교육데이터 정제를 위해 다양한 밀도분포를 고려한 개선된 DBSCAN 알고리즘

김정훈† · 나스리디노프 아지즈†
† 충북대학교 컴퓨터과학과

An Enhanced DBSCAN Algorithm to Consider Various Density Distributions for Educational Data

Jeong-Hun Kim† · Aziz Nasridinov†
† Dept. of Computer Science, Chungbuk National University

요 약

교육데이터마이닝은 다양한 교육 환경에서 생성되는 막대한 양의 데이터를 활용하여 학습자들의 학습 유형, 학습 진도를 분석, 예측하고 교육 성취를 효과적으로 향상시키는 것을 목적으로 한다. 효과적인 교육데이터마이닝 결과를 얻기 위해서는 교육데이터에 대한 정제 과정이 필요하며 DBSCAN 클러스터링을 통해 교육데이터에 포함된 노이즈 데이터를 제거하고 생성된 각 클러스터에서 동일한 비율로 데이터를 추출함으로써 편향되지 않은 표본 데이터를 생성할 수 있다. 하지만 DBSCAN은 두 개의 전역 매개변수에 의해 다양한 밀도분포를 가지는 클러스터를 생성할 수 없다는 문제점이 있으며 이는 교육 데이터를 정제함에 있어 치명적인 문제점이 될 수 있다. 본 논문에서는 DBSCAN의 문제점을 개선하고 클러스터링 정확도를 향상시키기 위해 고정된 매개변수를 사용하지 않고 각 밀도분포에 대해 최적의 입력 매개변수를 결정함으로써 다양한 밀도분포를 가지는 클러스터들을 효과적으로 생성하는 *C-DBSCAN*을 제안한다.

1. 서 론

교육데이터마이닝은 교육 환경에서 생성되는 데이터를 활용하여 의미 있는 정보를 추출해내는 분석 방법이다[1]. 주로 LMS(Learning Management System)에서 제공하는 학습자의 접속, 토론방 사용, 시험 결과 등의 활동 정보들을 기록한 로그 데이터 형태의 데이터를 분석함으로써 학습자들의 학습 유형, 학습 진도를 예측하거나 학업 성취 유무를 판단한다. 이와 같은 분석결과는 학습자들의 교육 성취도를 효과적으로 향상시키는 것을 목적으로 사용할 수 있다. 최근 교육데이터는 오프라인뿐만 아니라 다양한 온라인 교육 환경에 의해 교육데이터가 폭발적으로 늘어나고 있는 추세이다. 이는 분석이 필요한 데이터뿐만 아니라 우연에 의한 학습 성공, 편향적인 교육 성취도와 같은 노이즈 데이터 또한 증가함을 의미한다. 따라서 정확한 교육데이터마이닝을 위해서는 막대한 양의 교육데이터를 정제하여 분석에 사용되지 않는 불필요한 데이터(노이즈 데이터)를 제거하거나 전체 데이터를 대표할 수 있는 표본 데이터를 추출해야할 필요가 있다[2].

노이즈 데이터를 제거하고 편향적이지 않은 표본데이터를 추출하기 위한 방법으로 대표적인 데이터마이닝 기법 중 하나인 클러스터링을 사용할 수 있다. 클러스터링은 비지도 학습 기법 중 하나로 서로 유사한 데이터들을 동일한 그룹으로 분류한다[3]. 클러스터링에서 사용하는 유사성 척도는 거리 기반 척도와 밀도 기반 척도가 있으며 밀도 기반 척도를 사용하는 대표적인 클러스터링 알고리즘인 DBSCAN은 유사한 데이터들의 그룹과 노이즈 데이터를 동시에 찾는다[4]. 따라서 생성된 각 클러스터를 구성하는 데이터를 동일한 비율로 추출하여 편향적이지 않은 표본데이터를 확보할 수 있으며 노이즈 데이터를 제거할 수 있다. 하지만 DBSCAN은 각 데이터의 밀도를 정의하기 위해 두 가지 입력 매개변수가 필요하며 오직 하나의 밀도분포에 대한 클러스터만을 찾는 문제가 있다[5]. 이는 서로 다른 밀도분포를 가지는 클러스터가 존재할 경우 클러스터를 찾을 수 없거나 하나의 클러스터로 병합되는 심각한 문제를 유발하며[6], 정제된 교육데이터의 문제로 직결된다.

본 논문에서는 기존 DBSCAN 알고리즘을 구성주의적 교육 패러다임에 근거한 교육데이터의 특성을 클러스터링 제약조건으로 활용하여 문제점을 개선하고 클러스터링 성능을 향상시키는 *C-DBSCAN(Constraints based DBSCAN)*을 제안한다. *C-DBSCAN*은 각 클러스터의 밀도분포에 적합한 매개변수를 사용하여 다양한 밀도분포의 클러스터를 찾는다.

본 논문의 구성은 다음과 같다. 2장에서는 대표적인 밀도 기반 클러스터링 알고리즘인 DBSCAN에 대해

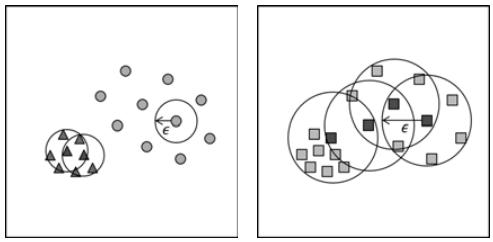
설명한다. 3장에서는 다양한 밀도분포를 가지는 클러스터를 찾기 위해 제안하는 방법인 *C-DBSCAN*을 설명한다. 마지막 장에서 결론을 내리고 향후 연구 방향을 제시한다.

2. 관련연구

본 장에서는 밀도기반 클러스터링 알고리즘인 DBSCAN의 기본적인 개념과 장단점을 간단히 설명한다. 특히, DBSCAN의 핵심개념을 설명한 후 두 개의 전역 파라미터에 의해 DBSCAN이 가지는 문제점에 대해 논의한다.

2.1 DBSCAN 알고리즘

DBSCAN(Density-based spatial clustering of application with noise)은 밀도기반 클러스터링 알고리즘으로 두 데이터 포인트에 대한 밀도 연결 관계를 제시하여 노이즈 데이터가 포함된 데이터집합 상에서 연결된 모든 데이터들의 집합을 클러스터로 정의한다. 이를 위해 DBSCAN은 두 개의 전역 매개변수인 ϵ과 *MinPts*를 가진다. ϵ은 이웃 데이터를 찾기 위한 반경이며 *MinPts*는 클러스터를 생성하기 위해 이웃 반경에 포함해야하는 최소 데이터의 수이다. DBSCAN은 데이터집합의 각 데이터 포인트에 대한 ϵ 내의 이웃 데이터 포인트의 수를 확인한다. 만약 ϵ 내의 이웃 데이터 포인트의 수가 *MinPts* 이상이면 새로운 클러스터를 생성하고 모든 이웃 데이터 포인트들을 같은 클러스터에 할당한다. 이후 각 이웃 데이터 포인트들의 ϵ 내의 이웃 데이터 포인트의 수를 확인하는 과정을 반복하여 클러스터의 크기를 점진적으로 증가시킨다. 클러스터에 포함된 각 데이터 포인트 중 더 이상 *MinPts* 이상의 이웃 데이터 포인트의 수를 가지는 데이터 포인트가 없을 경우 최종 클러스터로 확정된다. DBSCAN은 거리 기반 유사성 척도를 사용하지 않기 때문에 구 형태의 클러스터뿐만 아니라 다양한 형태의 클러스터를 찾을 수 있고 클러스터의 수를 입력할 필요가 없다는 장점을 가진다. 하지만 대부분의 데이터 집합에 포함된 클러스터들은 두 개의 전역 매개변수만으로 모두 찾을 수 없다. 다음 [그림 1]은 서로 다른 밀도분포를 가지는 두 개의 클러스터에 대해서 DBSCAN이 가지는 두 개의 전역 매개변수에 의한 클러스터링 문제점을 나타낸다. 먼저, 고밀도 클러스터를 기준으로 ϵ을 선택할 경우 저밀도 클러스터를 찾지 못하는 문제가 발생한다. 이 경우 저밀도의 클러스터는 모두 노이즈 데이터로 분류된다. 반대로 저밀도 클러스터를 기준으로 ϵ을 선택할 경우 고밀도 클러스터와 저밀도 클러스터가 하나의 클러스터로 병합되는 문제가 발생한다. 이와 같은 DBSCAN의 문제점은 고정된 전역 매개변수가 단 하나의 밀도분포만을 고려하여 클러스터를 찾기 때문이다. 따라서 DBSCAN은 다양한 밀도분포를 가지는 클러스터들을 생성하지 못하며 클러스터링 정확도에 한계점을 가진다.



[그림 1] DBSCAN의 문제점(*MinPts=4*)

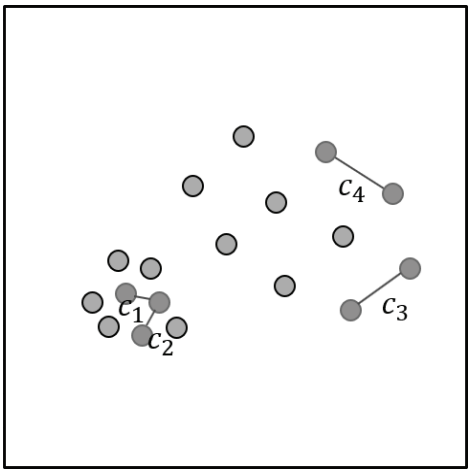
3. C-DBSCAN

본 장에서는 2.1절에서 언급한 DBSCAN의 문제점을 개선하여 클러스터링 정확도를 향상시키기 위해 *C-DBSCAN(Constraints based DBSCAN)*이라는 접근방법을 제안한다. *C-DBSCAN*는 기존의 DBSCAN에서 클러스터를 생성하는 과정에 제약조건을 추가하여 다양한 밀도분포를 가지는 클러스터를 생성한다. 3.1 절에서는 각 밀도분포에 대한 클러스터링 매개변수를 결정하는 방법을 설명한다. 3.2 절에서는 결정된 매개변수를 사용하여 다양한 밀도분포를 가지는 클러스터를 생성하는 *C-DBSCAN*을 설명한다.

3.1 클러스터링 매개변수 결정

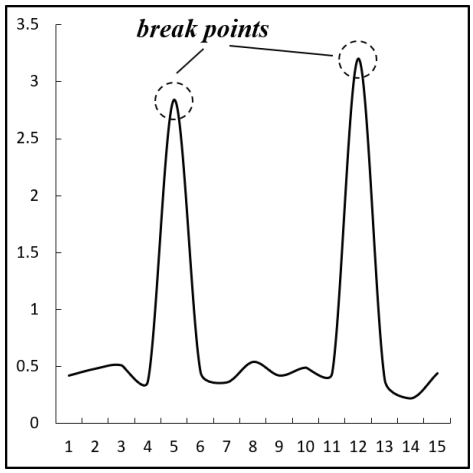
*C-DBSCAN*은 다양한 밀도분포를 가지는 클러스터를 생성하기 위해 고정된 전역 매개변수가 아닌 각 밀도분포에 대한 적응형 매개변수를 사용한다. 적응형 매개변수를 결정하기 위해 구성주의적 교육 패러다임에 근거한 교육데이터의 특성을 활용한 클러스터링 제약조건을 정의한다. 구성주의적 교육 패러다임은 특정 개념을 습득하기 위해서는 학습자 스스로가 경험으로부터 지식과 의미를 구성해내는 것으로 시행착오를 거쳐가며 요구되는 사전지식을 새로 습득하거나 수정해 나가는 것을 뜻한다. *C-DBSCAN*은 이러한 시행착오의 과정이 서로 유사한 학습자들은 학습능력과 교육성취도 또한 유사하다는 사실을 기반으로 하여 해당되는 학습자들이 항상 같은 클러스터에 속하도록 클러스터링 제약조건을 정의한다. 클러스터링 제약조건은 각 학습자를 데이터 포인트로 하여 벡터 공간 상에서 두 개의 데이터 포인트를 연결하는 형태로 정의하며 클러스터링 제약조건에 해당되는 데이터 포인트는 동일하거나 유사한 밀도분포 영역에 위치한다. 여기서 클러스터링 제약조건은 사용자에 의해 미리 정의하며 본 논문에서는 정의된 제약조건을 통해 클러스터링을 위한 적응형 매개변수를 결정하는 방법만을 다룬다. 다음 [그림 2]는 임의의 데이터집합 *D*와 서로 유사한 학

습자들을 확인하여 미리 정의한 클러스터링 제약조건을 나타낸다.



[그림 2] 클러스터링 제약조건

[그림 2]의 c_1, c_2, c_3, c_4는 데이터집합 D의 클러스터링 제약조건이며, 각 제약조건에 의해 연결된 데이터는 같은 클러스터에 포함된다. 각 제약조건에 대해서 연결된 데이터들을 같은 클러스터에 포함시킬 수 있는 ϵ을 결정하기 위해서 k-nearest neighbour (kNN) 알고리즘을 사용한다. 임의의 데이터에 대한 kNN을 통해 k번째 이웃과의 거리(k-dist)를 계산하면 해당 데이터가 포함된 영역의 밀도 분포를 대략적으로 확인할 수 있다. 또한 각 데이터의 k-dist는 해당 데이터가 클러스터를 생성할 수 있는 최소 반경을 나타낸다. 이러한 k-dist의 특성을 통해 각 제약조건의 ϵ을 다음과 같은 순서를 통해 결정한다. 먼저, 제약조건에 해당되는 두 개의 데이터 p와 q 중 p를 선택하여 k-dist를 계산한다. 여기서 k는 DBSCAN의 입력 매개변수인 *MinPts*와 동일하다. 이후 계산된 k-dist와 *MinPts*를 통해 제약조건의 또 다른 데이터 q와 밀도 연결 관계인지 확인한다. 반대로 q의 k-dist를 계산한 뒤 동일한 과정을 통해 데이터 p가 밀도 연결 관계인지 확인한다. 만약 p의 k-dist와 q의 k-dist 모두 밀도 연결 관계를 만족하면 두 데이터의 k-dist의 평균을 ϵ으로 결정한다. p의 k-dist만 만족할 경우 p의 k-dist를 ϵ으로 결정하며 그 반대의 경우 q의 k-dist를 ϵ으로 결정한다. 만약 p와 q 모두 밀도 연결 관계를 만족하지 못할 경우 p와 q의 이웃 데이터 중 임의의 데이터 o를 선택하여 k-dist를 계산하고 두 데이터 p와 q 모두 밀도 연결 관계인지 확인한다. 모든 이웃 데이터들에 대해 동일한 과정을 반복한 후 두 데이터 p와 q 모두 밀도 연결 관계를 만족하는 k-dist의 평균을 계산하여 ϵ으로 결정한다. 미리 정의된 모든 제약조건들에 대해 ϵ을 결정하면 각 제약조건들에 대한 데이터집합의 밀도 분포를 확인할 수 있다. 고밀도 영역의 제약조건은 작은 ϵ을 가지며 반대로 저밀도 영역의 제약조건은 큰 ϵ을 가진다. 결정된 ϵ의 집합을 통해서 데이터 집합의 전체 밀도분포를 확인할 수 있지만 다양한 밀도분포를 가지는 클러스터를 찾기 위해서는 결정된 ϵ 중 각 클러스터의 밀도분포를 효과적으로 반영하는 최적의 ϵ을 찾아야한다. 여기서 서로 같은 밀도분포에 있는 제약조건은 유사한 ϵ을 가진다. 따라서 ϵ의 변화량을 통해 각 클러스터의 밀도분포의 변화를 확인할 수 있다. 다음 [그림 3]은 ϵ의 변화량 그래프를 나타낸다.



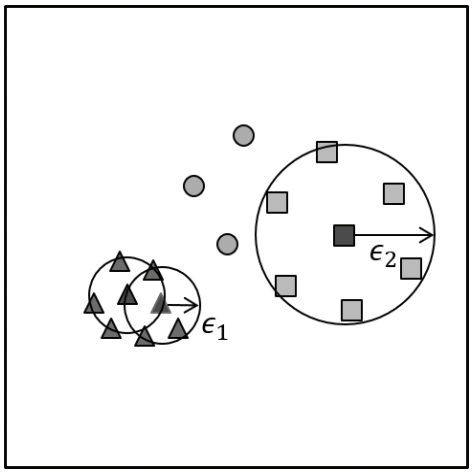
[그림 3] ϵ 변화량 그래프

[그림 3]과 같이 ϵ의 집합을 오름차순으로 정렬한 후 변화량을 그래프로 나타냄으로써 데이터 집합에 포함된 각 클러스터의 밀도분포를 확인할 수 있다. ϵ의 변화량이 급격하게 증가하는 부분을 *break point*라 하며 각 제약조건이 서로 다른 밀도분포에 위치하고 있음을 알 수 있다. 따라서 ϵ의 변화량이 급격하게 증가하는 부분을 분기점으로 하여 각 밀도분포에 대한 최적의 ϵ을 선택한다. [그림 2]의 경우 두 개의 밀도분포에 대한 ϵ을 선택하게 되는데 하나의 *break point*에 의해 두 개의 ϵ 부분집합으로 분할된다. 이후 각 ϵ 부분집합에서 가장 큰 ϵ을 해당 밀도분포의 최적 매개변수로 결정한다.

3.2 데이터 클러스터링

본 장에서는 데이터 집합에 포함된 다양한 밀도분포를 가지는 클러스터를 찾기 위해 각 밀도분포에 대한 최적의 ϵ을 사용한 *C-DBSCAN*에 대해 설명한다. 우선, 각 밀도분포 별로 결정된 최적의 ϵ 집합을 오름차순으로 정렬한 뒤 가장 작은 ϵ부터 선택하여 순차적으로 DBSCAN 클러스터링을 진행한다. 여기서 *MinPts*는 앞선 kNN의 k와 동일한 값을 사용한다. 가장 작은 ϵ은 데이터 집합에서 가장 고밀도의 클러스터를 생성

하며 클러스터에 포함되지 못한 데이터들은 다음 순서의 ϵ에 대한 입력 데이터 집합으로 사용된다. 이러한 과정을 모든 ϵ에 대해 반복한 후에도 클러스터에 포함되지 않은 데이터들은 노이즈 데이터로 결정하고 더 이상 클러스터에 속한 데이터가 남아있지 않거나 모든 ϵ에 대한 클러스터링을 진행하면 종료한다. 다음 [그림 4]는 *C-DBSCAN*의 클러스터링 진행과정을 나타낸다.



[그림 4] *C-DBSCAN* 클러스터링 진행과정

[그림 4]의 ϵ_1은 높은 밀도분포를 가지는 클러스터에 대한 최적의 ϵ이며 가장 먼저 클러스터링을 진행하여 생성된다(삼각형 모양의 데이터). 이후 [그림 4]와 같이 데이터 집합에서 클러스터에 포함되지 않은 데이터들에 대해 ϵ_2를 사용하여 클러스터링을 진행한다. 따라서 서로 다른 밀도분포를 가지는 클러스터에 대해 다른 ϵ을 사용함으로써 클러스터링 결과의 정확도를 향상시킬 수 있다.

4. 결론

본 연구는 효과적인 교육데이터마이닝을 위해 막대한 양의 교육데이터에 포함된 노이즈 데이터를 제거하고 편향되지 않은 표본 데이터를 추출하기 위한 기존의 DBSCAN 클러스터링 알고리즘이 가지는 문제점을 개선하여 클러스터링 정확도를 향상시킬 수 있는 *C-DBSCAN*을 제안하였다. *C-DBSCAN*은 다양한 밀도분포를 가지는 클러스터들에 대해 각 밀도분포 별로 다른 ϵ을 사용하여 클러스터링을 진행함으로써 가장 높은 밀도분포를 가지는 클러스터부터 순차적으로 클러스터를 생성한다.

본 연구의 향후 계획은 *C-DBSCAN*을 수행하기 위해 필요한 클러스터링 제약조건을 사용자가 미리 정의하지 않고 알고리즘 자체적으로 정의할 수 있도록 데이터 집합을 통계적으로 분석하여 특징을 추출하는 데이터 전처리 과정에 대한 연구를 계속할 것이다.

Acknowledgement

이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A3B03035729)

참 고 문 헌

[1] 이설화・지혜성・임희석 (2015). 교육데이터 마이닝을 위한 온라인 학습 활동 수집 모델 개발. **한국정보과학회 2015년 동계학술발표회 논문집,** 1358-1360.

[2] 박호진・권영현・안영민 (2013). 빅데이터와 빅데이터 정제 기술. **한국컴퓨터정보학회지, 21**(1), 1-8.

[3] Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann, Inc.

[4] Easter, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231.

[5] Zhu, Y., Ting, K. M., & Carman, M. J. (2016). Density-ratio based clustering for discovering clusters with varying densities. *Pattern Recognition, 60*, 983-997.

[6] Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016). An efficient and scalable density-based clustering algorithm for datasets with complex structures. *Neurocomputing, 171*, 9-22.