고차원 데이터에서 2차원 프로젝션을 이용한 클러스터링

장미희[0] 이혜명 박영배
명지대학교 컴퓨터공학과
leja@unitel.co.kr, hmlee@kmc.ac.kr, parkyb@mju.ac.kr

A Clustering using Two-Dimensional Projection in High-Dimensional Data

Mi-Hee Jang[0] Hye-Myung Lee Young-Bae Park
Dept. of Computer Engineering, Myongji University

요 약

데이터마이닝 기법 중의 하나인 클러스터링은 대용량 데이터베이스에서 유사한 특징을 가진 객체들을 집단화하는데 사용되는 매우 유용한 분석방법이다. 그러나 대부분의 클러스터링 알고리즘들은 고차원 데이터에서는 성능이 급격히 저하된다. 이것은 고차원 데이터 집합이 상당한 양의 잡음을 포함하고 있기 때문이며 고차원 데이터 고유의 희소성에 기인한다. 이에 따라 고차원 데이터의 구조와 특성을 지원하는 데 적합한 클러스터링 기법이 개발되고 있다.

본 논문에서는 고차원 클러스터링에서 잡음 데이터를 효과적으로 제거하기 위한 새로운 알고리즘을 제안하는데, 이 일고리즘은 고차원 데이터의 저차원으로의 변환에 기초한다. 저차원으로 변환을 위해 2차원 프로젝션을 이용하며, 반복적으로 2차원 프로젝션을 적용하여 잡음을 단계적으로 최소화한다. 이와 같은 2차원 프로젝션은 잡음을 점차적으로 줄여줄 뿐 아니라, 데이터 분포에 대한 시각화 작업에도 용이하다.

1. 서 론

1995년 이후 급속하게 보급되는 인터넷과 정보기술의 발전은 정보량의 증가를 급격하게 만들고 있다. 대용량 데이터의 증가와 저장능력의 발전으로 인하여 데이터베이스 내의 유용한 정보를 찾는 기술의 중요도는 더욱 증대되고 있다. 대용량의 데이터베이스에서 유용한 정보를 찾기 위해 데이터마이닝 기법에는 Decision Tree, Association Rule, Neural Network, Clustering, Data Visualization[11] 등이 사용된다. 또한, 데이터마이닝은 OLAP과 데이터웨어하우징(Data Warehousing)을 구축할 때 중요한 도구로 사용되고, 구체적으로는 의약분야의 경우 질병을 자동 진단, 환자의 병력관리와 투약관리, 통신분야에서 통화기록 분석이나 고객 성향분석, 금융업에서는 보험상품과 고객관계, 고객에 적합한 상품설계 등 다양한 분야로 활용될 수 있다. 여러 분야에서 활용도를 높여가고 있는 데이터마이닝은 데이터베이스를 구성하는 데이터가 갖는 특성을 찾아 유사한 그룹을 찾고 그 그룹을 분석하여 데이터간의 연관성과 특징을 추출하는 작업의 정확성이 매우 중요하다. 데이터마이닝 기법 중에서 클러스터링은 바로 유사한 특징을 갖는 데이터들의 그룹을 찾는 출발점을 제공한다. 기존의 클러스터링 알고리즘으로는 CLARANS[6], DBSCAN[3], OPTICS[4], BIRCH[8], CLIQUE[1], PROCLUS[2], CLIP[9] 등이 있는데, 이러한 알고리즘들은 고차원 데이터에서 다양한 방법으로 클러스터를 탐사한다. 데이터베이스를 구성하는 데이터들의 애튜리뷰트 수가 많아질수록 차원이 증가하여 고차원이 되면 클러스터를 탐사하기 어려워진다. 차원이 증가하면 데이터의 점들이 고차원 공간에서 갖는 희소성(sparsity)과 잡음(noise)으로 인하여 정확한 클러스터 형태 탐사에 실패할 수 있다. 클러스터링 알고리즘 중에서 CLIQUE, PROCLUS, CLIP 기법 등은 고차원 데이터에서 프로젝션 방법을 사용하여 연관성 있는 차원만 선택하여 클러스터를 탐사하는 기법으로 희소성과 잡음으로 인한 클러스터 탐사의 실패를 개선한 알고리즘들이다.

본 논문에서는 대용량 고차원데이터에서 효과적인 잡음 제거를 위한 클러스터링기법으로 2차원 프로젝션을 제안한다. CLIP은 1차원씩 점진적인 프로젝션을 통하여 클러스터 형성에 연관성이 적은 차원은 제외시키면서 후보영역을 탐색하여 정확한 클러스터의 형태를 찾는 클러스터링 기법이다. 이에 반해 본 논문에서는 고차원 데이터에 대해 반복적으로 2차원씩 프로젝션하여 클러스터의 형태를 찾아가는 기법을 적용해서 실세계 데이터에서 클러스터의 형태를 찾는데 있어 효과적으로 잡음(noise)을 제거시켜 정확한 형태의 클러스터를 얻고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 연구된 클러스터링 알고리즘에 관한 관련연구를 알아본다. 3장에서는 제안하는 클러스터링 기법에 대해서 소개하고, 4장에서는 결론 및 향후 발전방향을 고찰한다.

2. 관련 연구

고차원데이터에서 클러스터링을 위해 제안된 기존의 연구들에 대해 분석한다.

2.1 차원 전체를 고려하는 클러스터링

클러스터링 방법 중에서 CLARANS, DBSCAN, OPTICS, BIRCH는 고차원 데이터의 공간에서 거리를 기반으로 모든 차원을 고려하여 클러스터링을 한다. 이 중 CLARANS (Clustering Large Applications based upon RANdomized Search)는 전체 데이터 집합에서 임의의 샘플 데이터를 뽑아 k개의 클러스터를 대표할 수 있는 k-medoid들의 집합인 그래프를 검색하는 과정으로 처음으로 소개된 분할기법 알고리즘이다. 데이터 집합의 패턴이나 분포도가 복잡해질수록 충분한 공간 정보를 제공하지 못하고, 임의 탐색을 사용하므로 최적의 클러스터링 결과를 보장할 수 없다[6][10].

DBSCAN은 잡음을 고려하는 대표적인 알고리즘으로서 클러스터의 밀도기반 개념을 이용하며 임의 형태의 클러스터를 탐사한다. 그러나 DBSCAN은 R*-트리 기반으로 구현되므로 고차원 공간에서 R-트리 기반 인덱스의 성능저하로 말미암아 효율적으로 수행하지 못한다. 모든 인덱스-기반 방법론들이 그렇듯이 DBSCAN도 효율성에 있어서 심각한 성능저하와 잡음을 포함하는 데이터 집합에 대해 효과성 문제를 보이고 있다[3].

BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)는 잡음 데이터를 다룬 첫 번째 알고리즘으로서, 클러스터를 탐사하거나 잡음으로부터 클러스터의 구별을 위해 몇 가지 경험적 정보를 사용한다. BIRCH는 클러스터 특징을 저장하는 균형트리인 CF-트리라 불리는 계층적 데이터 구조를 사용한다. 아직까지 BIRCH는 가장 효율적인 알고리즘의 하나

이며 데이터베이스를 오직 한번 스캔하는데, 이것은 고차원 데이터에 대해서도 그렇다. 그러나 BIRCH는 요약될 데이터 항목을 정의하는데 반지름이나 지름 등의 유사성 개념을 사용하므로 오직 구형의 클러스터만을 탐색하는 한계가 있으며, 또한 데이터의 입력순서에 민감한 단점을 갖고 있다[8][10].

2.2 관련 차원만을 고려하는 클러스터링

기존의 대부분 알고리즘들은 데이터의 모든 차원을 고려하였으나 CLIQUE, PROCLUS, CLIP은 클러스터 형태에 밀접하게 연관있는 일부 차원만을 고려하는 알고리즘이다. 데이터의 차원이 증가할수록 데이터 고유의 희소성 문제와 상당량의 잡음 때문에 모든 차원을 고려하여 클러스터링을 하게 되면 성능이 현저히 떨어지는 것을 볼 수 있다.

CLIQUE는 부분차원 클러스터링에 관한 첫 번째 연구로 의의가 있다. CLIQUE는 밀도와 그리드 기반의 클러스터링 기법으로 고차원 데이터에서 차원간의 교차-곱으로 이루어진 단위(unit)에 포함된 점들의 수가 기준 밀도를 초과하면 밀집(dense)하다라고 정의한다. 클러스터는 차원들이 연관된 밀집 단위들의 집합이다. CLIQUE는 밀집하지 않은 차원이 있는 경우에 대비해 프로젝션을 이용해서 고차원 데이터의 밀집공간에 관여하는 일부 차원으로 이루어진 클러스터의 형태를 찾아내는데 효과적인 방법을 제시했다. 그러나 탐색된 밀집영역사이에 큰 오버랩이 존재하고, 그리드 기반으로 인하여 셀의 수가 지수적으로 증가하는 문제점이 있다[1][9].

PROCLUS(Projected Clustering)는 CLARANS에 관련 차원을 찾는 과정을 결합한 방법으로 CLARANS 기법처럼 k개의 medoid를 선택한 후에, 각 medoid와 관련이 높은 차원을 찾아내 클러스터링을 한다. 주어진 반복회수만큼 이 과정을 반복해서 각 medoid에 대응하는 차원의 집합을 계산한다. 이와 같이 PROCLUS는 각 medoid에 대응하는 차원의 집합을 계산하므로서 차원을 감소시킬수 있으나 대용량의 데이터 집합의 경우 데이터의 희소성과 잡음 데이터로 인하여 최상의 medoid를 탐색하는 시간을 예측하기 어려우며 적절한 medoid를 구하지 못한 경우에는 데이터 손실로 클러스터링 결과의 신뢰도를 저하시킬 가능성이 있다[9]. Projected Cluster란 전체 차원 중에서 부분 차원에 존재하는 클러스터를 의미한다.

CLIP(Clustering based on Incremental Projection)은 k-차원 데이터 공간에서 한 차원씩 점진적으로 프로젝션하면서 클러스터를 탐색한다. 즉, 하나의 차원에서 시작하여 밀집영역을 구한 뒤, 그 차원의 밀집영역에 의존적인 그 다음 차원의 밀집영역을 찾아내는 방법으로 최종 k차원까지 반복해 나가는 것이다. 프로젝션하는 순서는 임의적일 수 있어 차원의 중요도에 따라 우선 순위를 부여하여 순서를 결정할 수 있다. 클러스터링 과정은 부분 차원 탐색 과정과 클러스터 식별 과정으로 구성되어 있다. 부분차원 탐색과정에서는 한 차원에 대해서 축-평행하게 프로젝션한 후, 데이터의 분포를 계산하여 데이터가 기준밀도 이상인 밀집영역을 찾아 밀집영역에 해당하는 초월사각형(hyper-rectangle)부분에 존재하는 데이터에 대해서만 그 다음 차원의 값을 프로젝션하는 방법으로 k차원까지 진행하여 후보영역을 구한다. 클러스터 식별과정에서는 후보영역에서 클러스터의 형태를 영역에 속한 데이터 점들의 평균값을 이용하여 구체화한다[9]. 1차원 프로젝션 방법은 프로젝션하는 1차원에서는 완전한 클러스터를 찾으므로 더 이상 1차원적인 잡음은 존재하지 않으나 고차원의 잡음은 여전히 존재할 수 있다.

3. 제안하는 클러스터링 기법

본 논문에서는 효과적인 잡음 감소를 위한 클러스터링 알고리즘으로 k차원의 데이터에서 임의로 두 개의 차원을 선택하여 2차원씩 축-평행한 프로젝션을 반복하며 클러스터를 탐사하는 방법을 제안한다. 고차원 데이터는 상당한 양의 잡음(noise)을 포함하고 있으므로 클러스터링의 효과성이 급격히 저하된다. 제안하는 알고리즘은 이와 같은 고차원 문제를 해결하기 위해 고차원을 저차원으로 변환하는 2차원 프로젝션을 사용한다. 즉 데이터 집합에 대해 두 개의 차원을 프로젝션한 데이터의 분포를 조사하여 밀도 임계값 이상 되는 영역만을 선택한다. 그 다음은 선택된 영역에 속한 데이터들에 대해 또 다른 2차원적 프로젝션을 적용하여 밀도 임계값 이상의 영역을 선택한다. 이러한 단계를 반복하면서 고차원에 존재하는 잡음(noise) 데이터를 효과적으로 제거할 수 있다.

3.1 2차원 프로젝션 정의

직접적인 고차원 데이터의 프로젝션은 수식의 복잡성에 의하여 매우 큰 계산비용이 소요되므로 실현 가능한 2차원 프로젝션을 반복적으로 적용하여 고차원 프로젝션을 근사적으로 구현하고자 한다. 이 과정에서 고차원 데이터의 잡음이 점차적으로 제거된다. 2차원 프로젝션은 다음의 정의1에 근거하여 수행한다.

정의1 : 2차원 프로젝션

(i,j)축의 2차원 프로젝션 $P_{i,j}$는 다음과 같이 정의한다.

$P_{i,j}(x^1, x^2, \cdots, x^k) = (x^i, x^j)$ 혹은

$P_{i,j}(x_1, x_2, \cdots, x_k) = (0, \cdots, 0, x_i, 0, \cdots, 0, x_j, 0 \cdots, 0)$

즉, (x_i, x_j)만 제외하고 모두 0(zero)이다.

전체 k차원 데이터의 프로젝션 위하여 임의 2차원의 쌍 (i,j)에 대하여 반복적으로 프로젝션을 하는 것이다. 이 경우 연산횟수가 총 ${}_kC_2 = k(k-1)/2$ 번이라는 많은 횟수의 프로젝션이 필요하다. 따라서 본 논문에서는 이 중 (i,j) 쌍을 선택하는 프로젝션의 하나의 예로서, (1,2) (2,3) (3,4) (4,5) (5,6) (6,7) (7,8) ...(k-1,k) 쌍의 차원에 대한 프로젝션을 반복적으로 수행한다. 차원의 쌍을 선택하는 경우의 수는 임의로 변경될 수 있는 확장성을 고려한다.

정의2 : 단계적으로 반복적인 2차원 프로젝션

(P_i, P_j)프로젝션 결과와 (P_m, P_n) 프로젝션의 결과를 쌍으로 2차원 프로젝션한다. $P_{Px,Py}$는 다음과 같이 정의한다.

$P_{Px,Py}(P^1, P^2, \cdots, P^{k-1}) = (P^x, P^y)$ 혹은

$P_{Px,Py}(p_1, p_2, \cdots, p_{k-1}) = (0, \cdots, 0, p_x, 0, \cdots, 0, p_y, 0 \cdots, 0)$

단계적으로 반복적인 프로젝션은 1단계에서 2차원 프로젝션이 차원의 모든 쌍을 대상으로 이루어지면, 2단계 프로젝션에서는 1단계 프로젝션의 결과 쌍을 대상으로 이루어진다.

2단계 프로젝션이 된 후에도 프로젝션 된 결과의 쌍이 있다면 더 이상 프로젝션 할 쌍이 없을 때까지 반복적인 2차원 프로젝션을 수행한다. 이 때 클러스터 형성에 관련이 없다고 식별된 차원은 단계적으로 반복적인 프로젝션에서 제외한다. 그림 1은 단계적으로 이루어지는 반복적인 프로젝션을 표시한다.

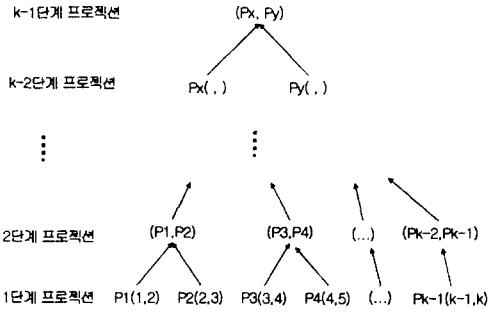

그림 1 단계적으로 반복적인 프로젝션

3.2 2차원 프로젝션으로 관련차원 식별

5차원의 100,000개 레코드를 가진 데이터 집합에 4개의 차원

이 관련된 클러스터가 1개 있는 경우를 예로 2차원 프로젝션 알고리즘을 시뮬레이션 했다. 입력되는 각 차원의 범위는 [0,100]이며, 클러스터 형성에 연관된 차원은 2, 3, 4, 5 차원이고, 잡음 데이터는 전체 데이터의 10%이다. 실험에서 선택한 프로젝션의 쌍은 (1,2) (2,3) (3,4) (4,5)이다. 그림2는 1차원과 2차원을 각각 X, Y축으로 2차원 프로젝션 한 $P_{1,2}$의 결과로 Y축만 클러스터와 관련이 있고, X축의 1차원은 클러스터 형성과 관련이 없다는 것을 알 수 있다. 그림3은 2차원과 3차원이 클러스터 형성에 관련이 있다는 것을 의미한다. 그림4는 3차원과 4차원이 클러스터 형성에 관련이 있다는 것을 의미한다. 그림5는 4차원과 5차원이 클러스터 형성에 관련이 있다는 것을 알 수 있다. 이 때 잡음 데이터도 함께 표시되므로 고차원데이터에서 잡음 데이터를 제거해야 잡음 데이터와 식별되는 클러스터의 형태를 정확하게 확인할 수 있다.

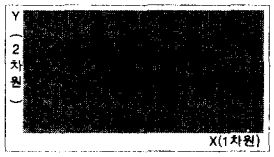

그림2 1,2차원 프로젝션 결과

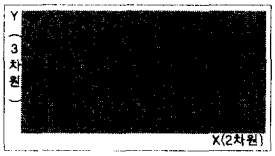

그림3 2,3차원 프로젝션 결과

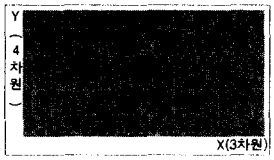

그림4 3,4차원 프로젝션 결과

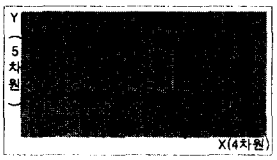

그림5 4,5차원 프로젝션 결과

그림2, 그림3, 그림4, 그림5의 프로젝션 결과에서 그림2의 1,2차원의 프로젝션 결과에서는 클러스터를 발견할 수 없으므로 2단계 프로젝션에서 제외한다. 이 때 1차원이 클러스터 형성에 연관되지 않았다는 것을 검증하기 위해 (1,5)차원의 쌍으로 프로젝션한다. 여기서 2차원은 (2,3) 차원의 쌍에서는 클러스터가 발견되었으므로 2차원은 클러스터에 연관되어 있다는 의미이므로 1차원만 새로운 쌍으로 검증해 본다. 그림6에서처럼 (1,5) 차원의 쌍으로 2차원 프로젝션했을 때 클러스터의 형태가 발견되지 않으므로 1차원이 클러스터 형성에 연관되지 않았다는 것이 검증된다.

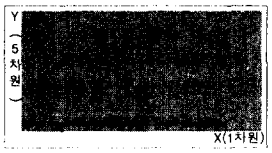

그림6 1,5차원 프로젝션 결과

3.3 단계적인 반복적 프로젝션으로 잡음 제거

1단계에서 2차원 프로젝션 된 결과인 (2,3)차원 결과와 (3,4)차원 결과를 2단계에서 다시 프로젝션한 후에 (2,3)차원의 결과를 그림7로 표시하였다. 2, 3, 4차원이 연관된 데이터 집합에서 잡음이 제거된 클러스터의 형태를 확인할 수 있다.

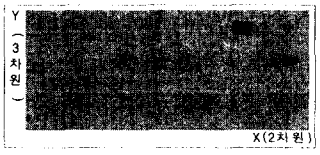

그림7 2단계 프로젝션 후 2,3차원

이와 같이 단계적으로 반복적인 프로젝션을 2차원씩 수행하면 잡음이 제거된 클러스터의 형태를 확인할 수 있다.

4. 결론 및 향후 연구

본 논문에서는 고차원 데이터의 잡음을 제거하는 클러스터링 기법을 제안하였다. 반복적인 2차원 프로젝션을 통하여 고차원 데이터의 잡음을 효과적으로 제거하면서 정확한 형태의 클러스터를 탐사할 수 있으며, 또한 클러스터 형성에 관련되는 차원과 관련되지 않는 차원을 식별할 수 있었다. 또한, 이러한 클러스터링 결과는 2차원 그래프를 이용하여 시각적으로도 확인이 가능하였다. 본 연구는 현재 알고리즘의 효율성과 최적화를 위하여 2차원 프로젝션의 쌍을 효율적으로 선택하는 방법과 3차원 프로젝션을 활용할 수 있도록 확장하는 방법을 구현 중이다. 앞으로 다양한 실험을 통하여 다른 방법들과 비교해야 할 것이다.

참고문헌

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan, "Automatic subspace Clustering on High Dimensional Data Mining Applications," *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pp. 94-105, 1998

[2] Charu C. Aggrawal, Cecilia Procopiuc, Joel L. Wolf, Philip S. Yu, and Jong Soo Prk, "Fast Algorithms for Projected Clustering," *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pp. 61-72, 1999

[3] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial database with noise," *Proc. of Int. Conf. on Knowledge Discovery and Data Mining*, 1996

[4] Mihael Ankerst, Markus M. Breunig, Han-Peter Kriegel, and Jorg Sander, "OPTICS: Ordering points to identify the clustering structure," *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 1999

[5] Hinneburg A., Keim D. A, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," *Proc. of 4th Int. Conf. on Knowledge Discovery and Data Mining*, 1998

[6] Raymond T. Ng, Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," *Proc. of 20th Int. Conf. on VLDB*, pp. 144-155, 1994

[7] Wei Wang, Jiong Yang, and Richard Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," *Proc. of 23rd Int. Conf. on VLDB*, pp. 186-195, 1997

[8] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pp. 103-114, 1996

[9] 이혜명, 박영배, "고차원 데이터에서 점진적 프로젝션을 이용한 클러스터링," 한국정보과학회 가을학술발표논문집(Ⅰ), 2000

[10] 손은정, 강인수, 김태원, 이기준, "클러스터링 분석에 의한 공간 데이터마이닝 방법," 한국정보과학회 가을 학술발표논문집(2), 1998

[11] Pieter Adriaans, Dolf Zantinge, "데이터마이닝," (그린), 1998

[12] Kaushik Chakrabarti, Sharad Mehrotra, "Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces," *Proc. of 26th Int. Conf. on VLDB*, pp. 89-100, 2000.