데이터마이닝을 위한 셀-기반 클러스터링 방법의 성능비교

진두석[0] 장재우
전북대학교 컴퓨터공학과
{dsjin, jwchang}@dblab.chonbuk.ac.kr

Performance Comparison of Cell-based Clustering Method for Data Mining Applications

Du-Seok Jin[0] Jae-Woo Chang
Dept. of Computer Engineering, Chonbuk National University

요 약

최근 데이터마이닝 응용분야에서 대용량의 고차원 데이터가 증가하고 있기 때문에 이를 효율적으로 처리할 수 있는 방법이 요구된다. 이를 위해 CLIQUE 방법과 셀-기반 클러스터링 방법이 제안되었다. 본 논문에서는 대용량의 고차원 데이터에 적합한 클러스터링 방법을 선택하기 위해, 셀-기반 클러스터링 방법을 CLIQUE 방법 및 CLIQUE 방법에 근사정보(Approximation) 를 결합한 방법과 성능 비교를 수행한다. 성능비교 결과, 셀-기반 클러스터링 방법이 데이터 클러스터링 및 데이터 검색시간에서 가장 우수한 성능을 보이며, 정확율은 CLIQUE 방법에 비해 다소 뒤떨어지나 전체적인 효율성에서 매우 우수한 성능을 보인다.

1. 서론

데이터마이닝은 대용량의 데이터베이스에서 숨겨진 지식, 패턴, 연관된 규칙 등을 발견하는 방법이다. 데이터들을 분석하여 서로 유사한 그룹으로 데이타를 클러스터링하는 방법은 데이터마이닝의 중요한 분야 중 하나이다. 일반적으로 클러스터링 방법은 클러스터들 사이에 계층구조를 가지고 있는 계층적(hierarchical) 클러스터링과 데이터 공간을 정해진 수의 클러스터에 따라 분할하고 클러스터링 하는 공간분할(space partitioning) 클러스터링 방법이 있다. 기존 클러스터링 알고리즘은 데이터 셋이 메모리 상주(memory resident)데이터로 제한되어 있었기 때문에 수 백만건 이상을 가진 대용량의 데이터베이스에는 적합하지 못한 단점이 있다. 아울러 기존 연구는 저차원 데이터의 클러스터링에는 적합하지만 데이터 셋의 차원이 높아질수록 급격한 성능 저하를 보이고 있어 고차원 데이터에는 부적합한 단점이 있다.

최근 데이터마이닝 응용분야에서 대용량의 고차원 데이터가 증가하고 있기 때문에, 이를 효율적으로 처리할 수 있는 방법이 요구된다. 따라서, 고차원 데이터를 효과적으로 처리하는 CLIQUE 방법과 셀-기반 클러스터링 방법이 제안되었다. 본 논문에서는 대용량의 고차원 데이터에 적합한 클러스터링 방법을 선택하기 위해, 셀-기반 클러스터링 방법을 CLIQUE 방법 및 CLIQUE 방법에 근사정보(Approximation) 을 결합한 방법과 성능 비교를 수행한다.

본 논문의 구성은 다음과 같다. 2 장에서는 데이터마이닝을 위한 클러스터링 방법에 대해 기술하고, 3 장에서는 성능평가 환경을 제시한다. 4 장에서는 세가지 방법의 성능을 비교하고, 마지막으로 5 장에서는 결론 및 향후연구방향을 제시한다.

2. 데이터마이닝을 위한 클러스터링 방법

본 장에서는 기존의 클러스터링 방법에 대해 소개한다. 첫째, BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)[1] 방법에서는 대용량의 데이터 집합을 위해서 CF(Clustering Feature)-tree 를 이용한다. CF-tree 는 B(Branching factor)와 T(Threshold)값을 가진 균형 트리로서 모든 점들을 저장하는 대신에 서브클러스터의 요약정보를 사용한다. CF-tree 의 중간노드는 그 노드의 모든 서브클러스터들로 구성된 하나의 클러스터를 나타내며 리프노드에 있는 각각의 엔트리의 지름은 T 보다 작은 값을 가지고 있어야 한다. 따라서, T 값에 따라 트리의 크기가 결정된다. CF-tree 는 데이터의 삽입과정에서 동적으로 구축되며 데이터 점들을 한번만 읽어서 트리를 구성할 수 있다는 장점이 있다.

둘째, CLIQUE(CLustering In QUEst)[2] 방법은 관련된 차원만을 선택하여 선택된 부분공간에서 클러스터링을 처리하는 방법으로 고차원 데이터에서 큰 밀도를 가진 영역을 찾는 효과적인 방안을 제시하였다. CLIQUE 는 밀도 및 격자 기반 클러스터링 방법이다. 즉 하나의 클러스터는 주위보다 높은 밀도를 가진 영역을 의미한다. CLIQUE 는 데이터 점들의 근사 밀도값을 구하기 위해

서 각 차원을 일정한 간격(ξ)으로 분할하고 , 각 차원에서 이러한 분할 공간의 교차-곱으로 이루어진 Unit 에 포함된 점들의 수가 기준밀도(τ)를 초과하면 밀집(dense) 하다고 정의한다. 따라서, 각 클러스터는 서로 인접한 밀집 Uint 들의 집합을 의미한다. 이와 같이 CLIQUE 는 부분공간 클러스터링 방법을 이용하여 고차원 데이터 공간에서 차원의 감소를 시도하여 고차원의 데이터의 클러스터링에 효과적인 방법을 제시하였으나, 각 차원을 일정한 간격으로 분할하기 때문에 클러스터의 정확한 형태를 표현하기 어려운 단점이 있다.

셋째, 셀-기반 클러스터링[3] 방법은 데이터의 효율적인 셀 구성 알고리즘을 통해 적은 수의 클러스터를 생성하고, 빠른 검색 처리를 위해서 근사 방법을 적용하며 이를 위해서 필터링 기법을 이용한 저장 인덱스 구조를 사용한다. 따라서 기존 클러스터링 방법들에서 발생하는 메모리의 제한 문제와 차원이 증가함에 따라 셀의 수가 지수적으로 증가하는 문제점을 개선한 방법으로 대용량 고차원 데이터에 대한 효율적인 클러스터링이 가능하다. 고차원 대용량 데이터의 경우 구성된 클러스터의 수가 매우 많아지기 때문에, 이를 저장한 클러스터정보파일(cluster information file)의 검색 과정에 많은 시간이 소요된다. 따라서 이러한 문제점을 보완하기 위해 셀-기반 클러스터링 방법은 근사정보파일(approximation information file)을 사용하여 필터링(filtering)을 수행함으로써 크기가 매우 큰 클러스터정보파일에 대한 검색 횟수를 줄여 전체 검색 성능을 향상시킨다.

3. 성능평가 환경

본 장에서는 셀-기반 클러스터링 방법, CLIQUE 방법 및 CLIQUE 방법에 근사정보(Approximation) 을 결합한 방법의 성능평가 환경에 대해 설명한다. 성능평가를 위한 시스템 환경은 CPU 650 MHz dual, 메모리 512MB 의 리눅스 서버에서 수행하였다. 사용된 데이터는 IBM Quest Data Mining project[4] 의 Synthetic Data Generation Code for Classification 을 이용하여 16 차원 100 만건의 데이터를 만들어 사용하였다.

표 1 은 본 논문에서 사용한 Synthetic Data 에 대한 설명이다. 실험에 사용한 8 차원의 데이터는 salary, commission, age, elevel, zipcode, hvalue, hyears, loan 으로 구성된 레코드 셋을 사용하였고, 16 차원의 데이터는 8 차원의 데이터에 area, children, tax, interest, ctype, cyear, job, balance 가 추가된 데이터 셋이다. 여기에서, salary, commission, age, hvalue, hyears, loan, tax, interest, cyear, balance 는 수치(numeric) 애트리뷰트에 속하고, elevel, zipcode, area, children, ctype, job 은 범주(categorical) 애트리뷰트에 속한다. 성능평가는 셀-기반 클러스터링 (Cell-based Clustering: CBC) 방법과 대용량 고차원 데이터의 처리가 효율적인 CLIQUE 방법과, 아울러 CLIQUE 방법에 근사정보파일을 사용한 방법(CLIQUE+A)에 대해 100 만 건의 데이터를 클러스터링하는 시간, 클러스터 정보의 검색시간 그리고 정확율과 검색시간의 가중치에 따른 효율성을 측정하였다.

표 1 실험 데이터의 애트리뷰트 속성

Attribute	Description	Value
Salary	Salary	uniformly distributed from 20000 to 150000
commission	Commission	salary >= 75000 => commission =0 else uniformly distributed from 10000 to 75000
age	age	uniformly distributed from 20 to 80
elevel	Education level	uniformly chosen from 0 to 4
zipcode	Zip code	uniformly chosen from 9 available zipcodes
hvalue	Value house owned	uniformly distributed from 0.5k100000to 1.5k100000 where k∈0...9 depends on zipcode
hyears	Years house owned	uniformly distributed from 1 to 30
loan	Total loan amount	uniformly distributed from 0 to 500000
Area	Area code	uniformly chosen from 20 available area codes
Children	The number of child	uniformly chosen from 0 to 4
Tax	Tax	salary < 60000 => tax =0 else = salary*0.01
Interest	Interest	loan < 100000 => interest = loan*0.01 else interest = loan*0.02
Ctype	Car type	uniformly chosen from 10 available car types
job	Job type	uniformly chosen from 20 available job types
cyears	Years car owned	uniformly distributed from 1 to 10
balance	Total balance amount	uniformly distributed from 0 to 500000

4. 성능평가

그림 1 은 16 차원 100 만건의 데이터를 클러스터링하는 시간을 나타낸다. 이때 세가지 방법 모두에 동등한 조건으로 클러스터링 시간을 측정하기 위하여 구간 임계값은 0 으로 설정하고 클러스터링 시간을 측정하였다. 그 결과 세가지 방법 모두 데이터가 증가함에 따라 클러스터링 시간이 선형적으로 증가함을 보이고 있다. 따라서 세 방법 모두 대용량의 데이터를 처리하기에 적합한 방법임을 알 수 있다. 실험 결과 100 만건의 데이터를 클러스터링하는 시간은 CLIQUE 방법은 약 730 초 정도이고, CLIQUE+A 방법은 약 1200 초, CBC 방법은 약 100 초 정도 소요된다. 즉, CBC 방법은 CLIQUE 방법에 비해 매우 적은 수의 셀을 생성하므로 클러스터링 시간이 약 85%가 감소되었고, CLIQUE+A 방법은 근사정보파일을 만드는 시간이 추가되어 클러스터링 시간이 CLIQUE 방법에 비해 약 160%정도 소요된다.

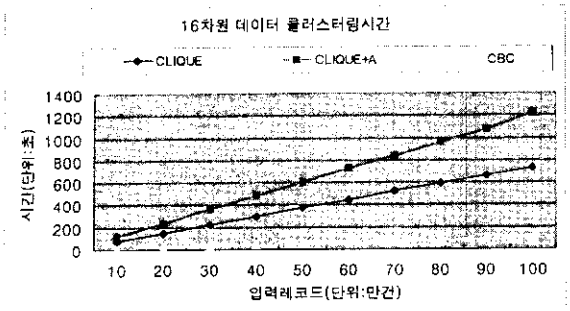

그림 1. 데이터 클러스터링 시간

그림 2 는 100 만건 데이터의 클러스터 정보를 저장한 후 질의에 대한 평균검색 시간을 나타낸다. 8 차원 데이터의 경우 CLIQUE 방법은 약 22 초 정도 소요되며, CLIQUE+A 방법은 약 7 초이며, CBC 방법은 약 1 초 정도 소요된다. CBC 방법이 검색시간이 빠른 이유는 셀 구성 알고리즘을 통해 적은 수의 셀을 만들며, 아울러 근사정보파일을 사용하여 필터링 효과를 얻기 때문이다. 그림 3 은 세가지 방법의 정확율을 나타낸다. CLIQUE 방법이 평균 94%정도의 정확율을 나타내며, CLIQUE+A 방법과 CBC 방법은 약 92~93%의 정확율을 나타낸다.

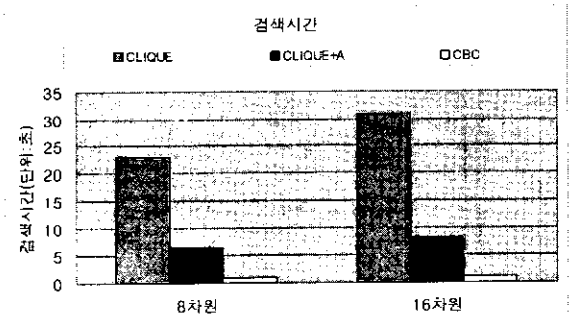

그림 2. 클러스터 정보 검색시간

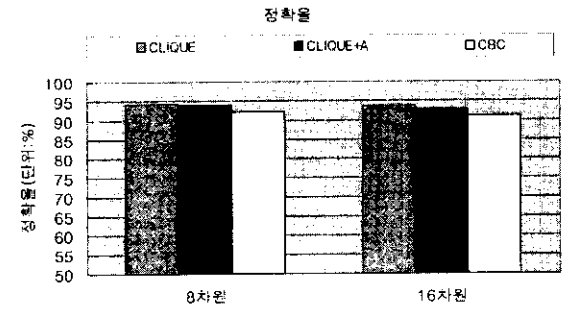

그림 3. 정확율

세가지 방법이 검색시간과 정확율 간의 트레이드오프(trade-off)관계에 있기 때문에, 다음 식을 이용하여 효율성을 측정하였다. E_{MD} 는 3 가지 방법에 대한 검색시간과 정확율 간의 효율성을 나타내고, 정확율에 대한 가중치는 Wp, 검색시간에 대한 가중치를 Wt 로 정의한다. P_{MD} 와 T_{MD} 는 3 가지 방법들에 대한 정확율과 검색시간을 나타낸다. 그림 4 는 Wt 가 1 로 고정되었을 때, Wp 가 1, 2, 4 로 변화시킬 때의 효율성을 나타낸다. CBC 방법이 Wp 가 2 일 때 (즉 정확률의 가중치가 검색시간의 2 배일 때)까지 타 방법에 비해 매우 우수한 성능을 보인다.

$$E_{MD} = W_p \cdot \frac{P_{MD}}{P_{MAX}} + W_t \cdot \frac{1}{\frac{T_{MD}}{T_{MIN}}}$$

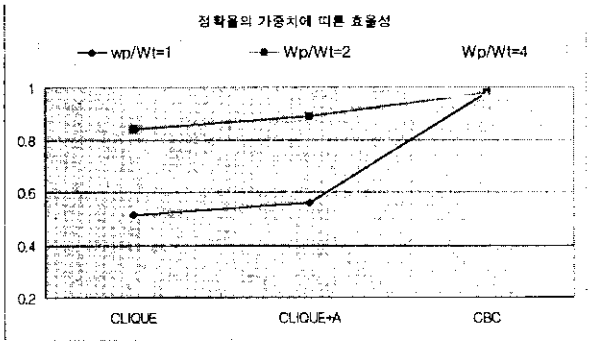

그림 4. 세가지 방법의 효율성

5. 결론 및 향후연구

최근 데이터마이닝 분야에서 고차원 대용량 데이터에 대한 클러스터링 방법이 중요한 이슈가 되고 있다. 따라서 본 논문에서는 대용량의 고차원 데이터에 적합한 클러스터링 방법을 선택하기 위해, 셀-기반 클러스터링 방법, CLIQUE 방법 및 CLIQUE 방법에 근사정보화일을 결합한 방법과 성능 비교를 수행하였다. 성능비교 결과, 셀-기반 클러스터링 방법이 데이터 클러스터링 및 데이터 검색시간에서 가장 우수한 성능을 보인다. 정확율은 CLIQUE 방법에 비해 다소 뒤쳐지나 전체적인 효율성에서 우수한 성능을 보인다. 향후연구로는 실제 데이터 마이닝 응용에 적용하여 성능비교를 수행하는 것이다.

참고 문헌

[1] T. Zhang, R. Ramakrishnan, and M. Linvy, "BIRCH : An Efficient Data Clustering Method for Very Large Databases", Proc. ACM SIGMOD, 1996, pp. 103-114.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data Mining Applications", Proc. ACM SIGMOD, 1998, pp. 94-105.

[3] 진두석, 장재우, "데이터 마이닝을 위한 대용량 고차원 데이터의 셀-기반 분류방법", 한국정보과학회 추계 학술발표논문집, 2000, pp.192-194.

[4] http://www.almaden.ibm.com/cs/quest.