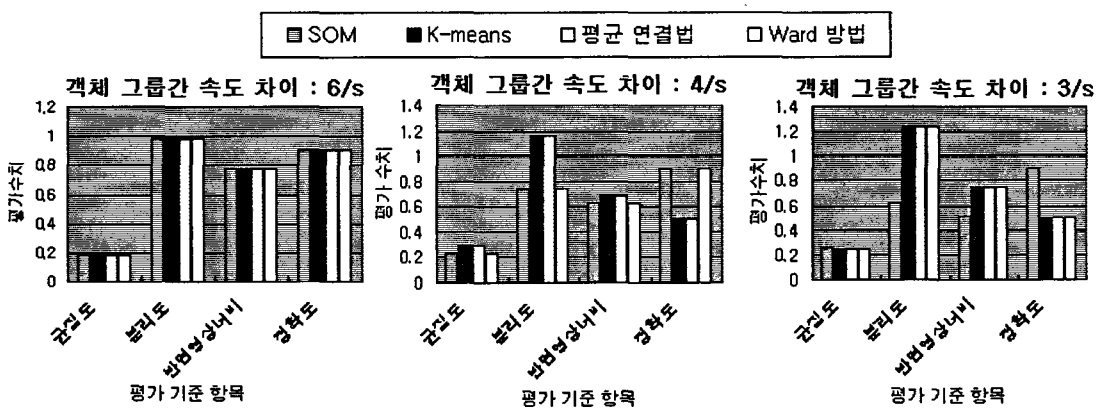
[그림 16] 속도차이가 0.3/s일 때 클러스터링 결과의 가시화

함을 알 수 있다.

실험을 통해 속도 차이에 따른 각 알고리즘의 방향 차이의 임계값을 구해 본 결과 SOM과 Ward 방법의 임계값은 0.3/s, K-means와 평균연결법의 임계값은 0.4/s로 측정되었다. 즉, 속도 차이가 작은 경우에도 다른 알고리즘보다 SOM의 클러스터링 성능이 우수함을 확인할 수 있었다.

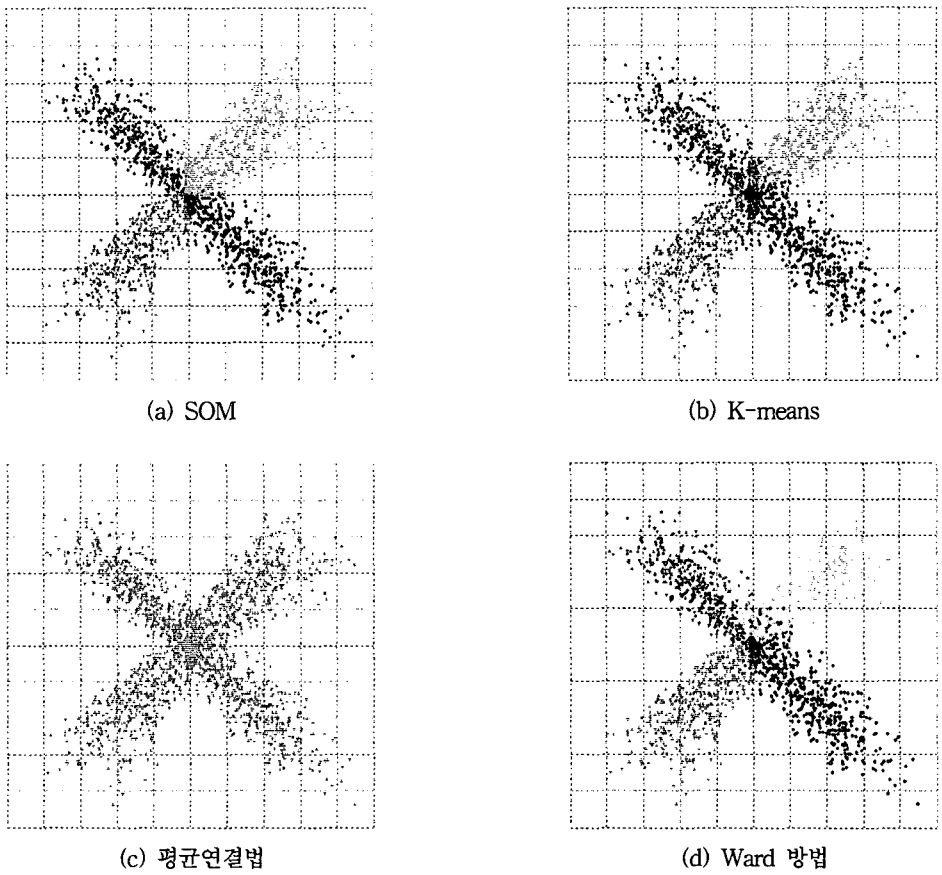
[그림 17]은 데이터세트 11, 13, 14에 대한 각 클러스터링 알고리즘의 성능 평가 수치를 나타낸 그래프이다. 데이터의 특성을 고려한 내적 기준인 정확도 측면에서 SOM이 다른 알고리즘보다 작은 속도 차이까지 구분하여 객체 그룹을 클러스터링 함을 알 수 있다.

[그림 18]은 각 객체들이 객체 그룹별로 서로 다른 방향을 향해 다양한 속도로 불규칙하게 움직이는 4개의 객체 그룹에 대한 데이터인 데이터세트 D17에 대해 클러스터링을 수행 결과이다. <표 7>과 같이 실제 네 가지 성능 평가 기준치를 측정한 결과 SOM, K-means, Ward 방법의



[그림 17] 데이터세트11, 13, 14에 대한 각 클러스터링 알고리즘의 성능 평가 결과

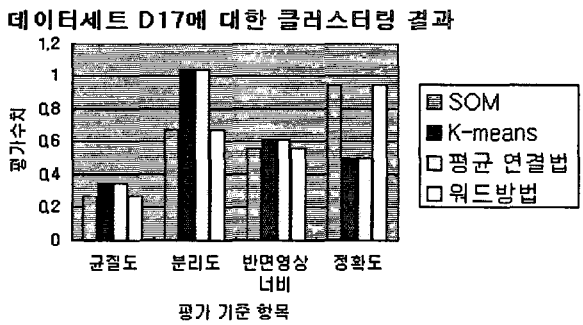
성능에 큰 차이가 없으나 [그림 18]에서 나타나는 것처럼 각 객체들이 움직임을 시작하는 좌표의 중심부분에서는 SOM의 결과가 다른 방법들보다 정확함을 알 수 있다.



(a) SOM (b) K-means

(c) 평균연결법 (d) Ward 방법

[그림 18] 데이터세트 17에 대한 클러스터링 결과의 가시화



[그림 19] 데이터세트 D17에 대한 클러스터링 결과의 성능 평가 수치

이렇게 다양한 데이터세트에 대해 각 알고리즘의 클러스터링 성능을 평가해본 결과 객체들 간의 시공간 속성의 차이가 큰 경우에는 모든 알고리즘의 성능이 우수하지만 객체 그룹들간의 시공간 속성의 차이가 작고 미묘한 경우에는 SOM의 성능이 가장 정확하게 평가되었다. 이는 SOM 입력 데이터의 시공간 속성을 가장 잘 반영하여 입력 데이터의 위상을 결과 클러스터 위상으로 정확히 투영하여 클러스터링을 수행하기 때문이며, 이러한 SOM의 특징은 데이터의 분류화(Classification) 작업에도 효과적이라고 할 수 있다. 하지만 방향과 속도의 차이가 아주 적은 경우 즉, 방향 차이가 30도 보다 작거나, 속도 차이가 0.3/s보다 작을 경우 SOM의 정확도도 떨어짐을 알 수 있었다.

각 알고리즘들 간의 성능을 비교해 볼 때 계층적 알고리즘에 비해 비계층적 알고리즘이 더 좋은 성능을 보였으며, 각 흐름 벡터의 요소간의 유사성을 단순한 유클리드 거리로 판단하는 K-means 기법에 비해 이를 기반으로 학습을 수행하는 SOM의 경우가 더욱 우수한 성능을 보여주었는데 이는 K-means가 SOM에 비해 지역적 최적 해결책을 결과로 도출하기 때문임을 유추해 볼 수 있다.

5. 결론 및 향후과제

본 논문에서는 시공간 데이터 마이닝에 대한 기존 연구들에서 사용되어 온 알고리즘들 중 패턴 인식과 클러스터링 능력이 뛰어나다고 알려진 SOM에 대해 분석하고 SOM 기반 시공간 데이터 클러스터링 모듈을 개발하였다. 구현된 SOM 기반 모듈과 K-means, 두 가지 종류의 응집계층 알고리즘의 클러스터링 결과를 크게 3가지 종류의 17가지 데이터세트를 기반으로 비교 분석하였다. 네 가지 알고리즘들의 클러스터링 성능 평가를 위해 균질도, 분리도, 반면영상 너비, 정확도 네 가지 평가 항목을 기준으로 다양한 방향과 속도 속성을 가지는 데이터에 대한 실험 한 결과 균질도, 분리도, 반면영상 너비와 같이 입력데이터의 특성을 고려하지 않고 결과 클러스터에 대한 통계 수치만을 반영하는 외적 기준에서는 K-means와 응집 계층 알고리즘이 SOM보다 성능이 우수한 부분도 있었다. 그러나 실제로 가시화를 통해 입력 데이터의 특성에 따른 클러스터링 결과를 확인한 결과 SOM이 다른 알고리즘들보다 더 정확하게 마이닝을 수행 하였음을 알 수 있었다. 그리고 내적 성능 평가 기준인 정확성과 임계값 측면에서도 SOM이 더 우수한 결과를 보임을 알 수 있었다.

본 논문의 연구내용을 기반으로 본 연구에서 사용한 네 가지 알고리즘 이외의 다양한 마이닝 알고리즘의 비교 연구나, 시공간 데이터의 속성을 고려한 복합적인 성능 평가 기준에 대한 연구 그리고 임계값보다 작은 속성치를 가지는 이동 객체 그룹을 클러스터링 할 수 있는 개선된 알고리즘에 대한 연구가 필요하다고 보여진다.

참고문헌

[1] J.F. Roddick, and M. Spiliopoulou, "A bibliography of Temporal, Spatial and Spatio-temporal Data Mining Research", Explorations 1(1), pp.34-38, 1999

[2] J.F. Roddick, and B. G. Lees, "Paradigms for Spatial and Spatio-Temporal Data Mining, Geographic Data Mining and Knowedge

Discovery", Miller, H & J. Han (eds). Taylor & Francis, pp.33-50, London, 2001.

[3] N. Johnson and D Hogg, "Learning the Distribution of Object Trajectories for Event Recognition", Image and Vision Computing, 14(8), pp.609-615, 1996

[4] J. Owens and A. Hunter, "Application of the Self-Organizing Map to Trajectory Classification", Third IEEE International Workshop on Visual Surveillance (VS'2000), pp.77-84, 2000

[5] Y. Theodoridis, J. R.O. Silva, and Mario A. Nascimento, "On the Generation of Spatiotemporal Datasets", In Proc. of the 6th Int'l Symposium on Large Spatial Database (SSD), pp.147-164, 1999

[6] http://www.insightful.com/

[7] 성유진, 박종수, "고차원 입력 데이터에 대한 클러스터링 알고리즘의 성능 분석", 제2회 데이터마이닝워크숍, 1999

[8] Chen G, Jaradat SA, Banerjee N, Tanaka TS, Ko MSH, Zhang MQ. "Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data", Statistica Sinica, pp.241-262, 2000

[9] Yeung, Haynor, Ruzzo: Validating Clustering for Gene Expression Data. Technical Report UW-CSE-00-01-01, 2000

[10] M. Erwig, R. H. Gutting, M. Schneider, M. Vazirgiannis, "Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases", Technical Report, CHOROCHRONOSTR -97-08, 1997

[11] P. Stolortz, H. Nakamura, E. Mesrobian, and et al., "Fast Spatio-Temporal Data Mining of Large Geophysical Datasets", In Proc. of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp.300-305, 1995

[12] U. Fayyad, D. Haussler, and P. Stolorz, "KDD for science data analysis: Issues and examples", In Proc. of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp.50-56. CA: AAAI Press, 1996.

[13] C. Shahabi, X. Tian,and W. Zhao. "TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries", In The 12th International Conference on Scientific and Statistical Database Management, pp.55-68 , SSDBM, 2000

[14] S. Rogers, P. Langley, and C. Wilson, "Mining GPS data to augment road models", In Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, ACM Press, pp.104-113, San Diego CA, 1999

[15] S. Gafney and P. Smyth, "Trajectory clustering with mixtures of regression models", In Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, ACM Press, pp.104-113, San Diego CA, 1999

[16] C. Stauffer and W. Eric L. Grimson, "Learning patterns of activity using real-time tracking", IEEE Trans. PAMI, vol. 22, pp.747-757, 2000

[17] N. Sumpter and A. Bulpitt, "Learning spatio-temporal patterns for predicting object behaviour", Technical report, University of Leeds, School of Computer Studies, The University of Leeds, UK. 1998

[18] J. Eisenstein, S. Ghandeharizadeh, L. Huang, C. Shahabi, G. Shanbhag and R. Zimmermann, "Analysis of Clustering Techniques to Detect Hand Signs", Int'l Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, 2001

[19] P. Remagnino, T. Tan, and K. Baker. "Agent orientated annotation in model based visual surveillance", In ICCV, pp.857-862, 1998

[20] S. Handley, P. Langley and F. Rauscher,

"Learning to predict the duration of an automobile trip", In Proc. of the 4th International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp.219-223, New York, 1998

[21] P.C. Juan and L.C. Ignacio, Discovering Similar Patterns in Time Series, In Proc. In Proc. of the 4th International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp.497-505, New York, 1998

[22] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000

[23] 김대주, 신경망 이론과 응용(Ⅰ), 하이테크정, pp.169-189, 2001

[24] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Information Systems, pp.107-145, 2001

Abstract

Performance Comparison of Clustering Techniques for Spatio-Temporal Data

Nayoung Kang* · Juyoung Kang** · Hwan-Seung Yong***

With the growth in the size of datasets, data mining has recently become an important research topic. Especially, interests about spatio-temporal data mining has been increased which is a method for analyzing massive spatio-temporal data collected from a wide variety of applications like GPS data, trajectory data of surveillance system and earth geographic data. In the former approaches, conventional clustering algorithms are applied as spatio-temporal data mining techniques without any modification.

In this paper, we focused to SOM that is the most common clustering algorithm applied to clustering analysis in data mining area, and develop the spatio-temporal data mining module based on it. In addition, we analyzed the clustering results of developed SOM module and compare them with those of K-means and Agglomerative Hierarchical algorithm in the aspects of homogeneity, separation, separation, silhouette width and accuracy. We also developed specialized visualization module for more accurate interpretation of mining result.

Key words : Data Mining, Spatio-Temporal Data Mining, Clustering, SOM, Performance Evaluation

* Engineering Information Group, Memory Division, Semiconductor Business, SAMSUNG ELECTRONICS CO., LTD

** Power Information Technology Group, Power System Research Laboratory, Korea Electric Power Research Institute

*** Dept. of Computer Science and Engineering, Ewha Womans Univ.