

시공간 데이터를 위한 클러스터링 기법의 성능 비교

강주영⁰ 이봉재 송재주 신진호 용환승
한국전력 전력연구원, 이화여자대학교

{jyang⁰, bjee, jjsong, jinho}@kepri.re.kr, hsyong@ewha.ac.kr

Korea Electric Power Research Institute, Ewha Womans University

Juyoung Kang⁰ Bongjae Lee Jaeju Song Jinho Shin Hwanseung Yong
Power Information Technology Group, KEPRI

Dept. of Computer Science and Engineering, EIST, Ewha Womans University

요 약

최근 GPS 시스템, 감시 시스템, 기상 관측 시스템과 같은 다양한 응용 시스템으로부터 수집된 시공간 속성을 가진 데이터를 분석하고자 하는 시공간 데이터 마이닝에 대한 관심이 더욱 높아지고 있다. 기존의 시공간 데이터 마이닝에 대한 연구는 문자-숫자 데이터를 기반의 마이닝 기법을 그대로 적용하고 있기 때문에 데이터의 시공간 속성을 충분히 고려한 분석으로는 한계가 많은 것이 사실이다. 본 논문에서는 패턴 인식과 클러스터링 능력이 뛰어나다고 알려진 SOM을 기반으로 시공간 클러스터링 모듈을 개발하고, 개발된 모듈의 성능과 클러스터링 정확성에 대하여 K-means, 응집 계층 알고리즘(Average Linkage, Ward)과 비교함으로써 시공간 데이터 마이닝을 위한 각 알고리즘들의 성능을 분석하였다. 또한 입력 데이터의 특성과 클러스터링 결과를 더욱 정확하게 나타내어 가시적인 분석을 도울 수 있도록 시공간 데이터 클러스터링을 위한 가시화 모듈을 개발하였다.

1. 서 론

최근 위성, 자연 과학 관측 시스템, 교통, 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 문자-숫자 데이터를 기반으로 하고 있는 반면, 시공간 데이터 마이닝의 대상 데이터는 시간과 공간 속성을 동시에 지니고 있기 때문에 분석 시 이를 적절하게 고려해 주어야만 한다[1]. 또한 기존의 데이터 마이닝과는 달리 지식 탐사의 절차와는 상관없이 입력의 형태나 속성, 결과의 의미에 대한 해석 등을 더욱 중요하게 다루어야 한다는 특징이 있다[2]. 따라서 데이터의 시공간 특성을 고려하지 않고 기존의 마이닝 기법을 그대로 적용하는 방법은 성능과 결과의 정확성 면에 있어서 한계가 있다고 할 수 있다[1].

현재까지 수행된 시공간 데이터 마이닝 연구는 주로 K-means, SOM(Self-Organizing Map), 응집 계층(Agglomerative Hierarchical) 알고리즘과 같은 문자-숫자 기반의 클러스터링 알고리즘들을 기반으로 하고 있는데[3,4], 이러한 알고리즘들이 시공간 데이터 마이닝에 있어서는 실제로 어느 정도의 성능을 보장하며, 데이터의 속성에 따라 수행 성능이 어떻게 변화하는지, 최적의 마이닝 알고리즘 선택 기준은 무엇인지 등에 대한 연구는 미흡한 실정이다. 따라서 시공간 데이터 마이닝이 응용 별로 더욱 정확한 결과를 도출하기 위해서는 기존의 마이닝 기법들을 적용함에 있어서 응용의 특성에 따라 최적의 알고리즘을 선택적으로 적용할 수 있도록 객관적인 기준을 제시할 필요가 있다.

본 논문에서는 선행 연구에서 가장 일반적으로 사용되어 온 SOM을 기반으로 시공간 클러스터링 모듈을 개발하고, 구현된 모듈의 클러스터링 성능을 K-means, Average Linkage, Ward의 세 가지 클러스터링 알고리즘들과 균질도, 분리도, 반면영상

결과를 비교 분석 한다.

본 논문의 구성은 다음과 같다. 1절의 서론에 이어 2절에서는 구현된 SOM 기반 시공간 데이터 시스템의 구성과 각 모듈에 대해 소개하고 3절에서는 성능 평가 환경과 수행 결과에 대해 기술한 후 마지막으로 5절에서 결론을 맺는다.

2. SOM 기반 시공간 데이터 마이닝 시스템

본 논문에서는 SOM을 기반으로 시공간 클러스터링 시스템을 구현하고 이의 성능을 분석하였다. 다음 [그림 1]은 본 논문에서 제안하는 시스템의 전체적인 구성도이다.

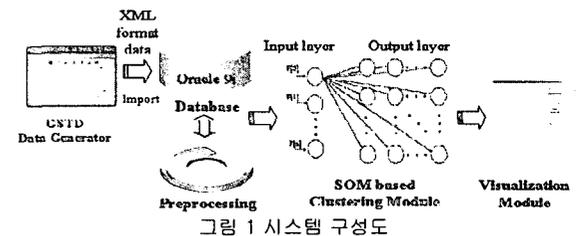


그림 1 시스템 구성도

시스템은 크게 데이터 생성기, 데이터베이스, 전처리 모듈, SOM 기반 클러스터링 모듈, 가시화 모듈로 구성되어 있다.

2.1 시공간 데이터 생성기

본 논문에서는 시공간 데이터에 필수적인 복잡한 사전 처리의 부담을 덜고 입력 데이터의 시공간 속성을 평가 목적에 맞게 적절히 설정하기 위하여, 실제 응용 시스템 데이터가 아닌 GSTD (Generate Spatio-Temporal Data) 시공간 데이터 생성기를 이용하였다. GSTD는 캐나다의 Alberta 대학에서 개발한 시공간 데이터베이스

너비, 정확도의 네 가지 항목을 기준으로 비교한다. 또한 더욱 정확한 마이닝 결과 분석을 위해 가시화 모듈을 개발하고 이를 이용해 통합 벤치마킹 시스템으로써 웹 상에서 사용자가 파라미터를 조절하여 다양한 형태와 움직임의 시공간 데이터를 XML 형태로 생성할 수 있도록 지원한다[5].

2.2 데이터 전처리 모듈

XML 형태로 생성된 원시데이터를 SOM에 적용하기 위해서는 다차원 벡터로 벡터화하는 작업이 필요하다. n 프레임 동안 움직인 이동 객체 i는 다음과 같이 n개의 흐름 벡터(Flow Vector)의 집합 Qi로 표현된다.

$$Q_i = \{f_1, f_2, f_3, \dots, f_n\}$$

또한 하나의 흐름벡터는 다음과 같이 4가지 요소로 이루어진다..

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

- x, y : 특정 시간에서의 x,y 좌표값
 - dx,dy : 객체가 특정시각에 각 축으로 움직인 상대순간속도
- 전처리 모듈에서는 각각의 연속되는 원시 데이터의 위치 좌표를 비교하여 상대 순간속도 dx와 dy를 계산하고 다음과 같은 다차원 흐름 벡터 형태로 변환하여 데이터베이스 내에 저장한다[표1].

[표1] 전처리 후의 시공간 데이터 테이블

필드명	데이터 타입	설명
ID	NUMBER(4,0)	이동 객체의 ID
TIME	NUMBER(18,6)	이동 객체가 움직인 시각(타임시스템)
X	NUMBER(18,2)	이동 객체의 위치 .x 좌표 값
Y	NUMBER(18,2)	이동 객체의 위치 .y 좌표 값
DX	NUMBER(18,2)	이동 객체가 x 축으로 움직이는 순간 속도
DY	NUMBER(18,2)	이동 객체가 y 축으로 움직이는 순간 속도
CLUSTER	NUMBER(4,0)	결과 클러스터 번호

2.3 SOM 기반 클러스터링 모듈

SOM은 Kohonen이 제안한 신경망 기반의 자기조직화 알고리즘으로 해부학적 이론에 근거하여 인간의 두뇌 구조를 모델링 한 방법이다. 즉, 인접한 출력노드들은 비슷한 기능을 수행할 것이라고 예측하여, 기존의 경쟁 학습을 개선하여 입력노드와 가장 가까운 출력노드들 뿐만 아니라 그 출력노드의 이웃노드들도 함께 학습시키는 알고리즘이다[6]. 구현된 SOM 기반 클러스터링 모듈은 다음과 같이 동작한다.

- 1) 데이터베이스에 접속한 후 클러스터 K의 수를 초기화한다
- 2) K개의 출력노드를 위상공간 내에 배치한다.
- 2) 학습률을 초기화한다. 학습률은 0과 1사이의 값을 가지며, 시간이 지남에 따라 감소한다.
- 3) 새로운 입력벡터를 입력노드에 제시한다.
- 4) 입력벡터와 모든 출력노드들과의 거리(Euclidian)를 계산하여 최소거리를 가지는 승자노드를 찾는다.
- 5) Gaussian 함수를 사용하여 승자노드와 이웃한 출력노드들의 가중치를 갱신한다.
- 6) 2번으로 가서 반복한다.
- 7) 훈련횟수만큼 네트워크를 훈련시킨 후 출력노드의 최종 가중치와 결과 클러스터 번호를 데이터베이스에 저장한다.

2.4 가시화 모듈

본 논문에서는 시공간 데이터의 특성과 클러스터링 결과를 더욱 쉽고 정확하게 나타내기 위하여 가시화 모듈을 구현하였다. 구현된 가시화 모듈의 특성은 다음과 같다.

- 2차원 좌표상에 입력 데이터 가시화(①)
- 입력에 대한 결과 클러스터 가시화(②)
- SOM 네트워크의 출력노드 표시(③)
- 균질도, 분리도, 반면영상너비 값 계산 (④)
- 객체의 정보조회(객체 ID, 소속 클러스터 번호, 좌표값) (⑤)
- 특정 클러스터의 정보조회(클러스터 번호, 중심좌표) (⑥)
- 특정 데이터가 속한 클러스터를 찾는 조회(⑦)
- 특정 데이터의 이동 궤적 가시화(시간에 따른 움직임)(⑧)

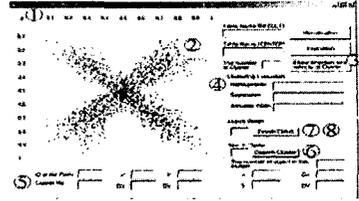


그림 2 가시화 모듈

3. 성능 평가

3.1 실험 데이터

성능 평가를 위해 사용한 실험 데이터는 크게 세 가지 종류로 나누어진다. 각각의 실험 데이터는 입력 데이터의 방향, 속도, 그리고 방향-속도의 복합 속성에 대해 알고리즘의 성능을 평가할 수 있도록 생성하였다. 첫 번째 그룹은 방향 속성에만 차이를 준 D1 ~ D6, 두 번째 그룹으로 속도에만 차이를 준 데이터 세트 D7 ~ D16, 마지막으로 임의의 복합 시공간 속성을 지닌 이동 객체 그룹인 D17의 세 가지 종류 데이터를 생성하였다. 각 데이터 세트는 2개의(D17의 경우 4개) 이동 객체 그룹들로 이루어져 있으며, 각 그룹 당 객체 수는 50개로 설정하였다 성능 평가에 사용된 테스트 데이터 세트의 특징을 정리하면 다음과 같다.

[표2] 실험 데이터의 시공간 속성

데이터 세트번호 (D1-D17)	방향 차이					속도 차이		방향 + 속도 (불규칙한 움직임)
	90도	75도	60도	45도	30도	1.0cm/s	0.2cm/s	
	활발함	동일	D1	D2	D3	D4	D5	D7-D16
	동일	비동일	D6					D17

3.2 성능 평가 기준

클러스터링의 성능 평가 즉 유효성 검사(validation)란 수치적, 객관적인 방식으로 클러스터 분석의 결과를 평가하는 작업이다. 일반적으로 클러스터링의 결과 평가하는 방법은 크게 외적 기준 분석과 내적 기준 분석의 두 가지로 나뉜다. 외적 기준 평가는 클러스터링의 결과를 대상 데이터를 분할하는 또 다른 최적 기준(Gold Standard)과 비교하는 작업이고 내적 기준 평가는 입력 데이터의 정보를 기반으로 입력 데이터와 클러스터링 결과 사이의 적합성을 평가하는 방법이다. 본 연구에서는 기존의 연구를 참고하여 다음과 같은 네 가지 항목을 평가 기준으로 선정하였다[7,8].

● 균질도(Homogeneity)

균질도는 클러스터의 중심점과 그에 속하는 점들간의 평균 거리이다. 균질도는 클러스터 내부의 분산을 나타내며 클수록 클러스터링이 잘되었다고 판단한다. 균질도 계산식은 다음과 같다.

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

D는 거리함수, P는 i번째 포인트, C(Pi)는 Pi가 속한 클러스터의 중심점, Npoint는 전체 포인트 수이다.

● 분리도(Separation)

분리도는 클러스터의 중심들간의 평균 거리로 클러스터 사이의 분산을 나타내며 수치가 작을수록 클러스터간의 분할이 명확하다고 판단한다. 분리도를 구하는 식은 다음과 같다.

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

Ci와 Cj는 i번째 클러스터와 j번째 클러스터의 중심점, Nci와 Ncj는 i번째 클러스터와 j번째 클러스터에 속한 포인트의 수이다

● 반면영상 너비(Silhouette Width)

반면영상 너비는 클러스터링 결과의 전체 질(quality)을 반영하

며 클러스터들이 얼마나 컴팩트하며 분리가 잘 되었는가를 나타낸다. 수치가 클수록 클러스터링이 잘 되었다고 판단한다.

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

$a(i)$ 는 i 번째 점과 이 점과 같은 클러스터에 속한 다른 점들과의 평균 거리이고 $b(i)$ 는 i 번째 점과 가장 근접해 있는 이웃 클러스터에 속한 점들의 평균 거리이다.

● 정확도(Accuracy)

정확도는 예상 클러스터와 결과 클러스터를 비교하기 위한 기준으로 각 데이터들이 얼마나 예상 클러스터에 정확하게 클러스터링 되었는지를 나타내어 결과의 정확성을 파악한다. 이 평가기준의 경우 예상 클러스터를 정확히 알고 있어야 평가가 가능하므로 실제로 데이터가 아닌 실험 데이터에만 적용 가능하다.

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

A 는 i 번째 클러스터에 정확하게 클러스터링 된 포인트 수이다. 3.3 성능 평가 결과

본 논문에서는 구현된 SOM 기반 클러스터링 알고리즘의 성능을 K-means, 응집 계층 알고리즘 방법인 Average Linkage, Ward의 세 가지 알고리즘들과 비교하였다. 구현된 SOM 이외의 알고리즘의 경우 S-PLUS 통계 패키지에서 제공하는 모듈을 이용하였다.

실험 데이터 그룹 D1~D6에 데이터의 방향 속성만을 달리한 그룹이다. 그림 3에서 볼 수 있듯이 방향 차이가 75도인 경우 (a), 모든 알고리즘이 두 객체 그룹을 명확하게 구분하여 클러스터링 하였지만, 차이가 30도로 줄어든 경우 (b) SOM 모듈만이 두 개의 그룹을 각각의 클러스터로 정확히 클러스터링 함을 알 수 있다.

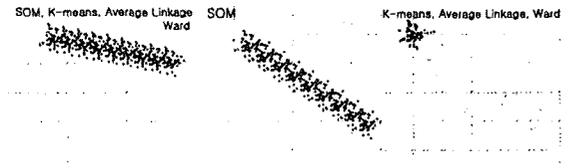


그림 3 데이터 그룹 D2, D5의 클러스터링 결과 비교
데이터 그룹간의 속도 속성에만 차이를 준 데이터 세트들 중 속도 차이가 0.6cm/s와 0.3cm/s인 두 가지 경우에 대해 클러스터링 결과를 가시화 시켜 보았다. 그림 4에서 볼 수 있듯이 이동 객체 그룹의 속도 차이가 적어질수록 구현된 SOM 모듈의 클러스터링 정확도가 다른 알고리즘들에 비해 우수함을 알 수 있다.

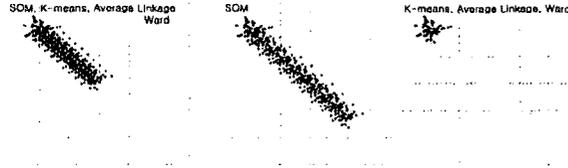


그림 5, 6은 각각 데이터세트 D2~D5 그리고 D11, D13, D14에 대한 각 성능 평가 수치를 나타낸 그래프이다. 그림에서 볼 수 있듯이 균질도, 분리도, 반면영상대비와 같이 데이터의 시공간 특성을 고려하지 않는 외적 기준 평가 항목인 경우 데이터 세트 별로 성능에 거의 차이가 없다. 하지만 두 그룹간의 시공간 속성에 차이가 적은 경우(D5, D14) 클러스터링 결과의 정확도에 있어서 SOM기반 모듈이 다른 알고리즘보다 우수함을 알 수 있다.

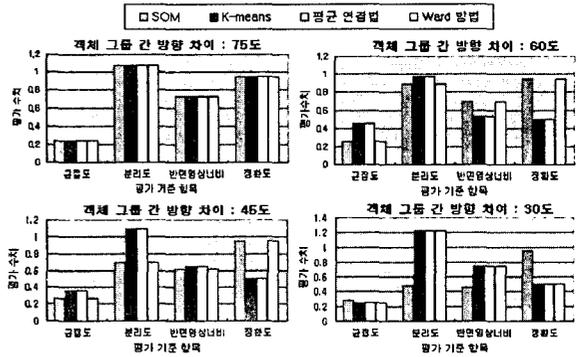


그림 5 클러스터링 알고리즘들의 성능비교(D2~D5)

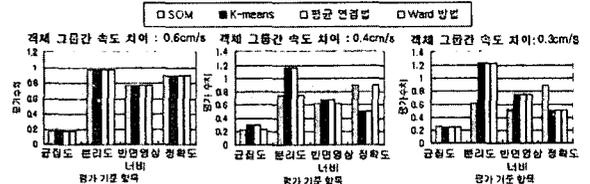


그림 6 클러스터링 알고리즘들의 성능비교(D11, D13, D14)

4. 결론

본 논문에서는 SOM 기반 시공간 데이터 마이닝 시스템을 구현하고 그 성능을 K-means, Average Linkage, Ward의 세 가지 알고리즘들과 비교 평가하였다. 균질도, 분리도, 반면영상대비와 같이 입력 데이터의 시공간 속성을 고려하지 않는 외적기준에서는 K-means와 구현된 SOM 기반 모듈이 비슷한 성능을 보였으나, 실제로 가시화를 통해 데이터 속성에 따른 클러스터링 정확도를 확인한 결과 구현된 모듈이 우수한 성능을 나타냄을 알 수 있었다. 이러한 결과를 기반으로 더욱 다양한 마이닝 알고리즘의 비교 연구나 시공간 속성에 따른 최적 알고리즘의 선택 기준을 제안하기 위한 지속적인 연구가 필요하다고 보여진다.

5. 참고 문헌

- [1] J.F. Roddick, and M. Spiliopoulou, " A bibliography of Temporal, Spatial and Spatio-temporal Data Mining Research", SIGKDD Newsletter, 1999.
- [2] J.F. Roddick, and B. G. Lees, " Paradigms for Spatial and Spatio-temporal Data Mining, Geographic Data Mining and Knowledge Discovery", Miller, H&J. Han. Taylor & Francis, London, 2001.
- [3] N. Johnson and D Hogg, " Learning the Distribution of Object Trajectories for Event Recognition", In Proc. of 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 Proc. of the 6th Int'l Symposium on Large Spatial Database(SSD), 1999
- [6] 김대주, 신경망 이론과 응용(1), 하이테크경, pp.169-189, 2001
- [7] Chen G, Jaradat SA, Banerjee N, Tanaka TS, " Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data", Statistica Sinica, pp.241-262, 2000
- [8] Yeung, Haynor, Ruzzo: Validating Clustering for Gene Expression Data. Technical Report UW-CSE-00-01-01, 2000