Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 1 pp. 717-721, 2022

https://doi.org/10.5762/KAIS.2022.23.1.717
ISSN 1975-4701 / eISSN 2288-4688

지지도 변화 감시를 이용한 효율적 부분 공간 클러스터링

박남훈
안양대학교 융합소프트웨어학과

Efficient Subspace Clustering with Monitoring Support Changes

Nam Hun Park
Department of Convergence Software, Anyang University

요 약 데이터 마이닝 분야에서 부분 공간 클러스터링은 항상 어려운 문제이다. 실제 응용 환경에서 데이터 스트림에 대한 데이터 마이닝 방법은 실시간으로 변화하는 동적 데이터 스트림 상의 지식을 효율적으로 추출해야 하며, 다차원 스트림의 모든 부분 공간 상의 정보 또한 효율적으로 탐색하여 추출해야 한다. 본 논문에서는 클러스터의 지지도 변화를 예측하는 격자기반 부분 공간 클러스터링을 제안한다. 각 격자셀에서 해당 범위 내의 데이터에 대한 정보가 유지되어, 고밀도 격자셀의 경우 범위를 분할하여 클러스터의 세부 경계를 정교하게 찾아간다. 데이터 스트림의 최근 변경 사항을 신속하게 식별하기 위해 격자셀의 지지도를 모니터링하고 변화를 예측하여 클러스터가 될 격자셀들을 사전 탐지한다. 지지도 예측을 통해 데이터 스트림의 변화를 실시간으로 탐지하여 반영한다. 제안된 방법은 일련의 실험을 통해 다양한 특성을 비교 분석하였으며, 지지도의 변화 속도를 예측하여 기존보다 빠르게 클러스터를 찾을 수 있다.

Abstract In the field of data mining, subspace clustering is always a difficult problem. In an actual application environment, a data mining method for a data stream should efficiently extract knowledge from a dynamic data stream that changes in real time, and should also efficiently search and extract information from all subspaces of a multi-dimensional stream. This paper proposes grid-based subspace clustering to predict change in cluster support. In each grid cell, information on data within the corresponding range is maintained, and in the case of a high-density grid cell, the range is divided up to precisely find a detailed boundary for the cluster. To quickly identify recent changes in the data stream, the proposed method monitors the support for grid cells and predicts changes in order to detect grid cells that will become clusters. Changes in the data stream are detected and reflected in real time through such support prediction. The proposed method compares and analyzes various characteristics through a series of experiments, and can find clusters faster by predicting the rate of change.

Keywords : Data Streams, Data Mining, Grid-Based Clustering, Real-Time Data Streams, Predicting Supports

*Corresponding Author : Nam Hun Park(Anyang Univ.)
email: nmhnpark@anyang.ac.kr
Received November 15, 2021 Revised December 13, 2021
Accepted January 7, 2022 Published January 31, 2022

1. 서론

실시간 스트림에 대한 데이터 마이닝 연구는 실제 인터넷 고객 스트림, 멀티미디어 데이터 및 실시간 센서 스트림과 같은 지속적으로 생성되는 방대한 데이터를 포함하는 최근 응용환경을 대상으로 수행되고 있다. 실제 환경에서의 데이터 스트림은 다차원 데이터로 구성되어 일부 차원 값이 누락되는 경우가 많다[1]. 데이터 스트림에서 실시간 변화를 탐지하려면 다차원 데이터의 모든 부분 공간에서 클러스터를 추적하고 변경을 예측하는 기능이 중요하다.

[2]에서는 k-median의 분할 기반 클러스터링 알고리즘을 사용하여 지속적으로 생성되는 데이터 스트림에서 먼저 각 부분 데이터 스트림의 클러스터를 찾는다. 새로 생성된 데이터 집합의 새 부분 데이터 스트림이 생성될 때마다 O(1)의 LSEARCH 루틴이 수행되어 부분 데이터 스트림의 클러스터 중심이 되는 k개의 데이터를 선택한다.

CluStream[3]은 변화하는 동적 데이터 스트림에서 생성된 데이터 집합의 클러스터를 찾기 위해 제안되었다. 기존의 k-means 방법을 이용하여 마이크로 클러스터라고 하는 초기 q개의 클러스터를 찾는다. 클러스터 특성 벡터[4]는 클러스터를 나타내는 데 사용된다. 새로운 데이터 집합이 생성되면 q 마이크로 클러스터를 지속적으로 갱신한다. 지정된 타임스탬프에서 모든 클러스터의 클러스터 특성 벡터는 스냅샷으로 저장되고, CluStream은 이러한 스냅샷의 마이크로 클러스터들에 대해 k-means 알고리즘을 다시 실행하여 매크로 클러스터라고 하는 k개의 최종 클러스터를 생성한다.

데이터 스트림에서 이러한 클러스터링 알고리즘들은 부분 공간 클러스터링을 대상으로 하지 않았다. CLIQUE[5]는 유한한 데이터 집합에서 상향식 방식으로 데이터 집합의 모든 부분 공간을 검색하여 부분 공간 클러스터를 찾는다. 동일한 목적으로 ENCLUS[6]은 데이터 항목들의 엔트로피를 측정하여 부분 공간 클러스터를 찾는 반면 FIRES[7]는 각 차원의 1차원 클러스터를 기반으로 다차원의 클러스터를 추측하는 근사화 방법을 사용한다. 이러한 기존의 부분 공간 클러스터링 방법들은 전체 데이터 집합을 여러 번 반복하여 검사해야 하는 것이다. 이들은 지속적으로 무한히 생성되는 온라인 데이터 스트림에 적용할 수 없다.

본 논문에서는 데이터 스트림에 대한 부분 공간 클러스터링을 위해 격자 기반 인덱스 구조를 채택하였다. $N_1 \rightarrow N_2 \rightarrow \cdots \rightarrow N_n$ 의 미리 정의된 차원 순서로 각 차원에 대한 독립적인 격자셀 리스트의 트리를 생성하고, 각 레벨에서 1차원 클러스터를 탐색한다. 차원 $N_k(1 \le k \le n)$에 대한 격자셀 리스트의 한 격자셀이 데이터 빈도가 높아지면 해당 격자셀의 자식으로 새로운 격자리스트 집합이 생성된다.

차원 N_k 내의 모든 다차원 부분 공간을 탐색하기 위해 차원 순서에서 N_k 뒤에 있는 차원에 대해서 새 격자리스트가 생성된다. 결과적으로 $(d-k)$개의 격자셀 목록이 격자셀의 자식으로 생성된다. k레벨에 있는 노드의 격자셀은 루트에서 자신을 포함하는 노드까지의 경로에서 각 차원의 격자셀 공간이 교차하여 형성된 직사각형 부분 공간에 해당한다. 또한 데이터 스트림의 변화를 실시간으로 반영하기 위해 밀도 변화 속도를 측정하여 그리드 셀의 지원을 모니터링하고 예측한다.

이 논문은 다음과 같이 구성되어 있다. 2장에서는 부분 공간의 데이터 항목 통계를 유지하기 위한 격자셀 구조를 제시한다. 3장에서는 격자셀 모니터링 및 지지도 예측 방법을 제시한다. 4장에서는 제안한 방법의 성능을 평가하기 위해 다양한 실험 결과를 비교 분석한다. 마지막으로 5장에서는 결론을 제시한다.

2. 데이터 스트림 모니터링

n차원 데이터 공간 $\boldsymbol{N}=N_1 \times \ldots \times N_n$ 의 데이터 스트림이 주어지면 k 차원 격자셀$(1 \le k \le n)$의 영역은 각각이 별개의 차원 $N_1, N_2, \cdots, N_k$로 정의된 k차원 그리드 셀의 직사각형 공간은 각 차원에서의 범위 I_k의 교집합으로 $\boldsymbol{RS}=I_1 \times I_2 \times \cdots \times I_k$ 이 된다. 이러한 격자 셀의 직사각형 공간에서 데이터 항목들의 분포 정보를 효율적으로 관리하기 위해 모니터링 트리가 사용된다.

정의 1. 부분 공간 격자셀 모니터링 트리

차원순서 $N_1 \rightarrow N_2 \rightarrow \cdots \rightarrow N_n$에 대해서, n차원 데이터 공간 $\boldsymbol{N}=N_1 \times \cdots \times N_n$상의 데이터 스트림 $\boldsymbol{D}$에서, m개의 자식노드를 가지는 부분 공간 격자셀은 다음과 같이 정의한다.

1) 트리의 각 노드는 항목 $E\langle min, max, G[1,\cdots,m], next_ptr\rangle$과 자식 배열 $U[1,\cdots m][1,\cdots,n-1]$, 격자셀의 차원을 나타내는 T_{dim} 으로 구성된다. 최대 m개의 자식 노드를 가진다.

2) j번째 레벨의 노드 p가 $(j-1)$번째 레벨의 노드 q의 격자셀 $q.G[i]$ $(1 \le i \le m)$의 자식이라면, 노드 p는 새로운 격자셀 리스트 L의 첫 번째 항목이 되고 노드 p는 목록 L의 헤드로 호출한다. 격자셀 $q.G[i]$는 격자셀 리스트 L에 있는 모든 격자셀의 부모로 호출된다.

3) D_g^t는 격자셀 $G[j]$의 범위에 있는 데이터 항목들의 분포 정도를 나타내도록 한다. 즉, $D_g^t = \{ e \mid e \in D^t$ 및 $e \in RS(G[j]) \}$이 된다. 공간 $RS(G[j])$에서 데이터 항목의 분포는 격자셀 $G[j](l,c,\mu,\sigma)$에 의해 모니터링 된다. 즉, D_g^t의 현재 데이터 요소 수는 $G[j].c^t$에 의해, D_g^t의 데이터 요소의 평균 및 표준편차는 각각 $G[j].\mu$ 및 $G[j].\sigma$에 의해 관리된다.

$N_1 \rightarrow \cdots \rightarrow N_n$ 차원 순서와 미리 정의된 노드 내 격자셀 수 h 에 대해, 격자셀 리스트 $L_1, \cdots, L_n$이 생성되어 각 차원 공간의 격자셀을 유지한다. 첫 번째 레벨의 격자셀 리스트는 h개의 격자셀을 유지하고 단일 노드가 생성되어 격자셀 리스트를 구성한다. 새 개체 o^t가 생성되면 이전 o^v 때의 분포정보는 다음과 같이 갱신된다.

$$g.\mu^t = (g.\mu^v \times g.c^v + e^t) / g.c^t$$

$$g.\sigma^t = \sqrt{\frac{g.c^v \times (g.\sigma^v)^2}{g.c^t} + \frac{(g.\mu^v)^2 + (e^t)^2}{g.c^t} - (g.\mu^t)^2}$$

데이터 스트림이 계속 진행됨에 따라 각 격자셀 리스트 L_v $(1 \le v \le n)$ 의 밀도가 높은 격자셀은 h개의 더 정교한 범위의 격자셀로 반복해서 분할된다. 밀도 높은 노드는 헤드 노드가 되어, 차원 N_l $(v+1 \le l \le n)$에 대해 영역을 분할하여 생성된 격자셀 리스트는 직사각형 부분 공간 공간 $G[1].l \times N_l$ 에서 데이터 항목의 분포 정보를 모니터링한다.

두 번째 레벨의 차원 N_l $(v+1 \le l \le n)$에 대한 격자셀 리스트의 격자셀이 밀도가 높아지면 더 정교한 격자셀로 다시 분할한다. 결과적으로 격자셀 리스트 내 격자셀 수가 증가하고, 최소크기의 격자셀이 되면 다음 차원에 대한 $(n-l)$개의 새로운 격자셀 리스트가 밀도높은 격자셀의 자식으로 다음 레벨에서 생성된다. 이런 식으로 모니터링 트리는 n번째 레벨까지 성장한다. 한편, 모니터링 트리에서 노드 n의 격자셀 g의 밀도가 낮아지면 상위 범위의 격자셀로 병합되고, 이 때. 하위 노드들은 밀도가 낮은 격자셀의 자식으로 상위 노드로 병합된다.

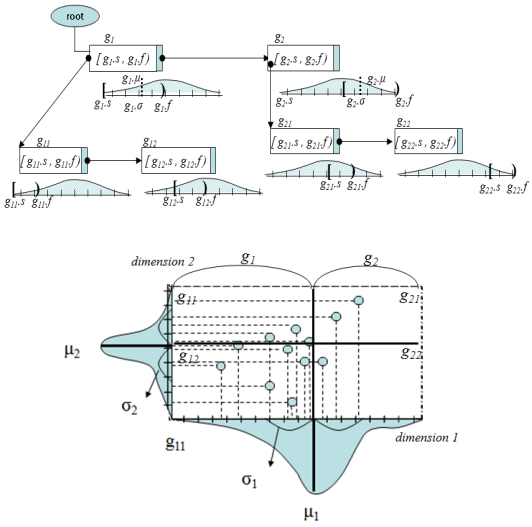


Fig. 1. An example of multi-dimensional grid-cells

3. 격자셀 지지도 변화

데이터 스트림이 진행되면서 각 격자셀의 지지도는 계속해서 변화한다. 격자셀의 객체 빈도수 변화를 분석하면 앞으로 지지도 변화를 예측하여 클러스터를 찾는 데에 반영할 수 있다. 지지도를 예측하고 패턴을 찾기 위해서는 각 격자셀에서 더 많은 정보를 유지해야 한다.

격자셀의 경우 지지도의 속도는 지지도 변화의 차이값으로 정의한다. 격자셀의 V_{count}는 최근 지지도의 속도를 의미한다. 데이터 스트림 D^t에서 이후 격자셀의 빈도수 변화 $V_{count}^{(t+1)}$를 구할 수 있다.

$V_{count}^{t+1} = count^t - count^{t-1}$

속도 V_{count}^{t+1}에서 $(t+1)$번째 시간에서 객체의 빈도 P_{count}^{t+1}는 다음과 같이 구할 수 있다.

$P_{count}^{t+1} = count^t + V_{count}^{t+1}$

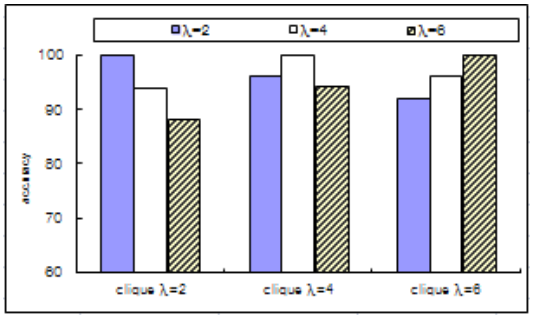
$(t+1)$번째 시점의 지지도는 P_{count}^{t+1}로부터 다음과 같이 예측할 수 있다.

$S^{t+1} = (count^t + P_{count}^{t+1}) / (|D^t| + count^{t+1})$

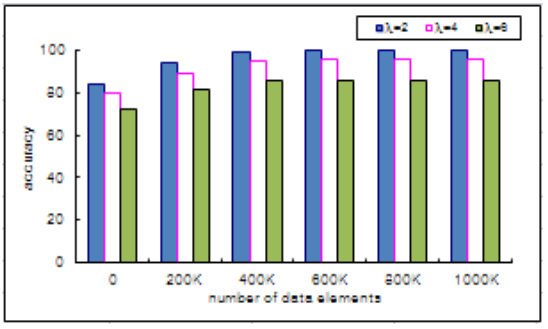
현재 지지도 S^t가 있는 격자셀이 v시간 후에 클러스터가 된다면, $S^{t+v} = (count^t + P_{count}^{t+v}) / (|D^t| + count^{t+v}) \ge S_{min}$ 을 만족해야 한다.

4. 실험

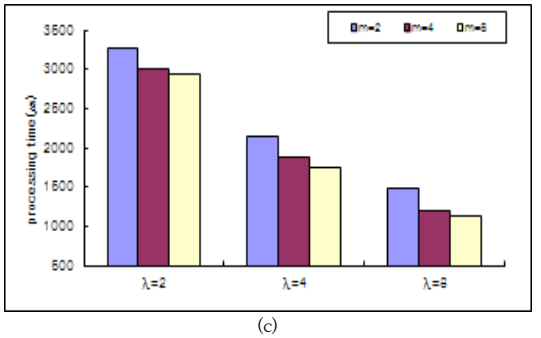
제안하는 방법의 성능을 분석하기 위해 100만 개의 20차원의 데이터 항목으로 구성된 데이터 집합을 ENCLUS의 데이터 생성기를 사용하여 준비하였다. 대부분의 데이터 항목은 각 차원별로 무작위의 범위의 10개의 영역을 중심으로 분포하였다. 실험 조건은 각 실험에서 다르게 지정하지 않는 한 S_{min}=0.01, 격자셀 최소크기 λ=2, m=4이다. 모니터링 트리의 각 레벨의 차원은 데이터 분포에 의해 동적으로 결정되게 하였다. 모든 실험에서 데이터 항목은 온라인 데이터 스트림의 환경을 시뮬레이션하기 위해 순서대로 하나씩 처리하였다.



(a)



(b)



(c)

Fig. 2. Performance Evaluation
(a) Accuracy comparison (b) Accuracy variation
(c) Memory usage

제안하는 방법의 정확도는 그림 2와 같다. 유한한 데이터 집합에 대한 부분 공간 클러스터링 알고리즘 CLIQUE의 수행결과와 비교하여 정확도를 측정한다. 전체 데이터 항목들 중 CLIQUE에 의해 클러스터로 판별된 데이터 항목 대비 제안된 방법에서도 CLIQUE와 동일한 클러스터로 판별된 데이터 항목으로 정확도를 정의한다.

Fig. 2(a)는 4개의 서로 다른 λ 값에 대해 정확도 변화를 보여준다. 제안하는 방법의 λ 값이 CLIQUE의 값과 같을 때 두 방법의 정확도는 같으며, 이는 데이터 스트림에서도 기존 CLIQUE와 동일하게 데이터 항목의 클러스터를 찾을 수 있음을 보여준다. Fig. 2(b)는 새로운 데이터 요소가 생성될 때 정확도의 변화를 보여준다. 제안한 방법은 부분 공간 클러스터를 찾는 과정에서 단위 격자셀을 찾을 때까지 많은 분할 과정이 발생하므로 초기의 정확도는 상대적으로 낮다. 그러나 연속적으로 정교한 격자셀을 생성하면서 정확도가 점점 높아지는 것을 보여준다. Fig. 2-(c)는 제안된 방법의 처리 시간을 보여준다. 분할하는 격자셀 수가 너무 작으면(m=2) 각 격자 리스트에서 반복하는 분할 수가 증가하여 상대적으로 처리 시간이 늘어난다. m이 크고, 지지도 변화를 반영했을 때 수행속도가 빠른 것을 확인할 수 있다.

5. 결론

데이터 스트림의 차원 수가 높을수록 부분 공간 클러스터링은 차원의 부분집합에서 관심있는 클러스터를 분석하는 데 유용하다. 그러나 기존의 부분 공간 클러스터링 방법은 가능한 모든 부분 공간 별로 클러스터를 후보를 생성하고 각 후보에 대해 데이터 집합 내 항목들을 반복적으로 검사해야 하기 때문에. 온라인 데이터 스트림에서는 사용할 수 없다. 본 논문에서는 데이터 스트림에 대한 부분 공간 클러스터링 방법을 제안하였다. 격자기반 구조를 유지함으로써 데이터 스트림의 현재 분포 정보를 주의 깊게 모니터링 한다. 각 격자셀의 지지도는 변화 속도를 예측하여 실시간 데이터 마이닝에서 차후 클러스터 변화를 예측하고 좀 더 빠르게 클러스터를 찾는 데에 활용할 수 있다.

References

[1] Ming Hua, Jian Pei, Xuemin Lin, "Ranking queries on

uncertain data", *The International Journal on Very Large Data Bases,* Vol. 20, No. 1, pp. 129-153, Feb. 2011.
DOI: http://dx.doi.org/10.1007/s00778-010-0196-4

[2] Mohammed Oualid Attaoui, Hanene Azzag, Mustapha Lebbah, Nabil Keskes, "Subspace data stream clustering with global and local weighting models", *Neural Computing and Applications,* Vol. 33, pp. 3691-3712, Aug. 2020.
DOI: https://doi.org/10.1007/s00521-020-05184-z

[3] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Yu, "A Framework for Clustering Evolving Data Streams", *In Proc. VLDB 29th*, Berlin, Sep. 2003.DOI: http://dx.doi.org/10.1016/B978-012722442-8/50016-1

[4] Tang MingJing, Li Tong, Zhu Rui, Ma ZiFei, "A Cluster Analysis Method of Software Development Activities Based on Event Log", *Recent Advances in Computer Science and Communications*, Vol. 14, Number 6, pp. 1843-1851, Aug. 2021.
DOI: https://doi.org/10.2174/2666255813666191204144931

[5] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", *In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 94-105. ACM Press, June 1998.
DOI: http://dx.doi.org/10.1145/276304.276314

[6] Chun-Hung Cheng, Ada Waichee Fu, Yi Zhang, " Entropy-based subspace clustering for mining numerical data", *In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 84-93, ACM Press, Aug. 1999.
DOI: http://dx.doi.org/10.1145/312129.312199

[7] Hans-Peter Kriegel, Peer Kroger, Matthias Renz, Sebastian Wurst, "A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data", *In Proceedings of the Fifth IEEE International Conference on Data Mining*, pp. 250-257, Nov. 2005.
DOI: http://dx.doi.org/10.1109/ICDM.2005.5

[8] M. Ester, H. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases", *Proceedings of the Second International Conference on Knowledge Discovery and Data MiningAugust*, Aug. 1996.
DOI: http://dx.doi.org/10.1.1.121.9220

박 남 훈(Nam Hun Park) [정회원]

- 2000년 2월 : 연세대학교 기계전자공학부 컴퓨터과학 (공학사)
- 2002년 2월 : 연세대학교 컴퓨터과학 (공학석사)
- 2007년 8월 : 연세대학교 컴퓨터과학 (공학박사)
- 2008년 9월 ~ 2010년 2월 : Worcester Polytec. Inst. Research Associate
- 2010년 3월 ~ 현재 : 안양대학교 융합소프트웨어학과 교수

〈관심분야〉
데이터 마이닝, 데이터 스트림