이화여자대학교 과학기술대학원
2002 학년도
석사학위 청구 논문

시공간 데이터를 위한 클러스터링 기법의 성능 비교

컴 퓨 터 학 과

강 나 영

2 0 0 3

시공간 데이터를 위한
클러스터링 기법의 성능 비교

이 論文을 碩士學位 論文으로 提出 함

2003 年 **7** 月

梨花女子大學校 科學技術大學院

컴퓨터학과　姜 那 榮

강 나 영의 碩士學位 論文 을 認准함

指導敎授　　용　환　승

審査委員　　조　동　섭　　_____

이　민　수　　_____

용　환　승　　_____

梨花女子大學校　科學技術大學院

목 차

논문개요 ---------- v

Ⅰ. **서론** ---------- 1

1.1 연구 배경 및 목적 ---------- 1

1.2 연구 내용 ---------- 2

Ⅱ. **관련 기술 및 연구 동향** ---------- 4

2.1 시공간 데이터 마이닝 ---------- 4

2.1.1 시공간 데이터의 정의 및 특징 ---------- 4

2.1.2 시공간 데이터 마이닝의 응용 도메인 ---------- 5

2.2 시공간 데이터 마이닝 기법 ---------- 6

2.2.1 클러스터링 기반 접근 방법 ---------- 6

2.2.2 패턴 탐사 기반 접근 방법 ---------- 8

2.3 클러스터링 알고리즘의 비교 분석 연구 ---------- 9

Ⅲ. **시공간 데이터 마이닝 클러스터링 기법** ---------- 10

3.1 K-means ---------- 10

3.2 응집 계층(Agglomerative Hierarchical) 알고리즘 ---------- 11

3.3 SOM(Self-Organizing Map) ---------- 11

3.4 SOM 기반 마이닝 모듈 설계 및 구현 ---------- 13

3.4.1 구현환경 ---------- 13

3.4.2 전처리 모듈 ---------- 14
3.4.3 SOM 모듈 ---------- 16
3.4.4 가시화 모듈 ---------- 17

Ⅳ. **성능평가** ---------- 20
4.1 실험 시공간 데이터 ---------- 20
4.1.1 기존 시공간 데이터들 ---------- 20
4.1.2 GSTD(Generate Spatio-Temporal Data) ---------- 21
4.1.3 실험 데이터의 특성 ---------- 23
4.2 성능 평가 기준 ---------- 26
4.2.1 균질도(Homogeneity) ---------- 26
4.2.2 분리도(Separation) ---------- 27
4.2.3 반면영상 너비(Silhouette Width) ---------- 27
4.2.4 정확도(Accuracy) ---------- 28
4.3 성능 평가 결과 ---------- 28

Ⅴ. **결론 및 향후 과제** ---------- 43
참고문헌 ---------- 45

그 림 목 차

[그림 3.1] Self-Organizing Map ---------- 12
[그림 3.2] 가시화 모듈의 사용자 인터페이스 메인 화면 ---------- 17
[그림 3.3] 입력데이터와 SOM 출력노드의 최종 가중치 가시화 ---------- 19
[그림 4.1] 객체 이동의 연속성 ---------- 22
[그림 4.2] GSTD 웹사이트 화면 ---------- 22
[그림 4.3] 출발위치가 동일한 시공간 데이터 ---------- 24
[그림 4.4] 궤적이 서로 교차되는 시공간 데이터 ---------- 24
[그림 4.5] 서로 다른 속도로 움직이는 시공간 데이터 ---------- 25
[그림 4.6] 움직임이 불규칙한 시공간 데이터 ---------- 24
[그림 4.7] 방향차이가 75 도일 때 클러스터링 결과의 가시화 ---------- 29
[그림 4.8] 방향차이가 60 도일 때 클러스터링 결과의 가시화 ---------- 29
[그림 4.9] 방향차이가 45 도일 때 클러스터링 결과의 가시화 ---------- 30
[그림 4.10] 방향차이가 30 도일 때 클러스터링 결과의 가시화 ---------- 30
[그림 4.11] 사례 A-2 의 클러스터링 결과의 가시화 ---------- 31
[그림 4.12] 속도차이가 0.6/s 일 때 클러스터링 결과의 가시화 ---------- 34
[그림 4.13] 속도차이가 0.4/s 일 때 클러스터링 결과의 가시화 ---------- 35
[그림 4.14] 속도차이가 0.3/s 일 때 클러스터링 결과의 가시화 ---------- 35
[그림 4.15] 사례 C 의 클러스터링 결과의 가시화 ---------- 38
[그림 4.16] 사례 C 의 성능 평가 기준치의 그래프(K=4, 9, 16, 25, 36) ------ 40
[그림 4.17] 사례 C 의 SOM 네트워크(K=16) ---------- 41
[그림 4.18] 위상적으로 근접한 4 개의 클러스터 가시화 ---------- 42

표 목 차

[표 3.1] 구현 환경 ---------- 14
[표 3.2] 원시 시공간 데이터 테이블 정의 ---------- 15
[표 3.3] 시공간 데이터 테이블 정의 ---------- 16
[표 4.1] 사례 A-1 의 클러스터링 결과의 성능 평가 수치 ---------- 32
[표 4.2] 사례 A-2 의 클러스터링 결과의 성능 평가 수치 ---------- 34
[표 4.3] 사례 B 의 클러스터링 결과의 성능평가 수치 ---------- 36
[표 4.4] 사례 C 의 클러스터링 결과의 성능 평가 수치 ---------- 39

論 文 槪 要

최근 데이터 양이 급증하면서 데이터 마이닝에 대한 연구가 활발하게 진행되고 있다. 특히 GPS 데이터, 감시 카메라의 궤적 데이터, 기상 데이터들과 같은 다양한 응용 시스템으로부터 수집된 시공간 데이터를 분석하고자 하는 시공간 데이터 마이닝 연구에 대한 관심이 더욱더 높아지고 있다.

기존 연구들에서는 SOM, K-means, 응집 계층 알고리즘과 같은 일반적인 클러스터링 기법들을 적용하여 시공간 데이터 마이닝을 수행하고 있다. 하지만 이러한 알고리즘들이 실제적으로 시공간 데이터에 이러한 기법들을 적용하는데 있어서 어느 정도의 성능을 보장할 수 있는지 혹은 데이터의 시공간속성에 따라 적절한 마이닝 알고리즘을 선택하기 위한 기준이 무엇인지 등에 대한 연구는 미흡한 실정이다.

본 논문에서는 기존의 시공간 데이터 마이닝 연구에서 주로 사용되어 온 알고리즘인 SOM 을 분석하여 SOM 기반 마이닝 모듈을 개발한다. 그리고 K-means 와 응집 계층 알고리즘과의 성능 비교를 통해 SOM 이 시공간 마이닝에 있어서 어느 정도의 성능을 보장하는지를 균질도, 분리도, 반면영상 너비, 정확도의 네 가지 기준에서 분석한다. 또한 시공간 데이터의 경우 입력 데이터의 속성에 따라 이러한 평가 기준 수치가 클러스터링 결과의 정확성 및 성능을 제대로 나타내지 못하는 경우가 발생할 수 있다는 점을 고려하여 시공간 데이터의 클러스터링 결과를 위한 특화된 가시화 모듈을 개발하고 이를 통해 결과 비교 및 분석을 수행한다.

I. 서 론

1.1 연구 배경 및 목적

최근 위성 시스템, 의학 진단 시스템, 감시 시스템, 자연 과학 시스템(natural science observation systems)이나 교통 시스템 등과 같은 다양한 과학 기술 응용 시스템의 에서 수집된 방대한 양의 시공간 데이터를 좀 더 빠르고 심도 있게 분석하고자 하는 시공간 데이터 마이닝 (Spatio-Temporal Data Mining) 연구에 대한 관심이 더욱더 높아지고 있다. 기존의 데이터 마이닝의 분석 대상이 되었던 데이터는 일반적으로 문자나 숫자 데이터(alphanumeric data)들을 기반으로 하고 있는 반면, 시공간 데이터는 시간과 공간의 속성을 동시에 지니고 있기 때문에 시공간 데이터 마이닝에서는 데이터 분석 시 이러한 속성들을 적절하게 고려해 주어야만 한다[1]. 또한 기존의 데이터 마이닝과는 달리 시공간 데이터 마이닝에서는 지식 탐사의 절차와는 상관 없이 입력의 형태나 속성, 도출된 결과의 의미 적절한 해석이 더욱 중요한 점으로 다루어져야 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[J1]. 현재까지는 주로 K-means, SOM(Self-Organizing Map), 응집 계층(agglomerative hierarchical) 알고리즘과 같은 기본적인 클러스터링 마이닝 알고리즘들을 기반으로 한 시공간 데이터 마이닝 연구가 일반적인데[3][4], 기존의 문자 및 숫자 대상의 마이닝 작업에 비교해 볼 때 실제적으로 기존의 비시공간 데이터를 위한 알고리즘들이 어느

정도의 성능을 보장하는지, 혹은 각 알고리즘 별로 데이터의 특정한 시공간 속성에 따른 알고리즘의 수행 능력은 어떻게 변화하는지, 시공간 속성에 따라 적절한 마이닝 알고리즘을 선택하기 위한 기준은 무엇인지 등에 대한 연구는 미흡한 실정이다.

그러므로 시공간 데이터 마이닝의 응용 별로 더욱 적합한 마이닝 결과를 도출하기 위해서는 시공간 데이터 마이닝만을 위한 적절한 데이터 마이닝 알고리즘, 모델링 기법, 인덱스 및 저장 기법 등의 연구가 필요할 뿐 아니라, 기존의 마이닝 기법들을 확장하여 적용함에 있어서 시공간 데이터의 특성 및 응용의 속성에 따라 적절한 해결 기법을 선택적으로 적용할 수 있도록 객관적인 기준을 제시할 필요가 있다.

1.2 연구 내용

본 논문에서는 시공간 데이터 마이닝에 대한 선행 연구들에서 일반적으로 사용되어 온 알고리즘들 중 패턴 인식과 클러스터링 능력이 뛰어나다고 알려진 SOM 에 대해 분석하고, 기존의 SOM 모듈을 수정하여 시공간 데이터를 기반으로 클러스터링을 수행하는 마이닝 모듈을 개발한다. 또한 실제로 시공간 데이터 마이닝에 있어서의 SOM 의 성능에 대한 객관적인 기준을 제시하기 위해서 K-means 클러스터링 알고리즘, 응집 계층 클러스터링 알고리즘과 개발된 SOM 모듈의 클러스터링 결과에 대해 성능 비교 및 분석 작업을 수행한다. 성능 비교를 위해 본 논문에서는 시공간 데이터 베이스의 벤치마킹을 위한 통합 시스템인 GSTD(Generate Spatio-Temporal Data) 툴[5] 에서 제공하는 시공간 데이터 생성 모듈을 이용하여 데이터를 생성하고, 본 논문의 SOM 모듈과 Insightful 사의 통계 분석 및 마이닝 프로그램인 S-PLUS[6]의 K-means 모듈에 대한 클러스터링 결과를 균질도(homogeneity), 분리도(separation), 반면영상 너비

(silhouette width), 정확도(accuracy)와 같은 네 가지 기준치를 기반으로 비교한다. 일반적인 문자 및 숫자 데이터의 클러스터링 결과의 분석을 위해서는 위와 같은 수치 기준들 만으로 평가하는 것이 일반적이지만, 본 연구에서 대상으로 하고 있는 시공간 데이터의 경우 입력 데이터의 속성에 따라 평가 기준 수치가 클러스터링의 결과의 정확성 및 성능을 제대로 나타내지 못하는 경우가 발생할 수 있다. 이러한 점을 고려하여 본 연구에서는 시공간 데이터 및 클러스터링 결과를 위한 가시화 모듈을 개발하고 이를 통해 가시화 작업을 통한 결과 비교 및 분석을 수행한다.

본 논문의 구성은 다음과 같다. 2 장에서는 시공간 데이터의 정의 및 특성과 시공간 데이터 응용 분야에 대해 소개하고 기존의 시공간 데이터 마이닝 기법들에 대한 선행 연구 및 클러스터링 알고리즘의 성능 비교 분석 연구에 대해 살펴본다. 3 장에서는 시공간 데이터 마이닝의 기법들 중 본 논문에서 시공간 데이터 마이닝 작업을 위해 사용한 SOM 과 K-means, 응집 계층 알고리즘을 중심으로 클러스터링 마이닝 기법들에 대해 살펴본 후, 개발한 SOM 기반 마이닝 모듈의 설계 및 구현에 대해 기술한다. 4 장에서는 시공간 데이터를 생성하기 위해 본 논문에서 사용한 GSTD 시공간 데이터베이스 벤치마킹 툴에 대해 소개하고, 본 논문의 성능 평가 작업을 위해 GSTD 데이터 생성기를 통해 생성한 테스트 데이터 집합의 특성에 대해 살펴보고, 성능 평가의 기준 항목들에 대해 살펴 본 후 이러한 기준들에 기반 한 두 가지 클러스터링 알고리즘의 성능 평가 결과에 대해 기술한다. 5 장에서는 결론 및 향후 연구 과제를 제시한다.

II. 관련 기술 및 연구 동향

데이터 마이닝과 지식 탐사에 대한 흥미와 중요성이 더해갈수록 데이터 마이닝을 지식 관리 방법의 하나로써 다양한 응용 분야에 적용시키고자 하는 연구가 더욱 활발해지고 있다. GPS 데이터, 감시 카메라의 궤적 데이터, 기상 데이터 등과 같은 다양한 응용 시스템으로부터 수집된 시공간 데이터를 분석하고자 하는 시공간 데이터 마이닝 또한 그 중요성을 더해가고 있다. 본 장에서는 시공간 데이터 마이닝과 관련된 다양한 응용 분야와 그로부터 생성된 시공간 데이터의 특성에 대해 살펴보고, 이러한 데이터들을 분석하기 위해 제안된 기존의 연구들에 대해 살펴보기로 한다. 또한 데이터 마이닝 알고리즘의 성능 평가 및 분석에 대한 연구 동향에 대해 살펴 본다.

2.1 시공간 데이터 마이닝

본 절에서는 시공간 데이터의 정의와 특성에 대해 살펴보고 이러한 데이터와 관련된 시공간 데이터 마이닝의 응용 분야에 대해 살펴 본다.

2.1.2 시공간 데이터의 정의 및 특징

시공간 데이터(Spatio-Temporal data)란 시간에 따라 변화하는 기하학 객체들에 대한 데이터이다. 데이터는 기하학적인 정보로 구성되며 이산적이거나 연속적인 위치 정보를 포함할 수 있다. 만약 객체 공간에서의 위치 자체만을 고려한다면 데이터는 이동 점(point)으로 나타내어질 수 있다. 반면 이동 지역(region)으로 나타낼 경우 데이

터는 증가(growing)하거나 줄어드는(shrinking) 객체의 크기에 관한 정보도 포함할 수 있다. 이러한 데이터들은 데이터를 공간과 시간의 흐름 상에 위치시킬 수 있는 거리 속성(distance attribute) 및 시간 속성(time attribute)를 갖는다는 점에서 비시공간 데이터와 다르다[7].

2.1.2 시공간 데이터 마이닝의 응용 도메인

현재까지의 시공간 데이터 마이닝의 대상이 되어왔던 응용 분야 중 가장 두드러지는 분야는 자연과학 분야이다. 자연 과학 분야에서의 시공간 데이터 마이닝은 다양한 과학 실험 및 관찰 시스템에서 수집된 지구 기상 데이터, 미생물, 식물 및 동물의 실험 데이터 등과 같은 방대한 양의 자연과학 데이터를 기반으로 하고 있다[8]. 오랜 기간에 걸쳐 수집되고 축적되어 온 자연과학 데이터들의 방대한 양에 비해 이를 기반으로 복잡한 과학적 연구 분석을 수행할 만 한 분석 기술은 부족하기 때문에 이 분야에서는 데이터 마이닝 기법을 통한 여러 가지 접근 방법들에 대한 연구가 이루어졌다. CONQUEST(CONTENT-based Querying in Space and Time)[9], TSA-Ttree[10], Quakefinder[8]와 같은 시스템들은 데이터 마이닝 기법을 기반으로 지구 기상 및 지질학 데이터로부터 사이클론, 지진, 기상 경향, 화산 분출과 같은 활동을 예측해내는 시스템으로 자연과학 분야의 대표적인 시공간 데이터 마이닝 시스템들이다.

자연과학 분야 이외에 움직이는 이동 객체를 추적하는(tracking) 감시 시스템이나 모니터링 시스템, GPS 시스템 혹은 교통 시스템과 같은 시스템에서 생성되는 위치 혹은 궤적(trajectory) 데이터를 다루는 이동 객체 관련 분야에서도 데이터 마이닝 기법을 기반으로 한 연구가 진행되고 있다. GPS 데이터를 마이닝하여 전자 지도 제련

(refinement) 작업을 수행하고자 하는 연구나[11], 객체 추적 시스템(object tracking system)이나 감시 시스템(surveillance system)등에서 수집된 이동 객체 데이터를 이용한 궤적 분류(trajectory classification) 작업이나 궤적 예측(trajectory prediction) 시스템들에 대한 연구가 진행되어 왔다[3,4,12,13,14].

2.2 시공간 데이터 마이닝 기법

문자나 숫자 기반의 기존의 데이터 마이닝과 지식 탐사 기법에 대한 연구가 다양한 접근 방법을 기반으로 풍부하게 이루어져온 반면 시공간 데이터 마이닝에 대한 연구는 이에 비해 매우 부족한 실정이다[1]. 시공간 데이터 마이닝에 대한 연구들은 그 접근 방법에 따라 크게 두 가지로 나누어 볼 수 있다. 첫째는 K-means나 신경망을 이용한 클러스터링 기반의 시공간 데이터 분석 방법이고, 둘째는 기존의 시계열 분석 방법에서 사용되어왔던 방법을 확장한 패턴 탐사(pattern discovery) 방법이다. 현재까지는 첫 번째 방법인 클러스터링 기반의 시공간 마이닝 기법에 대한 연구가 더 활발히 이루어지고 있다. 본 절에서는 시공간 데이터 마이닝에 대한 두 가지 접근 방법 별로 현재까지 연구되어 온 시공간 데이터 마이닝 기법들에 대해 살펴보기로 한다.

2.2.1 클러스터링 기반 접근 방법

클러스터링은 서로 간에 높은 유사도를 가지는 객체들을 같은 클러스터로 그룹화하는 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[15]. 일반적으로 자

주 사용되는 클러스터링 알고리즘들로는 분할 방법의 일종인 K-means, PAM(Partitioning Around Medoids)나 CLARANS(Clustering Large Applications based upon RANdomized Search), BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies)나 CURE(Clustering Using Representatives)와 같은 계층적 방법의 알고리즘, 그리고 밀도 기반 방법의 DBSCAN(A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density) 등이 있다[15]. 또한 이러한 클러스터링 알고리즘 이외에도 벡터 양자화 작업을 수행할 수 있는 SOM(Self-Organizing Map)과 같은 신경망이나 베이지안 네트워크(bayesian network)와 같은 인공 지능의 학습 기법들도 일종의 클러스터링 알고리즘으로 볼 수 있다.

시공간 데이터 마이닝 연구에서는 위와 같은 클러스터링 알고리즘 중 K-means 알고리즘과 SOM, 응집 계층(agglomerative hierarchical) 알고리즘이 일반적으로 많이 사용되어 왔다. 실시간 추적 시스템을 통해 얻어진 이동 객체 데이터들을 분석 하여 활동 공간에서의 객체 움직임의 패턴을 찾아내는 모니터링 시스템이나[13], 가상 현실에서 사용자의 수신호(手信號)를 입력하기 위해 착용하는 햅틱(haptic) 장갑과 같은 장치로부터 수집된 햅틱 데이터에 대해 K-means 알고리즘을 기반으로 한 클러스터링 작업을 수행하여 사용자의 수신호를 바르게 인식하고자 하는 연구가 수행된 바 있다[16]. Neil Johnson과 David Hogg [3]과 Jonathan Owens와 Andrew Hunter [4]는 고차원 벡터 데이터를 2차원 위상 구조로 매핑하는 SOM의 특징을 기반으로 객체 추적 시스템을 통해 수집된 이동 객체 데이터에 대한 시퀀스들을 유속 벡터(flow vector)라는 일종의 벡터 값으로 변환하여 시공간 데이터 클러스터링을 수행하였다. 또한 SOM 신경망을 확장하여 두 개의 신경망을 Leaky Integrator Layer라는 중간 계층으로 두 개의 SOM 신경망을 연

결한 복합 신경망을 기반으로 이동 객체 데이터에 대한 클러스터링 연구가 수행된 바 있다[J14]. 이외에 계층적 클러스터링 기법을 GPS 데이터의 분석에 적용하여 자동차 안전 시스템이나 편의 시스템을 위한 전자 도로 지도를 자동적으로 수정, 개선하고자 하는 연구가 진행되었다[11]. 또한 다양한 통계 기법이나 베이지안 네트워크, K-NN등과 같은 학습 기법들을 기반으로 한 이동 객체 데이터에 대한 궤적 클러스터링이나 목적지까지의 자동차 주행 시간 예측 시스템에 대한 연구가 진행된 바 있다[14,17,18].

2.2.2 패턴 탐사 기반 접근 방법

패턴 탐사 방법은 예전부터 시계열(time series) 데이터 및 시퀀스 데이터의 분석에 자주 이용되어 왔다. 서로 다른 길이의 주기적인 패턴을 찾아내기 위해서 Apriori 규칙을 기반으로 하여 후보 패턴들을 생성한 후 패턴의 등장하는 빈도와 신뢰도를 기반으로 가지치기(pruning) 작업을 수행하는 과정을 통해 빈번하게 발생하는 패턴이나 에피소드(episode)들을 발견해 낸다[19]. Juan P. Caraca-Valente와 Ignacio Lopez-Vhavarrias는 [20]에서 패턴 탐사 데이터 마이닝 방법을 기반으로 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 시스템을 개발하였다. 이외에도 이동 객체의 궤적 데이터를 일종의 확장된 시계열 데이터로 간주하여 패턴 탐사 방법을 적용하여 데이터를 분석하고자 하는 접근 방법들에 대해 연구가 수행된 바 있다[21].

2.3 클러스터링 알고리즘의 비교 분석 연구

본 논문에서는 이동 객체 데이터를 기반으로 시공간 데이터 마이닝 분야에서 가장 널리 사용되는 클러스터링 알고리즘인 K-means, 응집계층 알고리즘과 SOM의 성능에 대한 객관적인 기준을 제시하기 위해 세 알고리즘의 성능 평가 및 비교 작업을 수행한다. 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[7]. 클러스터링 결과의 성능은 확장성, 클러스터 모양의 다양성, 분석 대상 데이터의 융통성, 잡음 데이터의 처리등과 같은 다양한 기준에 있어서 평가되어야 한다[15]. [22]에서는 유전자 데이터 클러스터링을 수행하는 데 있어서 네 가지의 클러스터링 알고리즘의 성능을 비교 분석하였는데, 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 결과를 비교 하였다. 실제 클러스터링 결과의 정확성, 즉 예상 클러스터와 결과 클러스터의 일치도 또한 위의 기준들과 더불어 클러스터링 결과의 평가 기준으로 사용되기도 한다[23]

Ⅲ. 시공간 데이터 클러스터링 기법

본 장에서는 시공간 데이터의 클러스터링에 주로 이용되고 있는 클러스터링 기법 중에서 K-means 와 응집 계층 알고리즘, SOM 에 대해서 설명하고 본 논문에서 구현한 SOM 모듈에 대해 살펴본다.

3.1 K-means

K-means 는 객체를 K 개의 그룹으로 분할(Partitioning)하는 방법으로, 이 때 클러스터에 속하는 객체들의 평균값을 중심점으로 하여 분할한다.

K-means 가 수행되는 과정은 다음과 같다.

1. 군집의 수 K 를 결정한다.
2. 초기 K 개 군집의 중심을 선택한다.
3. 주어진 중심점을 기준으로 각 객체를 가까운 군집에 할당한다. 이 때 중심점과 객체간의 거리는 Euclidian 거리로 계산한다.
4. 새로 할당된 객체를 중심으로 각 군집의 새로운 중심점을 계산한다.
5. 만약 기존 중심점과 새로운 중심점간의 차이가 없으면 중지하고, 그렇지 않으면 2 번으로 돌아가 다시 수행한다.

본 논문에서는 S-PLUS 툴을 이용하여 시공간 데이터를 K-means 에 적용하였다.

3.2 응집 계층(Agglomerative Hierarchical) 알고리즘

응집 계층 알고리즘은 계층 알고리즘 중 bottom-up 알고리즘으로 우선 모든 n 개의 객체가 n 개의 서로 다른 그룹이라 가정한 후에 그룹간의 유사성(simlilarity)을 보고 가장 유사한 두 개의 그룹을 합병해 그룹 수를 줄여가는 과정을 전체 그룹 수가 k 개가 될 때까지 반복함으로써 k 개의 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) 등이 있다. 단일 연결법은 클러스터 사이의 거리를 각 클러스터내에 포함된 객체들 사이의 거리중 최소의 거리로 계산한다. 완전 연결법은 각 클러스터내에 포함된 객체들 사이의 거리중 최대 거리를 두 클러스터의 거리로 정의한다. 평균 연결법은 단일연결법과 완전연결법의 특징을 절충한 알고리즘으로, 각 클러스터 사이의 거리는 각 클러스터내에 포함된 객체들의 중심 사이의 거리로 계산한다. 시공간 데이터 마이닝에서는 연결법 중 평균 연결법이 주로 사용된다. 워드법은 전체 군집내 제곱합을 이용하여 군집의 수를 줄여가는 방법으로, 군집 수를 줄이기 위해 현재 군집 내의 객체들에 대해 전체 제곱합을 구한 후 이 값이 가장 작은 군집끼리 병합하는 방법이다.

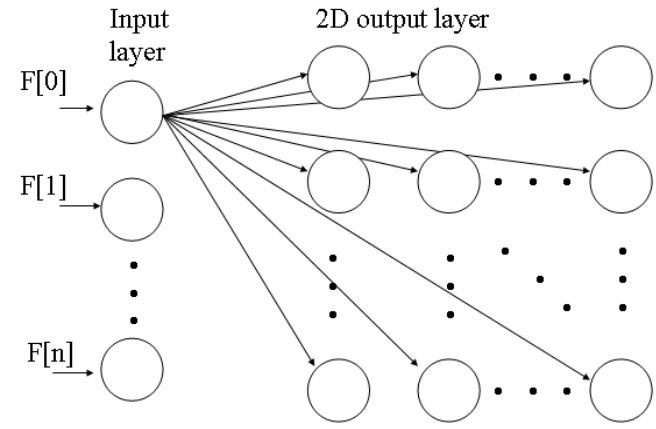
본 논문에서는 S-PLUS 툴을 이용하여 시공간 데이터를 응집 계층 알고리즘 중 평균 연결법과 워드법에 적용하였다.

3.3 SOM (Self-Organizing Map)

SOM 은 Kohonen 이 제안한 신경망 기반의 자기조직화 알고리즘으로 해부학적인

이론에 근거하여 인간의 두뇌 구조를 모델링한 방법이다. 즉, 인접한 출력노드들은 비슷한 기능을 수행할 것이라고 예측하여, 기존의 경쟁학습(Competitive Learning)을 개선하여 입력노드와 가장 가까운 출력노드들뿐만 아니라 그 출력노드의 이웃노드들도 함께 학습시키는 알고리즘이다[24].

[그림 3.1]와 같이 SOM 은 기본적으로 2 개의 층으로 이루어져 있다. 첫 번째 층은 입력층(Input Layer)이고, 두 번째 층은 2 차원의 출력층(Output Layer)이며, 모든 연결은 입력층에서 출력층의 방향으로 되어 있다.



[그림 3.1] Self-Organizing Map

SOM 은 패턴 인식과 클러스터링 능력이 뛰어나기 때문에 현재 진행되고 있는 시공간 데이터 마이닝 연구에 많이 응용되고 있다.

SOM 이 수행되는 과정은 다음과 같다.

1. K 개의 출력노드를 위상공간 내에 배치한다.
2. 학습률 $\alpha(t)$ 을 초기화한다. 학습률은 0 과 1 사이의 값을 가지며, 시간이 지남에 따라 감소한다.

3. 새로운 $x(t)$ 입력벡터를 입력노드에 제시한다
4. 입력벡터와 모든 출력노드들과의 거리를 계산하여 최소거리를 가지는 승자노드 $m_c(t)$ 를 찾는다. 이 때 거리는 Euclidian 거리로 계산한다.

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

5. 승자노드와 이웃한 출력노드들의 가중치를 갱신한다.
6. 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.

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

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

7. 2 번으로 가서 반복한다.

본 논문에서는 시공간 데이터를 SOM 에 적용하기 위해 SOM 기반 마이닝 모듈을 개발하였다.

3.4 SOM 기반 마이닝 모듈 설계 및 구현

본 절에서는 SOM 모듈 및 전처리 모듈, 가시화 모듈의 구현 내용에 대해 설명한다.

3.4.1 구현 환경

본 논문에서 시공간 데이터에 대한 클러스터링 알고리즘의 성능 비교를 위한 클러스터링 알고리즘 중 SOM 모듈을 구현하였다. 원시 시공간 데이터를 전처리하기 위한

데이터 전처리 모듈과 클러스터링 결과를 가시화하기 위한 가시화 모듈을 함께 구현하였다. 구현 환경은 [표 3.1]과 같다.

운영체제	Windows 2000 Advanced Server
DBMS	Oracle 9i
개발도구 및 언어	Oracle PRO*C Microsoft Visual C++ 6.0 Microsoft Visual Basic 6.0 ADO 2.6 Library

[표 3.1] SOM 기반 마이닝 모듈 구현 환경

데이터 전처리 모듈과 SOM 모듈은 Oracle Pro*C 를 이용하여 오라클 9i 에 저장되어 있는 시공간 데이터를 처리하고, 재컴파일을 위해 비쥬얼 C++ 6.0 을 사용하여 구현하였다. 그리고 가시화 모듈은 비쥬얼 베이직 6.0 과 ADO 2.6 라이브러리를 사용하여 구현하였다.

3.4.2 전처리 모듈

SOM 에 적용하기 위해서는 원시 데이터를 벡터화하는 작업이 필요하다. 데이터베이스에 저장되어 있는 원시 데이터의 테이블 구조는 [표 3.2]와 같다.

원시 시공간 데이터 테이블		
필드명	데이터 타입	설명
VALID	VARCHAR2(20)	포인트의 유효성 여부
ID	NUMBER(4,0)	이동객체의 ID
TIME	NUMBER (18, 6)	이동객체가 움직인 시각 (타임-스탬프)
X	NUMBER (18,5)	이동객체의 위치 - x 좌표값
Y	NUMBER (18,5)	이동객체의 위치 - y 좌표값

[표 3.2] 원시 시공간 데이터 테이블 정의

n 프레임(frame)동안 움직인 이동 객체 i 는 다음과 같이 n 개의 흐름 벡터(Flow Vector)의 집합 Qi 로 표현된다.

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

하나의 흐름벡터는 다음과 같이 4 가지 요소로 이루어진다..

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

x 와 y 는 이동 객체의 특정시간에서의 x 좌표값과 y 좌표값이고, dx 와 dy 는 객체가 특정 시간에 x 축으로 움직인 순간 속도와 y 축으로 움직인 순간 속도이다.

본 논문에서는 전처리 모듈을 통해 dx 와 dy 를 계산한다. 일반적으로 속도는 거리 변화를 시간변화로 나누어서 계산하여 절대속도로 구한다. 이렇게 구해진 절대속도는 1 보다 큰 값이 될 수도 있으며, 이 경우 SOM 에 적용하면 0 과 1 사이의 값을 가지는 x, y 에 비해 1 보다 큰 값을 가질 수 있는 순간 속도 dx, dy 에 더욱 민감하게 반응할 수 있다. 그러므로 본 논문에서는 dx 와 dy 를 절대속도가 아닌 상대속도로 계산한다. 즉 dx 는 각 데이터에 대해 x 축으로 움직인 절대속도를 구한 후 전체 데이터의 최대

속도로 나누어서 계산한다. dy 역시 y 축으로 움직인 데이터에 대해 동일한 방법으로 계산한다.

3.4.3 SOM 모듈

전처리 모듈을 통해 계산된 데이터는 [표 3.3]과 같은 구조의 테이블에 저장되어 K-means 와 SOM 에 적용된다.

시공간 데이터 테이블		
필드명	데이터 타입	설명
ID	Number(4,0)	이동객체의 ID
TIME	Number(18, 6)	이동객체가 움직인 시각 (타임-스탬프)
X	Number(18,5)	이동객체의 위치 - x 좌표값
Y	Number(18,5)	이동객체의 위치 - y 좌표값
DX	Number(18,5)	이동객체가 x 축으로 움직인 순간 속도
DY	Number(18,5)	이동객체가 y 축으로 움직인 순간 속도
CLUSTER	Number(4,0)	클러스터링 결과

[표 3.3] 시공간 데이터 테이블 정의

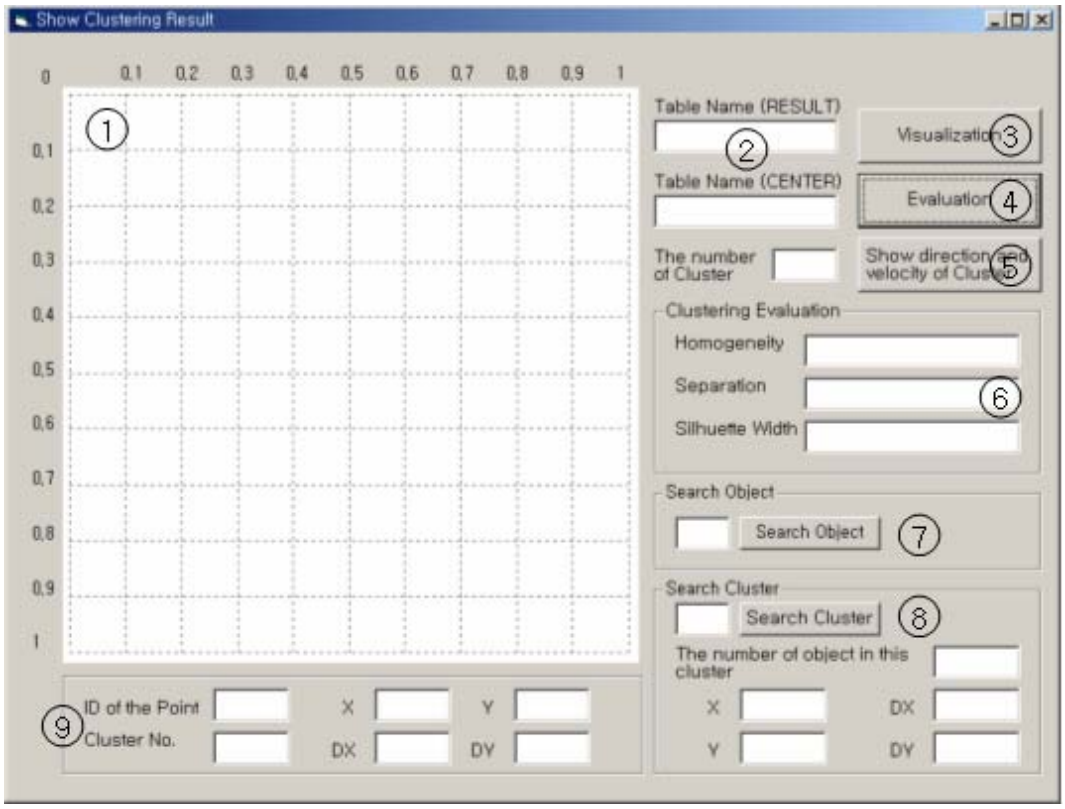
본 논문에서 구현한 SOM 모듈의 동작순서는 다음과 같다. 먼저 데이터베이스에 접속한 후 K 를 결정하고 초기 출력노드의 가중치를 설정하여 네트워크를 형성한다. 그리고 주어진 훈련횟수만큼 네트워크를 훈련시킨 후 출력노드의 최종 가중치의 값을 데이터베이스에 저장한다. 그리고 가시화 모듈에서 언제든지 클러스터링 된 결과를 보여주기 위해 각 데이터에 대한 클러스터 번호 즉, 각 데이터에 대해 최종적으로 선

택된 출력노드의 번호를 데이터베이스에 저장한다.

본 논문에서는 입력벡터들 중에서 K 개를 임의로 선택하여 SOM 의 초기 출력노드의 가중치로 한다. 훈련횟수는 일반적으로 신경망 알고리즘에서 수행되고 있는 훈련횟수인 10000 번으로 정한다.

3.4.4 가시화 모듈

[그림 3.2]는 K-means 과 SOM 의 클러스터링 결과를 시각적으로 보여주기 위해 구현한 가시화 모듈의 사용자 인터페이스 메인 화면이다.



[그림 3.2] 가시화 모듈의 사용자 인터페이스 메인 화면

1 : 클러스터링 결과를 점 및 화살표의 형태로 보여준다

2 : 클러스터링 결과를 가시화하기 위해 클러스터링 결과가 저장된 두 가지 테이블의 이름을 입력한다

3 : 3 번 버튼을 클릭하면 해당 테이블에 저장된 클러스터링 결과를 1 번 화면에 점 형태로 보여준다.:

4 : 4 번 버튼을 클릭하면 성능 평가 기준 중 균질도, 분리도, 반면영상 너비를 계산한다.

5 : 5 번 버튼을 클릭하면 각 클러스터의 방향과 속도크기를 화살표로 보여준다

6 : 계산된 균질도, 분리도, 반면영상 너비의 결과값을 보여준다

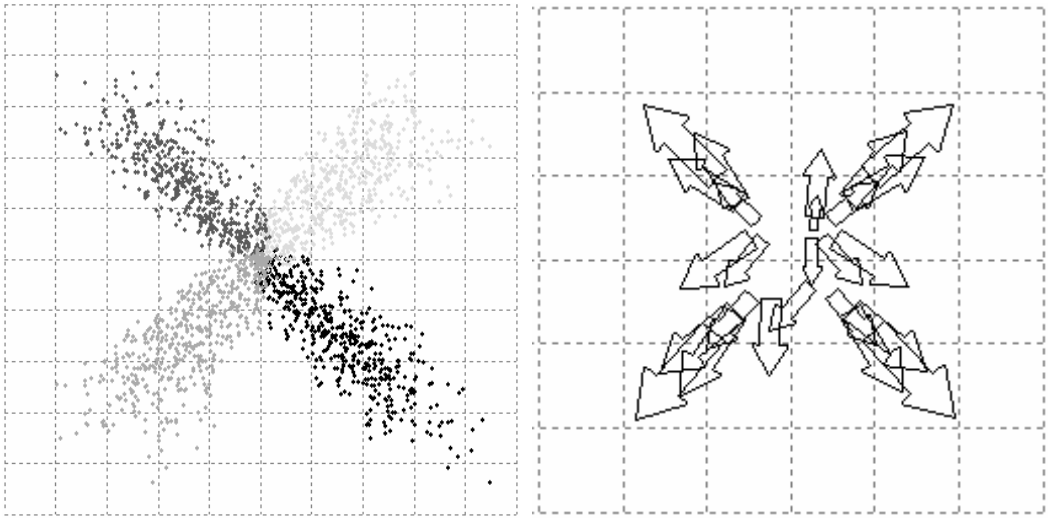
7 : 찾고자 하는 객체의 ID 번호를 입력하면 1 번 화면에 해당 객체가 움직인 궤적을 점과 직선의 형태로 시간순서로 보여준다

8 : 찾고자 하는 객체의 클러스터 번호를 입력하면 1 번 화면에 해당 클러스터에 속하는 포인트만 보여주고 8 번에 그 클러스터에 속하는 포인트 개수와 그 클러스터의 중심점을 보여준다.

9 : 1 번 화면에서 상세히 알고자 하는 포인트를 클릭하면 9 번에 그 포인트의 객체 ID, 속해 있는 클러스터 번호, 포인트의 x, y, dx, dy 값을 보여준다

본 논문에서 가시화 모듈은 시공간 데이터의 특성을 고려하여 클러스터링 결과를 보다 쉽고 정확하게 보여주도록 구현하였다. 클러스터별로 색상을 달리 표현하여 클러스터들간의 구분을 눈으로 쉽게 식별할 수 있고, 좀 더 정확하게 특정 클러스터에 속하는 포인트를 볼 수 있도록 클러스터 검색기능을 추가하였으며, 전체적으로 가시

화된 이미지만으로는 객체구분이 어려운 단점을 보안하기 위해 객체 검색 기능을 추가하여 해당 객체가 움직인 궤적을 보여주도록 하였다. 그리고 화살표의 방향과 크기를 통해 클러스터의 특성을 쉽게 파악할 수 있도록 하였다. [그림 3.3]은 입력데이터와 그 입력데이터에 대해 SOM 을 수행한 후 출력노드의 최종 가중치를 가시화 모듈을 통해 가시화한 한 예이다.



[그림 3.3] 입력데이터와 SOM 출력노드의 최종 가중치 가시화

[그림 3.3]과 같이 출력노드의 가시화를 통해 클러스터의 개수와 각 클러스터의 특성 즉, 각 클러스터 중심점의 위치와 방향, 속도크기를 파악 할 수 있다.

Ⅳ. 성능 평가

본 장에서는 시공간 데이터에 대한 클러스터링 알고리즘의 성능 평가를 위해 실험 데이터를 생성하고, 전 장에서 구현한 SOM 모듈과 S-PLUS 의 K-means 모듈과 응집 계층 클러스터링 모듈을 기반으로 균질도, 분리도, 반면영상 너비, 정확도의 네 가지 기준에서 성능을 비교한다.

4.1 실험 시공간 데이터

본 절에서는 기존 연구에서 다루고 있는 시공간 데이터에 대해 살펴보고, GSTD(Generate Spatio-Temporal Data) 툴을 이용하여 각 사례별로 시공간 데이터를 생성한다.

4.1.1 기존 시공간 데이터들

2 장에서 언급한 바와 같이 기존의 시공간 데이터 마이닝 연구들은 이미 축적되어 있는 방대한 양의 자연 과학 데이터 혹은 이동 객체들을 추적하는 감시 혹은 모니터링 시스템으로부터 수집된 시공간 데이터들을 기반으로 하고 있다. 보통 지구 지질학 데이터를 다루는 자연 과학 시스템 경우 속도, 기온, 산소 혼합률, 표면 온도, 지면 온도, 교류 경계층의 깊이와 같은 지질학 관련 다차원 정보로 이루어진 벡터 데이터를 대상으로 한다. 이러한 데이터들은 일반적으로 2 차원의 위치 정보와 시간 정보를 포함한 3 차원 혹은 그 이상의 표현 형식을 취하고 있는데, [4]에서 사용된 이동 객체 데이터의 경우 n 개의 2 차원 이미지 좌표로부터 수집된 각 이동 객체에 대한 n 개의 x, y 좌표, 그리고 각 연속되는 좌표간의 차이를 통해 계산해 낸 순간 속도의 4 차원 벡

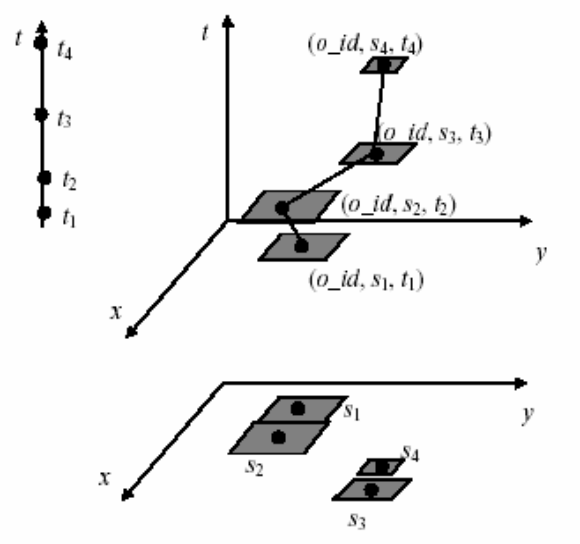
터로 이루어져 있다. 실제로 이러한 데이터들을 대상으로 한 데이터 마이닝의 경우 결과를 응용 분야의 문제 해결 방법으로 바로 적용할 수 있다는 이점이 있지만, 대부분 어느 정도 정확한 마이닝 결과를 얻기 위해서는 복잡한 데이터 사전 처리가 필수적이다. 본 논문에서는 클러스터링 결과의 정확한 평가를 위해 위와 같은 실제 응용 시스템으로부터 얻어진 데이터가 아닌 GSTD 라는 시공간 데이터 생성기를 통해 생성한 데이터를 사용하였다.

4.1.2 GSTD (Generate Spatio-Temporal Data)

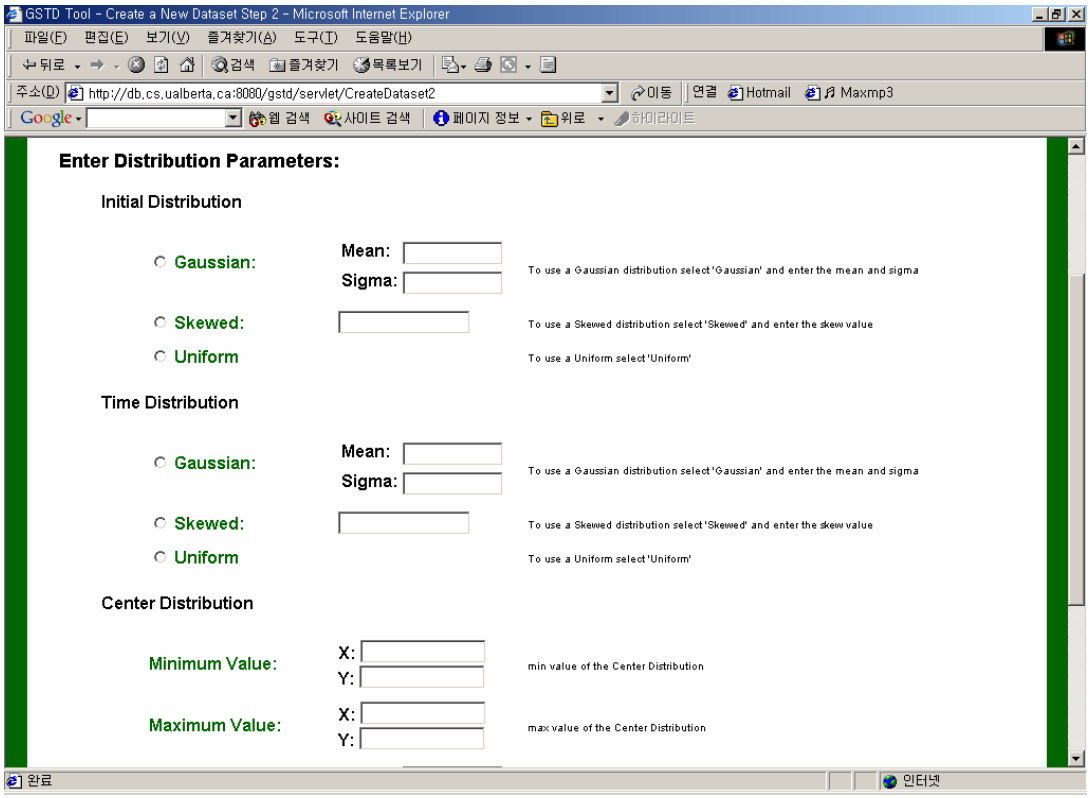
GSTD 는 캐나다의 Alberta 대학에서 개발한 시공간 데이터 생성기로서, 웹상에서 사용자가 직접 파라미터를 입력하면 다양한 형태와 움직임의 시공간 데이터를 XML 형태로 생성해준다[5]. GSTD 를 이용해서 point 타입의 객체와 rectangle 타입의 객체를 생성할 수 있는데, 본 논문에서는 point 타입의 객체를 생성하여 실험하였다.

GSTD 가 생성하는 데이터는 이동 객체의 ID 인 oid 에 의해 식별되는

의 집합이다. 는 시간이 일 때의 객체 o 의 위치이다. 이 때 와 는 모두 0 과 1 사이의 값을 가진다. [그림 4.1]은 시간이 진행됨에 따라 변하는 객체의 위치를 직선으로 연결하여 객체 이동의 연속성을 나타내고 있다.



[그림 4.1] 객체 이동의 연속성



[그림 4.2] GSTD 웹사이트 화면

[그림 4.2] 는 GSTD 웹사이트[25]에서 생성하고자 하는 시공간 데이터의 특성을 결정하기 위해 파라미터를 설정하는 화면이다. 분포 파라미터(distribution parameter)를 통해 객체들의 시작점 분포, 시간 분포, 중심점의 분포 등을 결정하여 다양한 특성을 가지는 시공간 데이터를 생성한다. 그리고 GSTD 는 생성한 시공간 데이터의 움직이는 모습을 가시화를 통해 확인할 수 있도록 한다.

4.1.3 실험 데이터의 특성

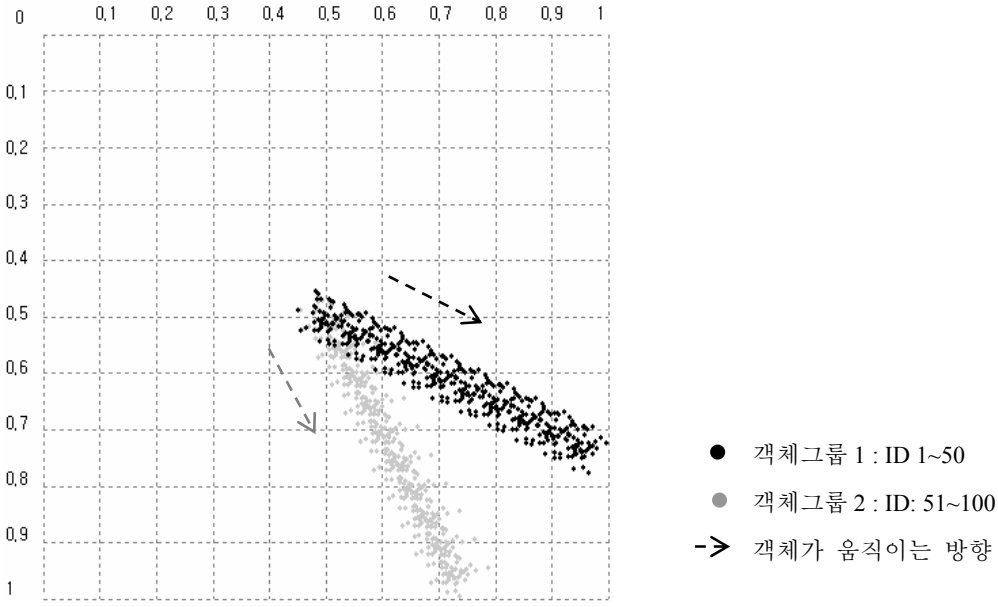
성능 평가를 위한 실험 데이터는 객체가 움직이는 방향과 속도를 기준으로 하여 다음과 같이 크게 3 가지의 사례로 분류한다.

[사례 A] 비슷한 속도로 움직이면서, 서로 다른 방향으로 움직이는 2 개의 객체 그룹 2 개의 객체그룹 1, 객체그룹 2 에 속한 객체의 수는 각각 50 개, 각 그룹의 포인트 수는 각각 500 개이며, 교차되는 위치에 따라 다음과 같이 두 가지로 분류하여 실험하였다.

A-1. 2 개의 객체그룹의 출발 위치가 동일한 경우

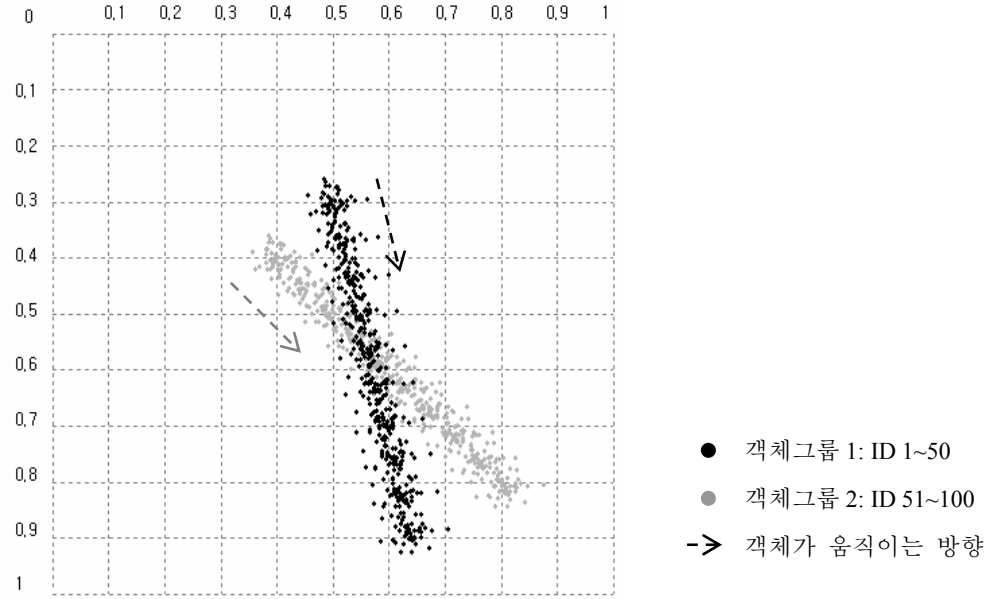
성능비교를 위해 두 객체 그룹간의 각도를 중심으로 방향 차이를 구분하였다. 두 객체 그룹간의 각도차이는 90 도, 75 도, 60 도, 45 도, 30 도로 하여 실험하였다.

[그림 4.3]은 두 객체 그룹간의 방향차이가 45 도일 때의 실험 시공간 데이터를 보여준다.



[그림 4.3] 출발위치가 동일한 시공간 데이터

A-2. 2 개의 객체그룹의 출발 위치는 다르지만, 궤적이 서로 교차되는 경우

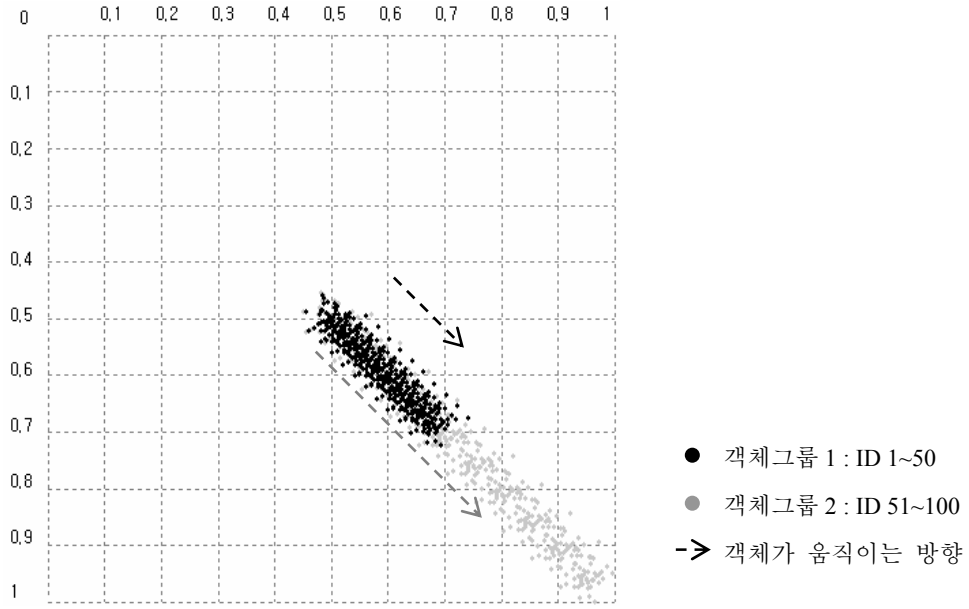


[그림 4.4] 궤적이 서로 교차되는 시공간 데이터

[사례 B] 동일한 방향으로 움직이면서, 서로 다른 두 속도로 움직이는 객체 그룹

2 개의 객체그룹 1, 객체그룹 2 에 속한 객체의 수는 각각 50 개, 각 그룹의 포인트 수는 각 500 개, 347 개이다. 두 객체그룹간의 평균속도 차이에 변화를 주면서 실험하였다. 두 객체그룹간의 속도차이는 1.0/s 에서 0.1/s 까지 0.1/s 씩 감소시킨다.

[그림 4.5]은 두 객체그룹간의 속도차이가 0.6/s 일 때의 시공간 데이터를 보여준다. 즉 객체그룹 1 의 평균속도는 0.4/s 이고, 객체그룹 2 의 평균속도는 1.0/s 이다.

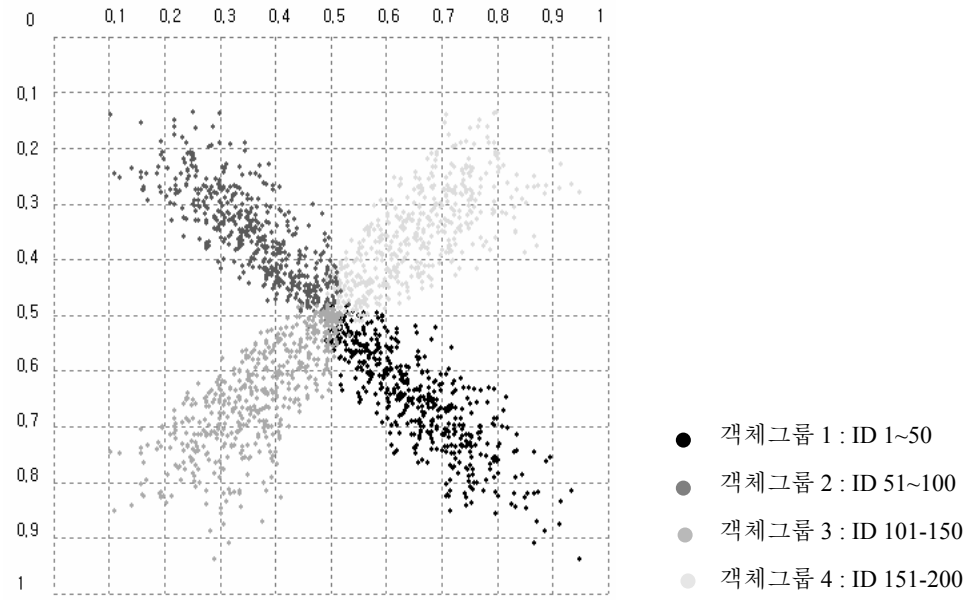


[그림 4.5] 서로 다른 속도로 움직이는 시공간 데이터

[사례 C] 움직임이 불규칙한 객체 그룹 - 다양한 속도로 움직이면서 방향이 서로 다른 4 개의 객체 그룹

4 개의 객체그룹 1, 객체그룹 2, 객체그룹 3, 객체그룹 4 에 속한 객체의 수는 각각 50 개, 각 그룹의 포인트 수는 각각 530 개이다. 4 개의 객체그룹은 모두 좌표의 중심으로

부터 각 모서리를 향해 움직인다.



[그림 4.6] 움직임이 불규칙한 시공간 데이터

4.2 성능 평가 기준

본 절에서는 본 논문에서 클러스터링 결과를 비교하기 위해 적용한 4 가지의 성능 평가 기준에 대해서 설명한다.

4.2.1 균질도(Homogeneity)

균질도는 클러스터의 중심점과 그 클러스터에 속하는 포인트들간의 평균거리로 계산한다[22]. 균질도의 수치가 클수록 클러스터링이 잘되었다고 판단한다. 균질도를 구하는 식은 다음과 같다.

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

D 는 거리함수, p_i 는 i번째 포인트, $C(p_i)$ 는 p_i 가 속한 클러스터의 중심점, N_{point} 는 전체 포인트 수이다.

4.2.2 분리도(Separation)

분리도는 클러스터의 중심들간의 평균 거리로 계산한다[22] 분리도의 수치가 작을수록 클러스터링이 잘되었다고 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_i 와 C_j 는 i번째 클러스터와 j번째 클러스터의 중심점, N_{ci} 와 N_{cj} 는 i번째 클러스터와 j번째 클러스터에 속한 포인트의 수이다.

4.2.3 반면영상 너비(Silhouette Width)

반면영상 너비는 클러스터링 결과의 전체 질(quality)을 반영한다[22]. 반면영상 너비는 계산된 수치가 클수록 클러스터링이 잘되었다고 판단한다. 반면영상 너비를 구하는 식은 다음과 같다.

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.

4.2.4 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_i 는 i번째 클러스터에 정확하게 클러스터링 된 포인트 수이다.

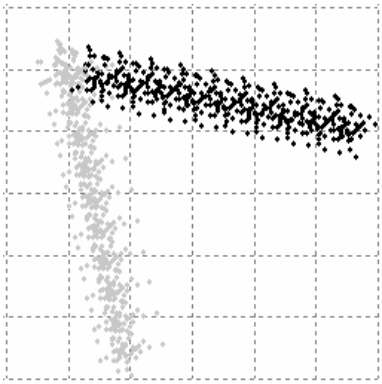
4.3 성능 평가 결과

본 절에서는 각 사례들에 대한 성능 평가 결과를 분석한다. 그리고 [사례 A-1]와 [사례 B]에 대한 각 알고리즘의 임계값(threshold)을 구하여 시공간 데이터를 위한 클러스터링 알고리즘의 성능을 비교한다. 임계값은 각 알고리즘의 클러스터링 성능이 우수할 때까지의 속성 차이값이며, 속성은 방향과 속도이다.

[사례 A] 비슷한 속도로 움직이면서, 서로 다른 방향으로 움직이는 2 개의 객체 그룹

A-1. 두 객체그룹의 출발 위치가 동일한 경우

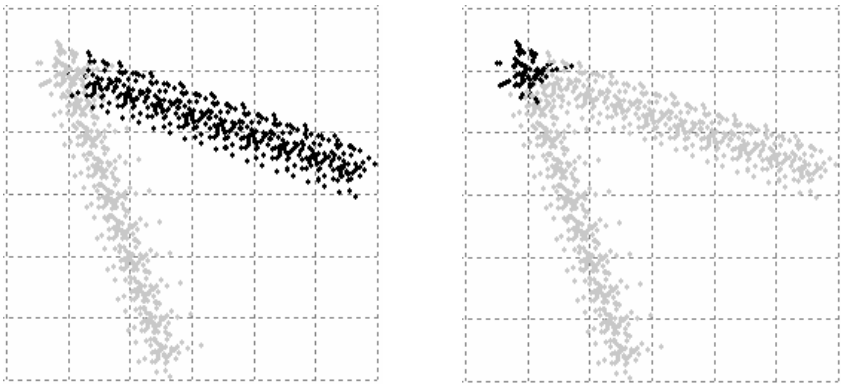
[그림 4.7]은 방향 차이가 75 도일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 방향 차이가 75 도 이상일 때는 각 알고리즘의 클러스터링 결과가 동일하다.



SOM, K-means, 평균 연결법, 워드방법

[그림 4.7] 방향 차이가 75 도일 때 클러스터링 결과의 가시화

[그림 4.8]은 방향 차이가 60 도일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 가시화된 결과를 보면, SOM 과 워드방법은 비교적 클러스터링 결과가 우수하지만 K-means 과 평균 연결법은 방향 차이가 60 도인 두 객체그룹을 제대로 클러스터링 하지 못한다는 것을 알 수 있다.

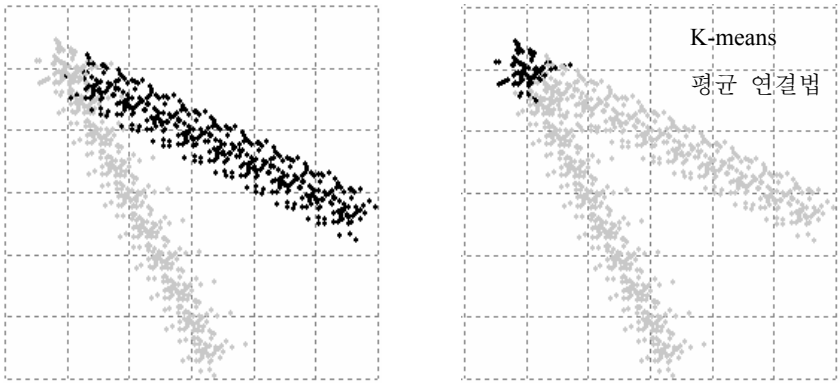


SOM, 워드방법 K-means, 평균 연결법

[그림 4.8] 방향 차이가 60 도일 때 클러스터링 결과의 가시화

[그림 4.9]은 방향 차이가 45 도일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 방향 차이가 60 도일 때와 결과가 동일하였다. 즉 SOM 과 워드

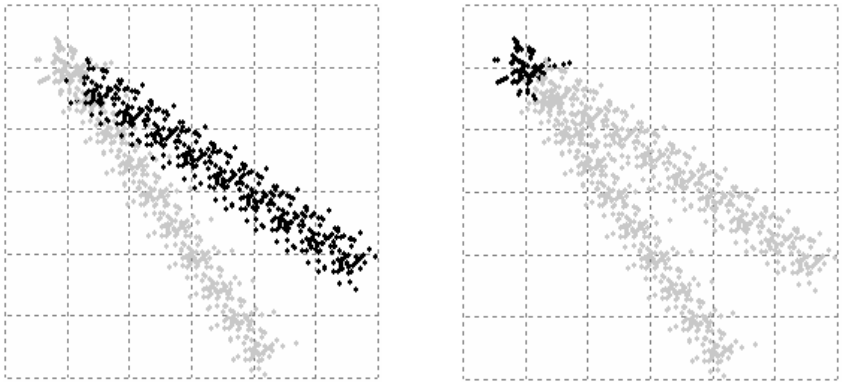
방법의 클러스터링 성능이 우수하였다.



SOM, 워드방법 K-means, 평균 연결법

[그림 4.9] 방향차이가 45 도일 때 클러스터링 결과의 가시화

[그림 4.10]은 방향차이가 30 도일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 방향 차이가 30 도일 때는 SOM 만이 비교적 정확하게 클러스터링을 수행함을 확인할 수 있었다.



SOM K-means, 평균 연결법, 워드방법

[그림 4.10] 방향차이가 30 도일 때 클러스터링 결과의 가시화

이 실험을 통해 방향 차이에 따른 각 알고리즘의 임계값을 구할 수 있었다. 여기서 임계값은 각 알고리즘의 클러스터링 성능이 우수할 때까지의 방향 차이를 의미한

다. SOM 의 임계값은 30 도, K-means 와 평균 연결법의 임계값은 75 도, 워드방법의 임계값은 45 도이다. 즉, 방향 차이가 작은 경우에도 다른 클러스터링 알고리즘보다 SOM 의 클러스터링 성능이 우수하였다.

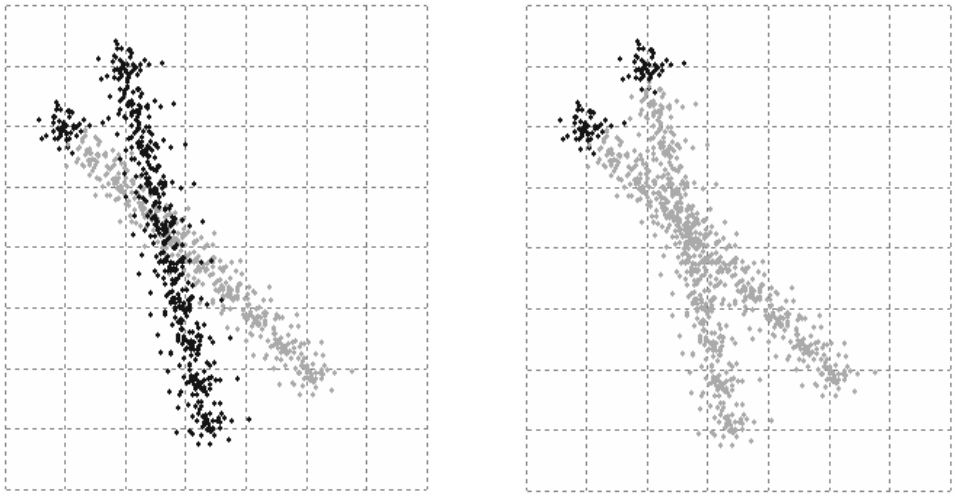
[표 4.1]은 사례 A-1 의 클러스터링 결과의 성능 평가 수치이다. 가시화된 결과와 성능 평가 수치가 일치하지 않음을 볼 수 있다. 즉 방향 차이가 30 도인 경우에 실제로 SOM 의 클러스터링 성능이 우수함에도 불구하고, 분리도와 반면영상너비 수치상으로는 이러한 결과를 제대로 반영하고 있지 못하고 있다. 이것은 균질도와 분리도, 반면영상너비는 실험 데이터의 특성을 고려하지 않고 클러스터링 결과만을 반영하는 수치이기 때문이다. 반면에 데이터의 특성을 고려하여 클러스터링 결과를 평가하는 정확도에서는 SOM 이 다른 알고리즘보다 우수하였다.

방향 차이	알고리즘	균질도	분리도	반면영상너비	정확도
75 도	SOM	0.2425	1.0706	0.7335	0.95
	K-means	0.2425	1.0706	0.7335	0.95
	평균 연결법	0.2425	1.0706	0.7335	0.95
	워드방법	0.2425	1.0706	0.7335	0.95
60 도	SOM	0.2510	0.8913	0.6917	0.95
	K-means	0.4520	0.9755	0.5327	0.5
	평균 연결법	0.4520	0.9755	0.5327	0.5
	워드방법	0.2510	0.8913	0.6917	0.95
45 도	SOM	0.2655	0.6908	0.6157	0.95
	K-means	0.3547	1.0919	0.6410	0.5
	평균 연결법	0.3547	1.0919	0.6410	0.5
	워드방법	0.2655	0.6908	0.6157	0.95
30 도	SOM	0.2868	0.4782	0.4595	0.95
	K-means	0.2534	1.2230	0.7491	0.5
	평균 연결법	0.2534	1.2230	0.7491	0.5
	워드방법	0.2534	1.2230	0.7491	0.5

[표 4.1] 사례 A-1 의 클러스터링 결과의 성능 평과 수치

A-2. 두 객체그룹의 출발 위치는 다르지만, 궤적이 서로 교차되는 경우

[그림 4.11] 은 서로 궤적이 교차되는 두 객체 그룹에 대해 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 가시화된 결과를 보면, SOM 과 워드방법은 비교적 클러스터링 결과가 우수하지만 K-means 과 평균 연결법은 두 객체그룹을 제대로 클러스터링 하지 못한다는 것을 알 수 있다.



SOM, 워드방법 K-means, 평균 연결법

[그림 4.11] 사례 A-2 의 클러스터링 결과의 가시화

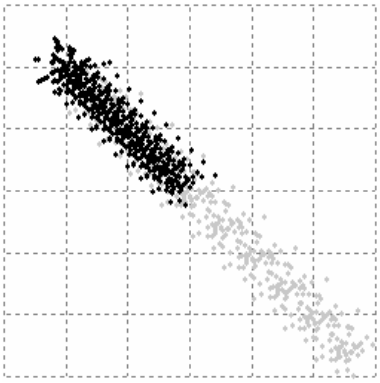
[표 4.2]은 사례 A-2 의 클러스터링 결과의 성능 평가 수치이다. 성능평가의 내적 기준인 정확도 측면에서 SOM 과 워드방법이 우수하였다.

알고리즘	균질도	분리도	반면영상 너비	정확도
SOM	0.2731	0.6719	0.5564	0.95
K-means	0.3459	1.0394	0.6108	0.5
평균 연결법	0.3459	1.0394	0.6108	0.5
워드방법	0.2731	0.6719	0.5564	0.95

[표 4.2] 사례 A-2 의 클러스터링 결과의 성능 평과 수치

[사례 B] 동일한 방향으로 움직이면서, 서로 다른 속도로 움직이는 2 개의 객체 그룹

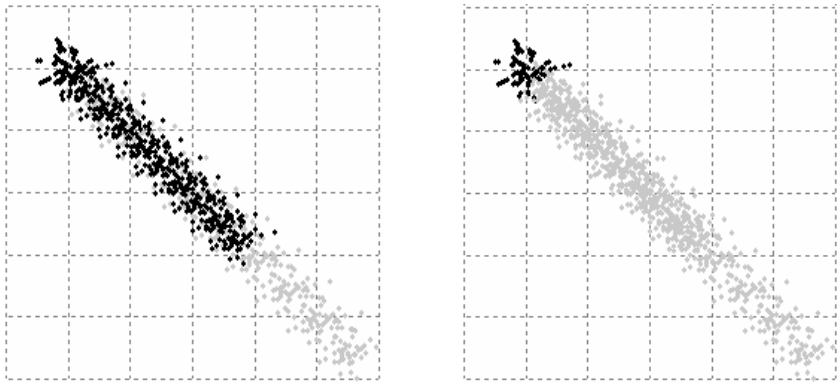
[그림 4.12]은 속도 차이가 0.6/s 일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 속도 차이가 0.6/s 이상일 때는 각 알고리즘의 클러스터링 결과가 동일하다. \



SOM, K-means, 평균 연결법, 워드방법

[그림 4.12] 속도차이가 0.6/s 일 때 클러스터링 결과의 가시화

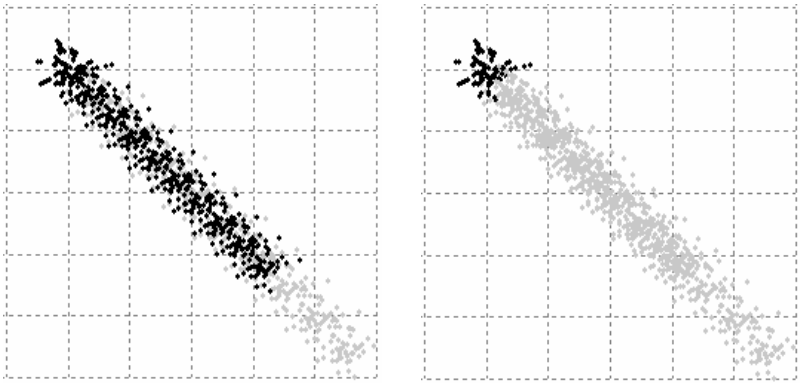
[그림 4.13]은 속도 차이가 0.4/s 일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 가시화된 결과를 보면, SOM 과 워드방법은 비교적 클러스터링 결과가 우수하지만 K-means 과 평균 연결법은 속도 차이가 0.4/s 인 두 객체그룹을 제대로 클러스터링 하지 못한다는 것을 알 수 있다.



SOM, 워드방법　　K-means, 평균 연결법

[그림 4.13] 속도차이가 0.4/s 일 때 클러스터링 결과의 가시화

[그림 4.14]은 속도 차이가 0.3/s 일 때 SOM, K-means, 평균 연결법, 워드방법의 클러스터링 결과를 보여준다. 속도 차이가 0.3/s 일 때는 SOM 만이 비교적 정확하게 클러스터링을 수행함을 확인할 수 있었다.



SOM　　K-means, 평균 연결법, 워드방법

[그림 4.14] 속도차이가 0.3/s 일 때 클러스터링 결과의 가시화

이 실험을 통해 속도 차이에 따른 각 알고리즘의 임계값을 구할 수 있었다. 여기서 임계값은 각 알고리즘의 클러스터링 성능이 우수할 때까지의 속도 차이를 의미한다. SOM 의 임계값은 0.3/s, K-means 와 평균 연결법의 임계값은 0.4/s, 워드방법의 임계

값은 0.3/s 이다. 즉, 속도 차이가 작은 경우에도 다른 클러스터링 알고리즘보다 SOM 의 클러스터링 성능이 우수하였다.

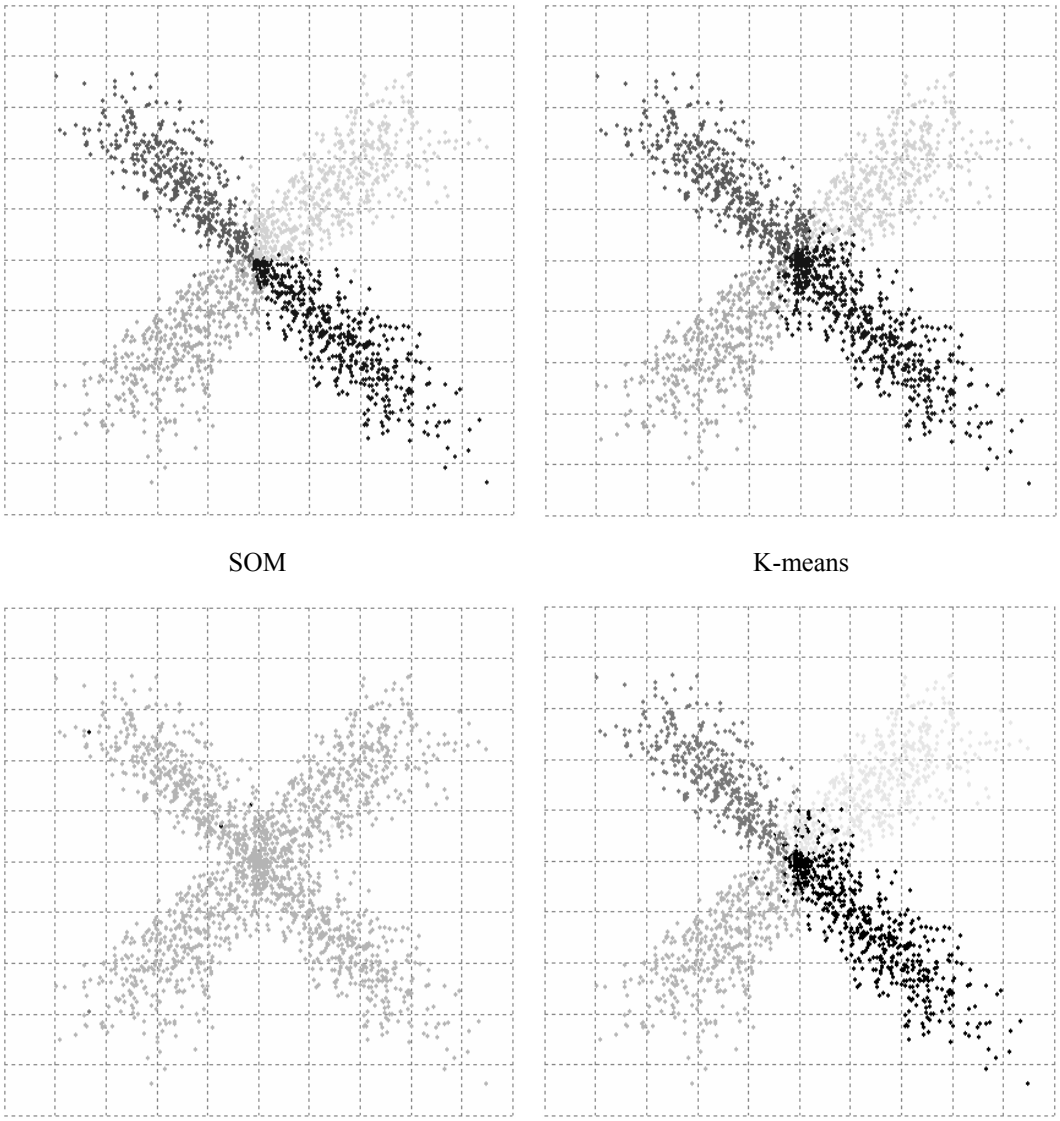
[표 4.3]은 사례 B 의 클러스터링 결과의 성능 평가 수치이다. 데이터의 특성을 고려한 내적 기준인 정확도 측면에서 SOM 이 다른 알고리즘보다 작은 속도 차이까지 구분하여 객체 그룹을 클러스터링할 수 있다는 것을 알 수 있다.

속도차이	알고리즘	균질도	분리도	반면영상너비	정확도
6/s	SOM	0.1784	0.9795	0.7777	0.9
	K-means	0.1784	0.9795	0.7777	0.9
	평균 연결법	0.1784	0.9795	0.7777	0.9
	워드방법	0.1784	0.9795	0.7777	0.9
4/s	SOM	0.2304	0.7412	0.6298	0.9
	K-means	0.2956	1.1648	0.6808	0.5
	평균 연결법	0.2956	1.1648	0.6808	0.5
	워드방법	0.2304	0.7412	0.6298	0.9
3/s	SOM	0.2564	0.6219	0.5092	0.9
	K-means	0.2414	1.2373	0.7521	0.5
	평균 연결법	0.2414	1.2373	0.7521	0.5
	워드방법	0.2414	1.2373	0.7521	0.5

[표 4.3] 사례 B 의 속도차이에 따른 클러스터링 결과의 성능평가 수치

[사례 C] 움직임이 불규칙한 객체 그룹 - 다양한 속도로 움직이면서 방향이 서로 다른 4 개의 객체 그룹

[그림 4.15] 은 각 객체들이 객체그룹별로 서로 다른 방향을 향해 다양한 속도로 불규칙하게 움직이는 4 개의 객체그룹에 대해 SOM 과 K-means 를 수행한 결과이다. 즉 하나의 객체그룹은 하나의 방향을 향하여 움직이는 객체들로 구성되지만 그 객체 각각은 다양한 속도로 불규칙하게 움직인다. [표 4.4] 를 보면 SOM, K-means, 워드방법의 성능에 큰 차이는 없지만 [그림 4.15]에서 좌표의 중심부분, 즉 각 객체들이 움직임을 시작하는 위치에서는 SOM 이 K-means 와 워드방법보다 더 정확한 클러스터링 결과가 나온다는 것을 보여준다.



[그림 4.15] 사례 C 의 클러스터링 결과의 가시화

[표 4.4]은 사례 C 의 클러스터링 결과의 성능 평가 수치이다. 균질도, 분리도, 반면 영상너비와 같은 외적 기준측면에서는 SOM 보다 다른 클러스터링 알고리즘이 더 우수하지만 실제 클러스티링 결과의 가시화와 정확도측면에서는 SOM 이 더 우수하다.

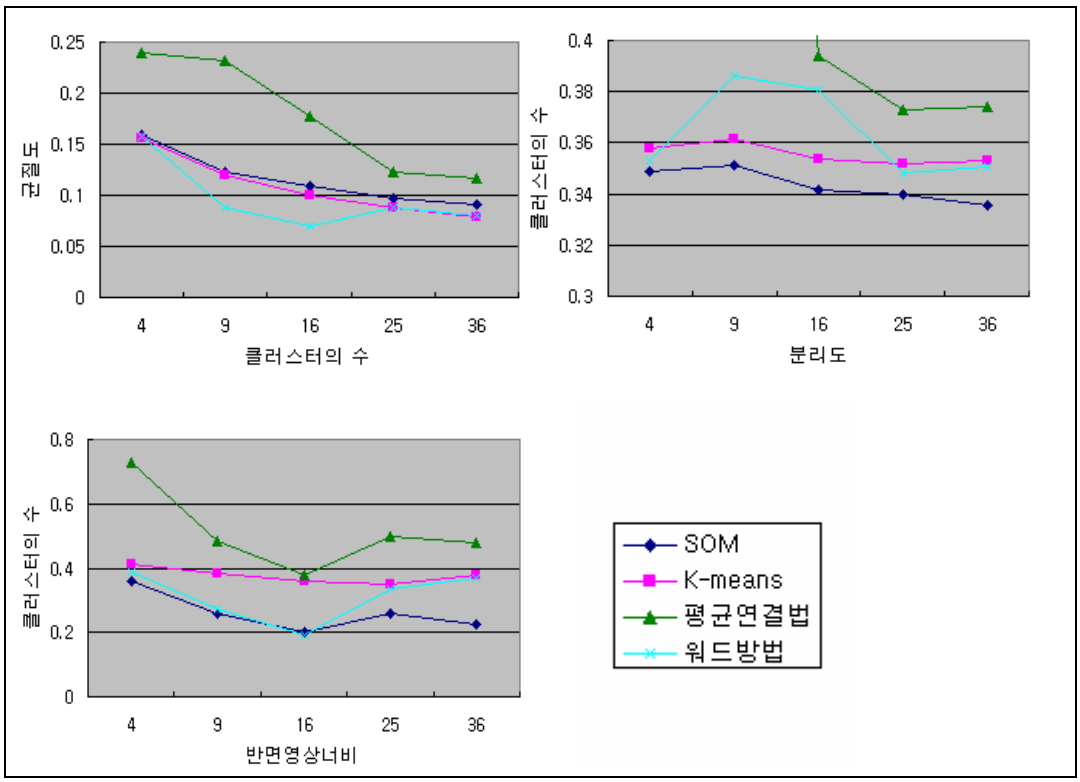
	균질도	분리도	반면영상 너비	정확도
SOM	0.1590	0.3489	0.3584	0.9
K-means	0.1556	0.3579	0.4111	0.85
평균 연결법	0.2393	1.1060	0.7293	0.25
워드 방법	0.1573	0.3531	0.3890	0.85

[표 4.4] 사례 C 의 클러스터링 결과의 성능 평가 수치

SOM 과 K-means, 그리고 2 가지의 응집 계층 알고리즘인 평균 연결법과 워드방법은 객체들간의 속성에 큰 차이가 있는 경우에는 모두 성능이 우수하지만 객체들간의 작은 속성 차이까지 구분하여 클러스터링을 하는 알고리즘은 SOM 이다. 하지만 SOM 도 아주 작은 속성의 차이 즉, 방향 차이가 30 도 보다 작거나, 속도 차이가 0.3/s 보다 작을 경우에는 제대로 클러스터링을 하지 못했다.

SOM, K-means, 응집 계층 알고리즘의 클러스터링 성능을 보다 더 자세히 비교하기 위하여 하나의 시공간 데이터 집합에 대해 클러스터 수를 변경하면서 클러스터링을 수행해보았다.

. [그림 4.16]은 사례 C의 데이터를 K=4, 9, 16, 25, 36 개로 클러스터링한 결과에 대해 균질도, 분리도, 반면양상 너비를 계산한 결과를 그래프로 나타낸 그림이다.

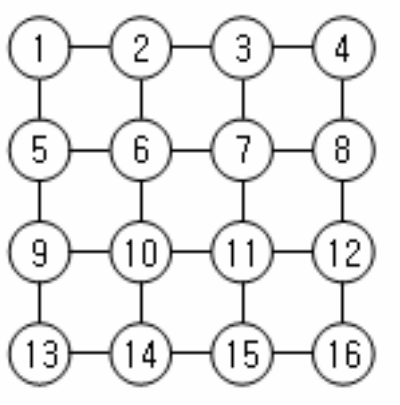


[그림 4.16] 사례 C의 성능 평가 기준치의 그래프(K=4,9,16,25,36)

[그림 4.16]에서 볼 수 있듯이 클러스터의 수가 다양한 경우에도 균질도, 분리도, 반면영상 너비의 수치상으로는 평균 연결법이 가장 좋은 결과가 나왔다. 그 이유는 이 세가지 성능 평가 기준은 입력데이터의 특성을 고려하지 않고 결과만을 반영하여 계산하기 때문이다. 하지만 실제로 클러스터링 결과를 가시화하면 앞에서 살펴보았듯이 SOM이 시공간 데이터의 특성을 고려하여 클러스터링을 수행함을 알 수 있다. 그리고 K-means와 두 개의 응집 계층 알고리즘인 평균 연결법과 워드방법의 클러스터링 결과를 통해서는 클러스터 상호간의 관계를 파악할 수 없다. 반면에 SOM은 클러스터간의 관계뿐만 아니라 클러스터의 위상을 통해 입력 데이터의 위상도 파악할 수 있어 클러스터링 결과를 이용하여 효과적으로 분류화(Classification)

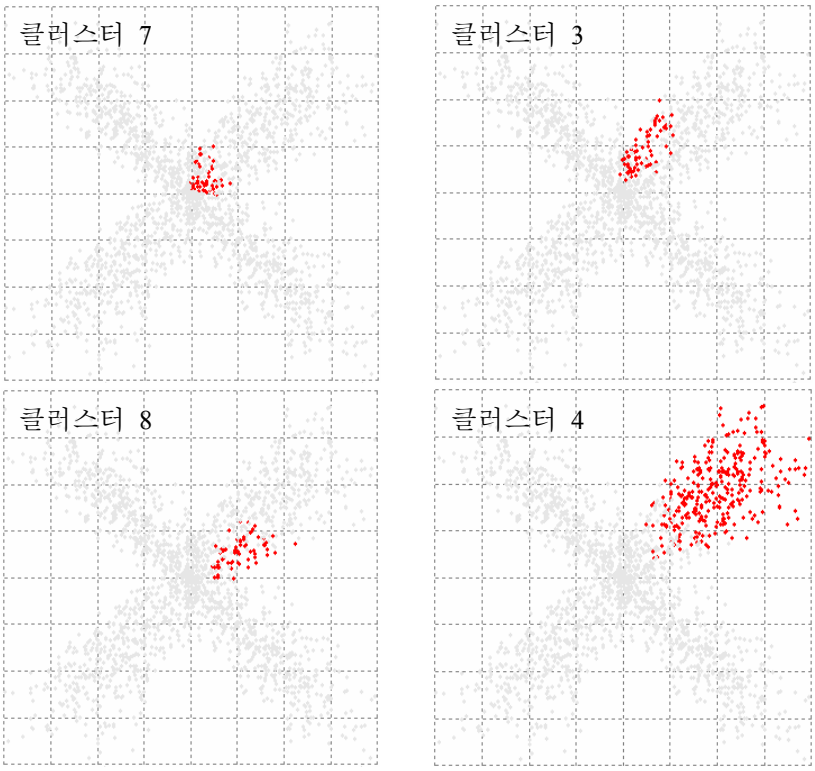
작업을 할 수 있다.

[그림 4.17]은 사례C의 데이터에 대해 K=16인 SOM을 수행했을 때 훈련된 네트워크 구조이다.



[그림 4.17] 사례 C의 SOM 네트워크(K=16)

이렇게 훈련된 네트워크의 하나의 출력노드는 하나의 클러스터를 의미하므로 위상적으로 가까이 있는 4개의 클러스터 3, 4, 7, 8을 가시화하면 [그림 4.18]과 같다.



[그림 4.18] 위상적으로 근접한 4개의 클러스터

[그림 4.18]에서 SOM의 출력노드 즉 클러스터의 위상과 입력 데이터의 위상이 일치함을 알 수 있다.

Ⅴ. 결론 및 향후과제

본 논문에서는 기존 시공간 데이터 마이닝에 대한 연구들에서 사용되어 온 알고리즘들 중 패턴 인식과 클러스터링 능력이 뛰어나다고 알려진 SOM 에 대해 분석하여 개발한 SOM 기반 마이닝 모듈과 S-PLUS 의 K-means 모듈과 응집계층 알고리즘 모듈에 대한 클러스터링 결과를 크게 3 가지 사례를 통해 비교하였다. SOM 기반 모듈은 3 가지의 모듈로 구성되어 있다. 데이터 전처리 모듈을 통해 GSTD 툴을 이용하여 생성한 원시 시공간 데이터를 SOM 에 적용할 수 있도록 벡터화하고, 전처리 된 데이터를 SOM 모듈에 적용하여 클러스터링을 수행하였다. 데이터의 시공간 속성을 고려하여 구현된 가시화 모듈을 통해 클러스터링 결과를 분석하였다.

클러스터링 결과의 성능 평가를 위해 균질도, 분리도, 반면영상 너비, 정확도 네 가지 기준치를 적용하였다. 3 가지 사례에 대한 실험을 통해 균질도, 분리도, 반면영상 너비와 같은 클러스터링의 외적 기준 즉, 입력데이터의 특성을 고려하지 않고 클러스터링 결과만을 반영하는 성능 평가의 외적 기준에서는 K-means 와 응집 계층 알고리즘이 SOM 보다 성능이 우수하다고 평가되었지만, 실제로 가시화를 통해 클러스터링 결과를 확인한 결과 SOM 이 K-means 와 응집 계층 알고리즘 보다 더 정확하게 마이닝을 수행하였음을 알 수 있었다. 그리고 성능 평가의 내적 기준인 정확성과 임계값 측면에서도 SOM 이 더 우수한 결과를 보였다. 또한 SOM 은 입력 데이터의 위상을 그대로 반영하기 때문에 분류화에도 효과적으로 이용될 수 있다.

본 논문의 연구내용을 기반으로 SOM, K-means, 응집 계층 알고리즘 외에 시공간 데이터에 대한 다양한 마이닝 알고리즘의 비교 연구와 시공간 데이터의 속성을 고려

하여 마이닝 결과의 성능을 평가할 수 있는 성능 평가 기준에 대한 연구, 그리고 임계값보다 작은 속성 차이를 가지는 객체그룹을 클러스터링 할 수 있는 알고리즘에 대한 연구를 향후 연구 과제로 제시한다.

참고문헌

[1] J.F. Roddick, and M. Spiliopoulou, A bibliography of Temporal, Spatial and Spatio-temporal Data Mining Research. SIGKDD Explorations **1**(1):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 (London: Yaylor and Francis), 2001.

[3] N. Johnson and D Hogg, Learning the Distribution of Object Trajectories for Event Recognition, In Proc. British Machine Vision Conference, vol. 2, 1995.

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

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

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

[7] 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

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

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

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

[11] S. Rogers, P. Langley, and C. Wilson, Mining GPS data to augment road models. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (pp. 104-113). San Diego, CA: ACM Press, 1999

[12] S. Gafney and P. Smyth. Trajectory clustering with mixtures of regression models. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

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

[14] 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

[15] J. Han and M. Kamber. Data Mining: Concepts and Techniques. (to be published by) Morgan Kaufmann, 2000

[16] 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

[17] P. Remagnino, T. Tan, and K. Baker. Agent orientated annotation in model based visual surveillance. In ICCV, 1998

[18] S. Handley, P. Langley and F. Rauscher, Learning to predict the duration of an automobile trip. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (pp. 219--223). New York: AAAI Press, 1998

[19] H. Mannila, H. Toivonen and A. I. Verkamo: Discovering frequent episodes in sequences. Proc. 1st International Conference on Knowledge Discovery and Data Mining (KDD'95), 210--215. AAAI Press 1995

[20] P.C. Juan and L.C. Ignacio, Discovering Similar Patterns in Time Series, In Proc. KDD-2000, 2000

[21] TC Fu, FL Chung, R. Luk and V. Ng, "Pattern Discovery from Stock Time Series Using Self-Organizing Maps," Workshop Notes of KDD2001 Workshop on Temporal Data Mining, 26-29 Aug., San Francisco, pp.27-37, 2001

[22] 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. 2002;12: 241-262.

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

[24] 김대주, 신경망 이론과 응용(Ⅰ), pp169-189, 2001

[25] http://db.cs.ualberta.ca:8080/gstd

ABSTRACT

Performance Comparison of Clustering Technique For Spatio-Temporal Data

Department of Computer Science & Engineering

Ewha Institute of Science and Technology

Kang Na Young

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 such as K-means, Agglomerative Hierarchical algorithm and SOM are commonly applied as spatio-temporal data mining techniques. However, researches on the performance of these approaches when they actually applied to spatio-temporal data mining and on what is the proper data mining algorithm for the input data sets considering the spatio-temporal properties, is sparse at present.

In this thesis, we analyze SOM, which is the popular clustering algorithm applied to clustering

analysis in data mining area, and develop the spatio-temporal data mining module based on it. In addition, we analyze the clustering results of SOM and compare it with those of K-means and Agglomerative Hierarchical algorithm in the aspects of homogeneity, separation, separation, silhouette width and accuracy. We also develop specialized visualization module for more accurate interpretation of mining result. This is because, without considering the properties of spatio-temporal data, numerical criterions of performance evaluation may not show properly the accuracy and performance of clustering results.