빅데이터 클러스터링을 위한 K-Means 초기 중심 선정 연구

김영주*, 허유경*, 백종상**, 정환종**, 이성로**, 정민아*
*목포대학교 컴퓨터공학과
**목포대학교 정보전자공학과
e-mail: xfile7@mokpo.ac.kr

A Study on Initial Seeds Selection of K-Means for Big Data Clustering

Yeong-Ju Kim*, Yu-Gyeong Heo*, Jong-Sang Back**, Hwan-Jong Jeong** ,
Sung-Ro Lee**, Min-A Jung*
*Dept of Computer Engineering, Mok-po University
**Dept of Information & Electronic Engineering, Mok-po University

요 약

K-Means 알고리즘은 구현이 쉽고, 패턴수가 n일 때 시간 복잡도가 O(n)인 장점을 가져 대용량 데이터에서 널리 이용된다. 그러나, K-Means 알고리즘은 초기 클러스터 중심을 어떻게 선정하는가에 따라 할당-재계산 횟수, 클러스터링 결과를 결정짓는다. 본 논문에서는 K-Means 알고리즘에서 클러스터 초기 중심 선정 연구를 살펴보고 계통임의추출법을 적용하여 K-Means 초기 중심 선정 방법을 제안한다. 제안한 방법은 대용량 데이터의 클러스터링 시간을 감소하고 정확도를 향상시킬 수 있다.

1. 서론

클러스터링은 여러개의 측정치들로 구성된 데이터데이터를 유사한 그룹으로 묶어

대규모 데이터에 대한 특성 값에 따라 몇 개의 클러스터로 군집화하는 클러스터링 기법은 계층적 클러스터링이나 분할 클러스터링 등 다양한 기법으로 나누어 설명할 수 있는데 현대 사회의 정보 대량화는 계층적 클러스터링이나 그래프이론 클러스터링으로는 처리할 수 있는 데이터에 한계가 있고 시간 복잡도 측면에서 비효율적이다[1].

이러한 이유로, 본 논문은 가장 잘 알려지고 일반적으로 사용되고 또한, 가장 많이 연구되고 있는 K-Means 알고리즘을 빅데이터에서 효율적인 클러스터링을 하기위해 연구한다. 우선, K-Means의 성능이 최대가 되기 위해서는 다음과 같은 문제를 개선해야 한다. ① 초기 중심 선정: 초기 중심 선정에 따라서 클러스터링 반복 횟수, 전체 클러스터링 연산횟수, 최종 클러스터링 결과가 Good 또는 Fail 된다. ②클러스터 K선정: n개의 객체를 K개의 군집으로 나눌 때 각 객체를 어떤 군집에 배정하는 것이 전체 거리를 최소로 하는가? (최적화 개선)

본 논문은 일반적으로 무작위로 선정하는 K-Means의 초기 중심 선정에 대하여 기존의 연구를 조사하고 빅데이터에서 효율적인 클러스터링을 위한 초기 중심 선정 방법을 제안한다.

2. 관련연구

2.1 K-Means 초기 중심 선정 알고리즘

K-Means는 초기 클러스터 중심의 선정이 무작위로 이루어짐에 따라 클러스터 성능 또한 이 초기 클러스터 중심에 종속적일 수밖에 없다는 한계를 가지고 있다. 이러한 문제를 개선하기 위해서 많은 연구가 진행되어져 왔다. 그 중에서 초기 중심들을 데이터 집합에 고르게 분포시켜 클러스터링의 성능을 개선하는 연구[2]를 보면 분산도가 높은 중심들을 얻기 위하여 중심 간의 거리를 최대로 하는 방법, 삼각형의 높이를 이용하는 방법, 삼각형의 넓이를 이용하는 방법으로 초기 중심을 선정하였다. ① 최대 거리를 이용한 방법은 초기 중심 간의 거리를 최대로 하는 방법으로 실험결과 초기 중심의 일부분이 밀집되는 현상이 발생되었다. 즉, 높은 분포도와 거리가 먼 결과를 도출하였다. ② 삼각형의 높이를 이용하는 방법은 ①의 바람직하지 못한 경우를 해결하기 위해 삼각형의 높이를 이용한 방법으로 중심 간의 거리 대신 높이를 계산하여 높으면 중심을 대체하는 방법이나 ①번과 같은 밀집현상이 나타났고 (클러스터 개수)K가 2일 경우 적용할 수 없다는 단점이 도출되었다. ③최대 평균 거리 알고리즘은 초기 클러스터 중심들을 가능한 멀리 선정하는 것으로 무작위 선정된 초기 클러스터 중심이 일부 영역으로 편향되는 현상을 막을 수 있고, 이에 따라 클러스터링 속도 향상과 클러스터의 정확도를 높이고자 하는 방법으로 실험결과 초기 클러스터 중심을 무작위로 선정하는 방식에서 벗어나 초기 중심들을 최대한 멀리 배치함으로써 클러스터링의 정확도가 향상되었고 초기 클러스터 중심에 종속적이던 현상을 해소하여 일관된 결과를 얻을 수 있었다.

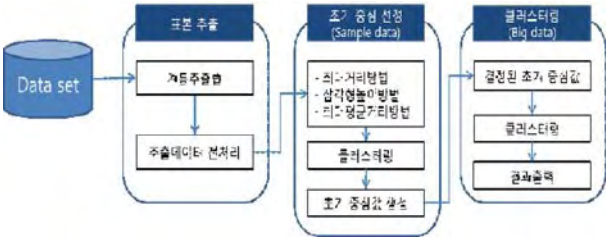
그 외에도 [3]에서는 임의의 한 패턴을 선택하는 대신 선택된 초기 클러스터에서 색인어와 가중치로 표현되는 세 개의 문서를 선택하여 초기 클러스터 중심 벡터로 설정한다. [4]은 클러스터간의 분리 크기에서 거리를 고려한다면, 각 최적 중심은 초기 센터를 가질 수 있을 것이라는 점에서 출발한다. [5]은 통신 보안 시스템에 적용하기 위하여 프로토콜을 대상으로 하는 K-Means 알고리즘을 연구하는데 이를 Two-Party K-Means 클러스터링 프로토콜이라 한다. 기본 아이디어는 문서의 중앙에서 출발하여 중심을 찾는 방법이다.

3. 빅데이터 클러스터링

3.1 실험환경

Amazon EC2를 사용하여 하둡을 이용한 병렬 컴퓨팅 환경을 구축한다. Instance는 약 60개 정도를 사용하여 대용량의 데이터셋을 유연하게 처리하도록 한다. 각 Instance의 사양은 m3.xlarge (13 ECUs, 4 vCPUs, 2.5 GHz, Intel Xeon E5-2670v2, 15 GiB memory, 2 x 40 GiB Storage Capacity)이다.

3.2 클러스터링 알고리즘



[그림1] 클러스터링 알고리즘

빅데이터 환경에서 클러스터링 알고리즘은 위 [그림1]과 같이 크게 세부분으로 나뉜다. 첫 번째, 표본추출과정은 계통추출법으로 표본데이터를 추출하고 추출된 표본 데이터는 클러스터링을 하기 위해 전처리 된다. 두 번째, 초기 중심선정 과정은 다양한 초기중심선정 방법으로 초기 중심 선정을 하고 클러스터링하여 초기 중심값을 생성한다. 세 번째, 클러스터링 과정은 두 번째에서 결정된 초기 중심값으로 클러스터링을 진행하여 결과를 출력한다.

3.3 제안한 초기 중심 선정

3.3.1 표본추출

큰수의법칙(Strong Law of Large Number)은 큰 모집단에서 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다는 통계와 확률 분야의 기본개념이다[6]. 즉, 빅데이터에서 무작위로 뽑은 표본의 평균은 빅데이터 전체의 평균을 나타낸다. 이와 같은 이론을 바탕으로 표본을 추출하고 추출된 표본의 평균과 분산을 이용하여 K-Means의 초기 중심을 선정 하는 것은 클러스터링 반복 횟수, 전체 클러스터링 연산횟수, 최종 클러스터링 결과에 영향을 준다.

3.3.2 초기 중심 선정 알고리즘

계통추출법은 체계적 표집(systematic sampling)이라고도 하며, 첫번째 요소는 무작위로 선정한 후 목록의 매번 k번째 요소를 표본으로 선정하는 표집방법이다. 모집단의 크기를 원하는 표본의 크기로 나누어 k를 계산한다. 여기서 k는 표집간격이라고 불린다[6].

모집단이 3,000,000이고 1,500을 표본으로 추출한다고 가정하면, (3,000,000/1,500=2,000) 임의로 선택된 출발점에서 시작하여 매 2,000번째 마다의 표본을 추출하는 것이다.

초기 중심 선정 알고리즘은 아래 [그림2]와 같다.

1. $\frac{X}{k}$, 표집간격(Sampling Interval) k를 구한다. 여기서, X는 전체 데이터 집합, N은 추출하려는 표본 집합이다.
2. $0 \sim K$ 사이의 임의의 표본 $e1$을 추출한다. $(X > K, e1 \in X)$
3. $e1 + k = e2$, 표본 $e1$에 표집간격을 더하여 다음 표본을 추출한다.
4. $do\,while(e \leq X)$, 전체 데이터 집합 X보다 작거나 같을 때까지 반복한다.$(i \leq n)$
5. 추출된 표본을 K-Means로 클러스터링 한다.
6. $C_1, C_2, C_3, \ldots, C_i$), $(E(X_1), E(X_2), E(X_3), \ldots, E(X_i))$, C_i개 클러스터와 각각의 중심값 $E(X_i)$가 출력된다.
7. $C_{init} = (E(X_1), E(X_2), E(X_3), \ldots, E(X_i))$, 6번에서 생성된 클러스터의 중심값을 빅데이터 클러스터링의 초기 중심값으로 선정한다.

[그림2] K-Means 초기 중심 선정 과정

4. 결론

본 논문은 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 상당히 효과적인 것으로 알려져 있는 K-Means의 성능이 최대가 되기 위한 개선 사항 중 초기 중심 선정을 통계적 확률 추출 방법을 적용하여 제안하였다. 향후 연구로 기존에 개선된 K-Means 중심 선정 방법과 제안한 K-Menas의 중심 선정 방법의 성능을 비교 평가 할 계획이다.

ACKNOWLEDGMENT

이 논문은 미래창조과학부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업(NIPA-2014-H0401-14-1009)과 2014년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2009-0093828)의 연구결과로 수행되었음.

참고문헌

[1] Jain, A. K. and Dubes, R. C., "Algorithms for Clustering Data". Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. 1988.

[2] Wonhee Lee, "Improved Method of Initial Seeds Selection in Large-Scale Document Clustering using K-Means Algorithm", Chonbuk University doctoral thesis, 2010.

[3] Shinwon Lee, "A Study on Hierarchical Clustering using Advanced K-Means Algorithm for Information Retrieval", Chonbuk University doctoral thesis, 2005.

[4] Rafail Ostrovsky, Yuval Rabani, Leonard J.

Schulman and Chaitanya Swamy, "The Effectiveness of Lloyd-Type Methods for then k-Means Problem", Proceedings of the 47th Annual IEEE Symposium on Foundaions of Computer Science, pp.165-176, 2006.
[5] Paul Bunn, and Rafail Ostrovsky, "Secure Two-Party k-Means Clustering", Proceedings of the 14th ACM conference on Computer and communications security, Alexandria, Virginia, USA, pp.486-497, 2007.
[6] www.wikipedia.org