碩士學位論文

강화된 k-평균 군집분석에서 초기치 선정의 복잡도 개선

濟州大學校 大學院

電算統計學科

金 京 彦

2012年 6月

강화된 k-평균 군집분석에서 초기치 선정의 복잡도 개선

指導敎授 金 鐵 洙

金 京 彦

이 論文을 理學 碩士學位 論文으로 提出함

2012年 6月

金京彦의 理學 碩士學位 論文을 認准함

審査委員長 _____ 印

委 員 _____ 印

委 員 _____ 印

濟州大學校 大學院

2012年 6月

Improving Complexity of Selection Initial Point in enhanced K-Means Clustering

Kyoung-Un Kim
(Supervised by professor Chul-Soo Kim)

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science.

2012. 6.

This thesis has been examined and approved.

Thesis director, _____

Thesis director, _____

Thesis director, _____

June 2012

Department of Computer Science and Statistics

GRADUATE SCHOOL

CHEJU NATIONAL UNIVERSITY

목 차

List of Tables ---- i

List of Figures ---- ii

Abstract ---- iii

Ⅰ. 서론 ---- 1

Ⅱ. 연구배경 ---- 3

1. 데이터 마이닝과 군집분석 ---- 3

1) 데이터 마이닝 ---- 3

2) 데이터 마이닝의 필요성 ---- 5

3) 군집분석 ---- 5

4) 군집분석의 장·단점 ---- 7

5) 군집분석의 필요성 ---- 8

2. 군집분석의 요구사항 ---- 10

3. 주요 군집분석 방법 ---- 14

Ⅲ. k-means 군집분석과 강화된 k-means 군집분석 ---- 22

1. k-means 군집분석 알고리즘 ---- 22

2. 강화된 k-means 군집분석 ---- 24

1) k-means 군집분석의 강약점 ---- 24

2) 강화된 k-means 군집분석 알고리즘 ---- 25

Ⅳ. 제안하는 K-means 군집분석 ---- 27
1. 강화된 k-means 군집분석의 강약점 ---- 27
1) 강화된 k-means 군집분석의 강약점 ---- 27
2) 시간복잡도 ---- 28
2. 제안하는 k-menas 군집분석 ---- 28
1) 제안하는 k-means 군집분석 알고리즘 ---- 28
2) 시간복잡도 ---- 30

Ⅴ. 실험 결과 및 분석 ---- 31
1. 실험환경 ---- 31
2. 실험 데이터 ---- 31
1) IRIS data set ---- 31
2) Image Segmentation data set ---- 32
3. 실험 결과 ---- 33
1) IRIS data set 실험 결과 ---- 33
2) Image Segmentation data set 실험 결과 ---- 37

Ⅵ. 결론 ---- 41

Ⅶ. 참고문헌 ---- 43

List of Tables

Table 1. 구간형 데이터에 대한 거리 ---------- 6

Table 2. IRIS data set k-means 군집분석 실행 결과 ---------- 33

Table 3. IRIS data set 강화된 k-means 군집분석 실행 결과 ---------- 34

Table 4. IRIS data set 제안하는 k-means 군집분석 실행 결과 ---------- 35

Table 5. IRIS data set 통합 결과 분석표 ---------- 36

Table 6. Image Segmentation data set k-means 군집분석 실행 결과 ---------- 37

Table 7. Image Segmentation data set 강화된 k-means 군집분석 실행 결과 ---------- 38

Table 8. Image Segmentation data set 제안하는 k-means 군집분석 실행 결과 ---------- 39

Table 9. Image Segmentation data set 통합 결과 분석표 ---------- 40

List of Figures

Figure 1. KDD(Knowledge Discovery in Database)의 과정 ---------- 4

Figure 2. k-means 군집분석 ---------- 22

Figure 3. 강화된 k-means 군집분석 ---------- 25

Figure 4. 제안하는 k-means 군집분석 ---------- 29

Abstract

In situations that a lot of data overflow, data mining is attracting attention because it is to extract useful informations and patterns from data.

Clustering is an important technique in data mining. It is to group data into clusters such that the similarities among data within the same cluster are maximal while dissimilarities among data from different clusters are maximal. As active subject of research, it is finding a way that can be an effective and efficient clustering.

K-means clustering proposed by MacQueen(1967)[1] is famous and useful method of partition-based clustering. It is simple and can be used for a variety of data types.

But k-meas clustering is quite sensitive to positions of initial points. If chosen initial points is too close, it lower accuracy and increase execution time of iterative relocation.

Enhanced k-means clustering proposed by Abdul Nazeer and Sebastian(2009)[2] complemented k-means clustering's defects. Its accuracy is higher and execution time of iterative relocation is lower than k-means clustering.

But total execution time is higher than k-means clustering.

In this paper, we propose an algorithm that improves time-complexity of selection initial points in enhanced k-means clustering proposed by Abdul Nazeer and Sebastian(2009).

Ⅰ. 서론

본 논문에서는 MacQueen(1967)이 제안한 k-means 군집분석[1]을 근간으로 발전시킨 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]의 초기 평균값 선택의 시간복잡도를 향상시키는 알고리즘을 소개한다.

MacQueen(1967)이 제안한 k-means 군집분석은 데이터의 유사성을 군집의 평균값과의 거리를 척도로 군집을 형성하는 군집분석의 분할기법(partitioning methods)에 해당하는 분석방법이다.

분할기법은 분할의 향상 즉, 클러스터링의 정확도 향상을 위해 반복 재배정(iterative relocation technique)을 사용한다. k-means 군집분석에서 군집의 평균값을 갱신하여 갱신된 평균값과 객체의 거리를 비교하여 가장 가까이에 위치한 군집에 재배정하는 것이 반복 재배정에 해당한다.

MacQueen(1967)이 제안한 k-means 군집분석은 간단한 구조를 가지고, 많은 환경에서 빠른 수렴을 통해 결과를 신속히 내며, 객체들 사이의 유사성(또는 비유사성)의 정도를 수치로 잘 표현해 낼 수 있다면 거의 모든 형태의 데이터에 적용이 가능하다. 또한 변수들에 대한 역할정의가 필요 없으므로 적용이 쉽고, 탐색적인 기법으로 사전 정보 없이 데이터의 내부구조에 대해 의미 있는 자료구조를 얻을 수 있는 장점이 있다.

하지만 초기 평균값에 대한 의존도가 너무 높아 임의 추출(random sampling)된 객체들이 서로 인접해 있는 경우에는 클러스터링 정확도 저하와 여러 가지 자료유형을 포함하는 데이터에서 객체 사이의 유사성 거리 정의와 가중치 결정의 어려움이 있다. 그리고 초기 군집수 k를 설정해야 하는데 군집수 k가 적합지 않으면 좋은 클러스터링 결과를 기대하기 어렵고, 클러스터링 결과의 최적을 보장할 수 없는 단점이 있다.

강화된 k-means 군집분석은 k-means 군집분석을 토대로 단점을 보완하며 발전시켜온 것으로, 여기서는 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석을 다룬다.

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 MacQueen(1967)에 의해 제안된 k-means 군집분석[1]의 안정적이지 못한 정확도를 안정적으로 정확도를 향상시킴과 아울러 객체 재배정 과정에서의 불필요한 재배정을 제거하는 방법을 제시하여 객체 재배정에서의 실행시간을 단축시켰다.

하지만 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석 역시 초기 평균값 선택에서의 높은 시간복잡도로 인해 군집분석 전체 실행시간을 증가시켰다.

본 논문에서 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 초기 평균값 선택에서의 시간복잡도 향상을 제안하기에 앞서 k-means 군집분석이 데이터 마이닝의 다양한 분석기법 중 어떤 분석기법에 해당하는지와 해당하는 분석기법의 요구사항 및 분류에 대해 먼저 소개한다.

그리고 MacQueen(1967)이 제안한 k-means 군집분석, Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석, 제안하는 k-means 군집분석의 소개와 실험결과 및 분석, 결론으로 끝맺는다.

Ⅱ. 연구배경

1. 데이터 마이닝과 군집분석

1) 데이터 마이닝

데이터 마이닝은 “대량의 데이터로부터 유용한 지식이나 패턴을 발견하는 과정”이다. 좀 더 상세히 정의한다면 “데이터 마이닝이란 의미 있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반자동화된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정”이다. (Berry and Linoff, 1997, 2000)

그리고 데이터 마이닝의 또 다른 정의로서 가트너 그룹은 다음과 같이 정의하고 있다.(2004년 1월 가트너 그룹 웹사이트)

“데이터마이닝은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정이다.”

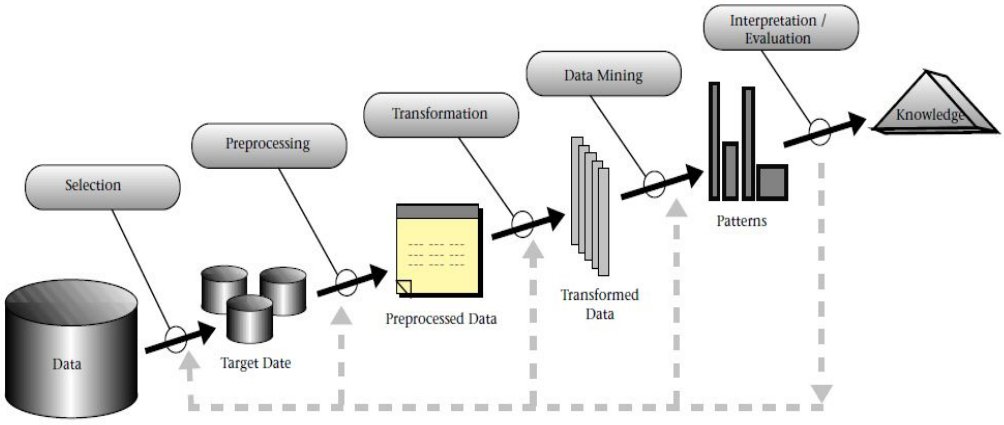
또한 1990년대 초 가트너 그룹의 Howard Dresner에 의해 만들어진 신조어로 비즈니스 인텔리전스(business intelligence)가 있다.

이는 최종사용자 질의 및 보고(end user query and reporting)를 포괄하는 의미로 경영진과 경영분석가들이 데이터를 통해 합리적 의사결정을 내릴 수 있도록 데이터를 수집, 저장, 처리, 분석하는 일련의 기술 및 응용시스템을 말하며, 데이터 웨어하우스(data warehouse), 데이터 질의 및 보고도구(data query and reporting tools), 데이터 마이닝(data mining), 비즈니스 성과관리(BPM: business performance management) 등의 요소들을 포함하고 있다.

데이터 마이닝은 비즈니스 인텔리전스(business intelligence)의 일부로서 경영자와 경영분석가들이 다양한 비즈니스 의사결정 문제를 해결해 주는 일련의 데이터 분석 과정이라고 할 수 있다.[7]

1995년 지식발견 및 데이터 마이닝(KDD : Knowledge Discovery and Data Mining) 국제학술대회가 처음 개최된 이후, 위와 같이 데이터 마이닝에 대한 정의가 다양하게 나오고 있는데, 대부분의 사람들이 데이터 마이닝을 "데이터베이스에서의 지식발견"(Knowledge Discovery in Database : KDD)과 동의어로 잘못 알고 있다.

데이터 마이닝은 KDD의 한 과정으로 <그림 1>에서 보는바와 같다.[4]



<그림 1> KDD(Knowledge Discovery in Database)의 과정
(출처: From Data mining to Knowledge Discovery in Databases, 1996)

(1) 데이터 정제(Data cleaning) : 잡음과 불일치 데이터 제거

(2) 데이터 통합(Data integration) : 다수의 데이터 소스들의 결합

(3) 데이터 선택(Data selection) : 분석 작업과 관련된 데이터들이 데이터베이스로부터 검색

(4) 데이터 변환(Data transformation) : 데이터 마이닝을 위해 적합한 형태로 데이터를 변환하거나 합병정리

(5) 데이터 마이닝(Data mining) : 데이터 패턴을 추출하기 위해 지능적 방법들이 적용되는 필수적 과정

(6) 패턴 평가(Pattern evaluation) : 의미 있는 패턴 식별

(7) 지식 표현(Knowledge presentation) : 의미 있다고 판별된 패턴 표현

데이터 마이닝은 평가를 위해 숨겨진 패턴을 찾아내는 필수적인 단계이면서도 전체 과정에서 보면 하나의 단계에 불과하다.

2) 데이터 마이닝의 필요성

데이터의 크기와 형태, 데이터에 존재하는 패턴의 유형, 데이터 잡음의 정도, 그리고 데이터에 따른 특수한 분석목적 등 다양한 요인과 요구사항, 그리고 비즈니스 인텔리전스(business intelligence)의 의사결정문제를 해결하는데 있어 데이터에서 의미 있는 패턴, 상관관계, 추세 등의 정보를 얻는 데이터 마이닝의 필요성은 더욱 증가되고 있다.

그리고 데이터 저장과 검색의 비용이 지속적으로 줄어들고 대용량의 데이터를 저장하고 생성하는데 필요한 설비를 구축하는 것이 가능하게 되었다. 즉 하드웨어의 발전으로 지속적인 연산능력 향상을 가져왔고 이는 데이터 마이닝을 더욱 발전시키는 원동력이 되었다.

또한 데이터 발생의 엄청난 증가 자체가 데이터 마이닝의 필요성을 더욱 증대시키고 있는데, 데이터가 증가한 이유는 단순히 경제나 지식베이스가 확장되어서만이 아니라 데이터를 자동적으로 얻는 데 소용되는 시간과 비용이 절감되고 이에 대한 가용성이 증가했기 때문이다. 즉 보다 많은 사건들이 기록될 뿐만 아니라 각 사건에 보다 많은 정보들이 수집되고 있기 때문이다.[7]

3) 군집분석

군집분석은 대량의 데이터를 처리하는 데이터 마이닝의 주요 분석 기법 중 하나로 특별한 가정 없이 객체들 사이의 유사성 또는 거리(비유사성)에 근거하여 자연스러운 군집을 찾고 분석을 하는 탐색적인 통계분석 기법이다.

즉 주어진 데이터에서 유사한 객체들을 몇몇의 군집으로 그룹화 하여, 각 군집의 성격을 파악함으로서 데이터 전체의 구조에 대한 이해를 쉽게 한다.

특히 대용량 데이터에 대해서는 개개의 객체를 요약하는 것보다는 전체를 유사한 객체들의 군집(cluster)으로 구분하여, 복잡한 전체보다는 그들을 대표할 수

있는 군집들을 관찰함으로서 전체 데이터에 대한 의미 있는 정보를 얻어낼 수 있기에 군집분석의 필요성은 더욱 증대되고 있다.[6]

(1) 거리(Distance)

거리(distance)는 비유사성(dissimilarity)의 척도로서 개체들 간의 먼 정도를 의미하며, 유클리드 거리가 가장 널리 사용된다.

그런데 이러한 거리들은 척도불변성(scale invariance)을 가지지 않으므로, 즉 객체들의 측정단위에 의존하므로 사용에 주의가 요구된다. 따라서 각 변수의 표준편차나 범위 등으로 나누어 측정단위를 없애는 표준화(standardized)를 고려하는 것이 더 바람직한 경우가 많다.

범주형 변수의 경우에는 거리라는 개념의 적용이 쉽지 않으므로 두 객체가 서로 다른 범주에 속한 회수를 이용하기도 한다. 즉 객체의 속성 값의 불일치 수를 이용하여 거리를 표현하기도 한다. 그리고 변수들이 범주형과 연속형이 혼합되어 있는 경우는 위와 같은 방법은 유용하지 않는다.[6]

다음은 구간형 데이터에 대한 거리이다.

거 리	공 식
유클리드 거리(Euclidean distance)	$\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$
유클리드 제곱거리(Squared Euclidean distance)	$\sum_{k=1}^{p}(x_{ik}-x_{jk})^2$
시티블록 거리(City Block distance)	$\sum_{k=1}^{p}\|x_{ik}-x_{jk}\|$
민코브스키 거리(L_m(Minkowski distance))	$[\sum_{k=1}^{p}\|x_{ik}-x_{jk}\|^m]^{\frac{1}{m}}$
체비세프 거리(Chebychev distance)	$\max_k\|x_{ik}-x_{jk}\|$

<표 1 > 구간형 데이터에 대한 거리

(출처:“고객관계관리를 위한 데이터마이닝 방법론“, 자유아카데미)

(2) 군집의 유형

■ 상호배반적(Disjoint) 군집

각 관찰치가 상호배반적인 여러 군집 중 오직 하나에만 속한다.

■ 계보적(Hierarchical) 군집

한 군집이 다른 군집의 내부에 포함되는 형태로 군집간의 중복은 없으며 군집들이 단계마다 계층적인 구조를 이룬다.

■ 중복(Overlapping) 군집

두 개 이상의 군집에 한 관찰치가 동시에 소속되는 것을 허용한다.

■ 퍼지(Fuzzy) 군집

관찰치가 소속되는 특정한 군집을 표현하는 것이 아니라 각 군집에 속할 가능성을 표현한다.[6]

4) 군집분석의 장·단점

(1) 군집분석의 장점

■ 탐색적인 기법

군집분석은 대용량 데이터에 대한 탐색적인 기법으로, 주어진 데이터의 내부구조에 대한 사전 정보 없이 의미 있는 자료구조를 찾아낼 수 있는 방법이다. 따라서 회귀분석, 의사결정나무분석, 신경망분석 등의 여러 가지 모형화를 위한 분석의 사전 기법으로 사용될 수 있다.

■ 다양한 형태의 데이터에 적용가능

분석을 위해 기본적으로 객체들 간의 거리를 데이터의 형태에 맞게 정의하면 거의 모든 형태의 데이터에 대하여 적용이 가능하다. 즉 객체들 사이의 비

유사성(또는 유사성)의 정도를 거리로 표현할 수 있다면 군집분석이 가능하다.

■ 분석방법의 적용 용이성

데이터에 대해 사전 정보를 거의 요구하지 않으므로 모형화를 위한 분석과 같이 사전에 특정 변수들에 대한 역할정의가 필요하지 않고, 객체들 사이의 거리만이 분석에 필요한 입력 자료가 된다.[6]

(2) 군집분석의 단점

■ 가중치와 거리의 정의

객체들 사이의 비유사성인 거리 또는 유사성을 어떻게 정의하는가에 따라 군집분석의 결과는 크게 좌우된다. 따라서 연속형, 범주형 등의 여러 가지 자료유형을 포함하는 데이터의 경우, 객체들 사이의 거리를 정의하고 각 변수에 대한 가중치를 결정하는 것은 매우 어려운 문제이다.

■ 초기 군집수의 설정

k-평균 군집분석에서는 사전에 정의된 군집수를 기준으로 동일한 수의 군집을 찾게 되므로, 군집수가 데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없다. 그러므로 군집분석의 절차는 여러 차례의 반복이 이루어져야 한다.

■ 결과해석의 어려움

사전 정보 없이 분석이 이루어지는 탐색적인 분석방법으로의 장점을 가지고 있지만, 사전에 주어진 목적이 없으므로 결과를 해석하는 데 있어서 어려움이 있다. 따라서 주어진 변수에 따라 잘 구분된 군집이라 하여도 그 결과를 충분히 이해하고 실제적으로 활용하기가 쉽지 않은 경우가 종종 있다.[6]

5) 군집분석의 필요성

대용량의 데이터에서 유용한 정보나 패턴을 알아보는데 있어 유사한 객체들을

동일한 군집으로 분류하는 일은 매우 보편적이며 필수적인 과정이다.

군집분석은 비교사 학습(Unsupervised Learning) 알고리즘으로 사전에 데이터가 명확하게 정의되어 있지 않은 상태에서 객체들이 갖고 있는 속성들의 값 사이의 유사성을 근거로 하여 군집으로 분류하기 위한 방법이다.

따라서 사전에 데이터에 대해 정의되어 있지 않은 경우는 우선 군집분석을 사용해야 한다.

데이터를 군집화 시킴으로서 데이터 전체가 아닌 데이터를 대표할 수 있는 군집들을 분석함으로서 전체 데이터에 대한 형태와 정보를 알아내기도 하지만, 군집 밖에 멀리 떨어져 존재하는 객체에 대해서는 이상치(outlier) 판별에도 적용할 수 있다. 또한 분류나 예측을 위한 선행 작업으로도 활용된다.

2. 군집분석의 요구사항

군집분석은 데이터를 데이터가 갖고 있는 유사성을 기반으로 묶어주는 과정을 말한다. 이 때, 동일한 군집에 속한 데이터들은 서로 유사성(similarity)이 매우 커야하며, 외부 군집과는 상이성(dissimilarity)이 커야 한다.

군집분석은 패턴인식, 데이터 분석, 이미지 처리, 그리고 시장조사를 포함한 매우 많은 응용에서 넓게 사용된다. 그 예로 판매자가 고객 내에서 특정한 그룹을 찾고, 구매 패턴에 기초하여 고객 그룹의 특성화, 생물과 동물의 분류법을 이끌어내고 비슷한 기능 군끼리 유전자를 분류하고 집단의 고유 특성에 대한 정보 획득, 토양에 관한 데이터베이스를 사용하여 비슷한 용도의 땅을 식별하거나 도시에서 집의 유형, 가치, 지리적인 위치에 따라 집의 그룹 식별, 보험 가입자 그룹에서 평균적으로 높은 배상비용을 가진 이들의 그룹을 구별, 통신사에서 충성 고객과 이탈고객 그룹의 구별, 웹상에서 정보탐색을 위한 문서 분류 등을 돕는데 이용된다.

데이터 마이닝 기술로서 군집분석은 데이터의 분포에 대한 지식을 얻고, 각각의 군집의 특징을 관찰하고, 추가적인 분석을 위해 특정 군집 집합에 초점을 맞추기 위한 독립적인 도구로서 사용된다.

대안적으로는, 찾아낸 군집들에 적용할 수 있는 특성화, 분류화 같은 다른 응용 분석기법을 위한 전처리 단계로 이용될 수 있다.

대용량 데이터베이스에서 효과적이고 효율적인 군집분석을 할 수 있는 방법을 찾고 있는 요즘, 연구의 활발한 주제는 군집분석 방법의 확장성, 복잡한 형태 및 데이터 유형의 군집분석 방법의 효율성, 고차원의 군집분석 기술, 그리고 대용량 데이터베이스 내에서 수리적인 자료와 범주형 자료가 혼합된 경우 군집분석방법 등에 초점을 맞춘다.

다음은 데이터 마이닝에 있어서의 군집석의 대표적인 요구사항들이다.[4]

1) 확장성(Scalability)

많은 군집분석 알고리즘들은 데이터 객체가 200개보다 작은 데이터집합에서 더 잘 작업된다. 그런데 대용량 데이터베이스는 수백만의 데이터 객체를 가지고 있다. 큰 데이터 집합의 표본에서의 군집분석은 편향된 결과를 초래할 수 있다. 높은 확장성을 가진 군집분석 알고리즘이 요구된다.

2) 속성이 다른 타입을 다루는 능력

많은 군집분석 알고리즘은 구간에 기초한 수치형 데이터에 대해 군집이 설계된다. 그런데 응용은 이항형이나 범주형 그리고 순서형 데이터 혹은 이런 데이터 타입이 혼합되어 있는 데이터 타입이 혼합되어 있는 데이터에서의 군집분석도 요구한다.

3) 임의의 형태에서 군집 발견

많은 군집분석 알고리즘은 유클리드 혹은 맨하탄 거리 척도에 기초하여 군집을 결정한다. 이러한 거리척도에 기초한 알고리즘은 비슷한 크기나 밀도를 갖는 구형의 군집을 찾는 경향이 있다. 임의의 형태의 군집을 찾을 수 있는 알고리즘을 개발하는 것이 매우 중요하다.

4) 입력 인자를 결정하기 위한 도메인에 관한 지식의 최소 요구사항

많은 군집분석 알고리즘은 사용자가 군집분석을 위해 (적절한 군집의 개수와 같은) 확실한 인자를 넣어야 한다. 군집분석 결과는 입력된 인자에 매우 민감할 수 있다.

인자는 특히 고차원의 객체를 포함한 데이터 집합에서 가끔 결정하기 어렵다. 이는 사용자에게 짐이 될 뿐만 아니라 군집분석의 질을 통제하기 어렵게 만들기도 한다.

5) 잡음 데이터를 다루는 능력

대부분의 실제 세계의 데이터베이스는 이상치나 결측치, 알려지지 않거나 오류인 데이터들을 가지고 있다. 몇몇 군집분석 알고리즘은 이러한 데이터에 민감하며 군집의 질을 낮출 수 있다.

6) 입력 레코드 순서에 둔감함

몇몇 군집분석 알고리즘은 입력된 데이터 순서에 민감하다. 예를 들어 그런 알고리즘에 다른 순서로 제시되면, 같은 데이터 집합이라도 극단적으로 다른 군집들을 생성할 수 있다. 입력의 순서에 둔감한 알고리즘의 개발이 매우 중요하다.

7) 고차원

데이터베이스나 데이터 웨어하우스는 여러 차원 혹은 속성들을 가지고 있을 수 있다. 많은 군집분석 알고리즘은 2차 혹은 3차원의 낮은 차원의 데이터를 다루는 데는 능숙하다. 사람의 눈은 3차원 이하 군집분석의 질을 판단하는데 좋다. 고차원의 공간에서 데이터 객체를 군집분석하는 것은, 특히 강한 희박성과 강한 편중성을 가진 경우를 고려하면 더욱 어렵다.

8) 제약기반 군집화

실제세계의 응용에선 다양한 종류의 제약 하에서 군집분석을 수행할 필요가 있다. 만약 어느 도시에 여러 개의 새 자동현금지급기(ATM)의 위치를 골라야 할 때, 이것을 결정하기 위해 강이나 고속도로망, 그리고 지역마다 고객의 요구 등과 같은 제약조건을 고려하여 세대를 나누게 된다.

요구되는 과업은 특정한 제약조건을 만족하는 좋은 군집화 행동을 갖는 데이터 그룹을 찾는 것이다.

9) 상호작동성과 유용성

사용자는 군집분석 결과가 설명하기 쉽고 이해하기 쉬우며 사용하기 쉽게 되는 것을 기대한다. 즉, 군집화는 특별한 의미의 해석과 응용으로 엮어있다.

응용의 목표가 군집화 방법의 선택에 어떻게 영향을 주는지 연구하는 것이 매우 중요하다.

3. 주요 군집분석 방법

군집화 알고리즘의 선택은 유용한 데이터 타입과 특정한 목표와 응용에 따라 달라진다. 군집분석이 서술 혹은 탐색 도구로 사용되면 데이터가 무엇을 찾아내는지 보기 위해 같은 데이터에 대해 여러 알고리즘을 시도해 볼 수 있다. 군집분석 방법은 다음과 같이 분류한다.[4]

1) 분할 기법(partitioning methods)

n개의 객체 혹은 튜플(tuple)이 주워졌을 때, k ≤ n 이 되도록 군집을 나타내는 데이터 분할을 k개로 만든다.

(1) 각 그룹은 적어도 하나의 객체를 가지고 있어야 한다.
(2) 각 객체는 정확히 하나의 그룹에 속해야한다.

위 두 가지 조건을 만족하도록 k개의 그룹으로 분할하는 기법이다.
그리고 분할을 향상시키기 위한 시도로 객체를 하나의 그룹에서 다른 곳으로 이동시키는 반복적인 재배정기법(iterative relocation technique)을 사용한다.
분할에 기초한 군집화에서 광범위한 최적성을 얻기 위해서는 모든 가능한 분할의 완전한 계산이 요구된다. 대신에 대부분의 응용은 두 개의 유명한 경험적 기법 중 하나를 사용한다.

(1) 각 군집이 군집의 평균값으로 대표되는 k-means 군집분석
(2) 각 군집이 군집의 중앙값으로 대표되는 k-medoids 군집분석

이 두 군집분석은 구(spherical) 형태의 군집을 찾는데 더 효율적이다.[4]

2) 계층적 기법(hierarchical method)

계층적 기법은 데이터 객체들의 군집들을 트리구조로 그룹화 하는 것을 말한다. 계층적 분류기법은 계층적 분해가 상향식(bottom-up) 혹은 하향식(top-bottom)으로 형성되는가에 따라 집괴적(Agglomerative)인 것과 분할적(Divisive)인 계층 군집화로 분류된다.

(1) 집괴적(Agglomerative) 계층 군집화

각 객체를 상향식(bottom-up)으로 자신의 군집에 배치하고, 그 원자 군집들을 더 큰 군집으로 만들어간다. 모든 객체가 하나의 군집을 구성하거나 어떤 종료의 조건이 만족되면 종료한다. 대부분의 계층 군집화 기법은 군집 내부의 유사도에 대한 정의만 다를 뿐 이 분류에 속한다.

(2) 분할적(Divisive) 계층 군집화

하향식(top-bottom)으로 모든 객체를 하나의 군집으로 여기면서 시작한다. 각 객체가 하나의 군집을 형성하거나 적절한 개수의 군집이 얻어질 때까지, 혹은 가까운 군집들 간의 거리가 어떤 한계를 넘어서는 것과 같은 특정한 종료 조건이 만족될 때까지 군집을 작은 조각으로 세분화한다.

계층적 군집화 기법은 단순하지만 합병과 분할점의 선택에 관련하여 어려움을 겪는다. 한번 결정이 되면 객체의 그룹이 합쳐지거나 분할되면 다음 단계의 과정은 해로 생성된 군집 하에서 이루어지며, 이전 단계로 되돌리거나 군집 간에 객체를 바꾸지도 못한다는 점이다. 따라서 합병과 분할에 관한 결정은 어떤 단계에서 제대로 이루어지지 않으면 군집의 질이 낮다.

게다가 합병과 분할의 결정시, 객체나 군집의 적절한 개수를 평가하고 점검해야 할 필요가 있기 때문에 확장적용이 쉽지 않다. 기법의 군집화 능력을 향상시키기 위해 다단계 군집화를 위한 다른 군집화 기법과 통합하는 것이다. 최근 연구는 계층적 집괴와 반복 재배정 기법의 통합을 강조하고 있다.[4]

3) 밀도기반 기법(density-based method)

대부분의 분할 기법은 객체간의 거리에 기초하여 객체들을 군집한다. 그러한 기법들은 구형(spherical)의 군집만을 찾을 수 있고, 임의의 형태의 군집을 찾는 데는 어려움이 있다.

밀도기반 기법은 "근처(neighborhood)"의 밀도가 어떤 한계점을 능할 만큼 주어진 군집이 커지도록 계속하는 기법이다. 즉 주어진 군집 내에서 각 데이터 포인트가 주어진 반경 근처에 최소한의 개수만큼은 갖도록 한다. 대표적인 기법으로 DBSCAN, OPTICS가 있다.

(1) DBSCAN : 밀도 기반 군집화 기법

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)은 밀도에 기초한 군집분석 기법이다. 충분하게 밀도가 높은 지역은 군집으로 키우고, 잡음값을 가진 공간적 데이터베이스에서 임의의 형태인 군집을 찾는 알고리즘으로 군집을 밀도 연결점의 최대화 집합으로 정의한다.[4]

(2) OPTICS : 군집 구조식별을 위한 순서화

DBSCAN 및 다른 군집분석 알고리즘은 사용자에게 군집의 발견을 이끌어낼 수 있는 인자 값을 선택해야하는 책임이 있다. 인자의 설정은 경험적으로 이루어지며 결정하기 힘든데, 실제 세계, 고차원의 데이터 집합에서는 더욱 그렇다.

대부분의 알고리즘은 인자 값에 매우 민감하여 아주 작은 차이의 설정이 아주 다른 데이터의 군집화를 이끌어 낼 수 있다. 게다가 고차원의 실제 데이터 집합이 전역적인 밀도 인자 수들에 의해 그들의 본질적인 군집화 구조가 묘사되지 않은, 매우 치우친 분포를 갖는다.

이러한 어려움을 극복하기 위해 OPTICS 군집분석 기법이 제안되었다. 명시적으로 데이터 집합 군집화를 만들어내는 대신에 자동적이고 상호작용하는 군집분석을 위해 점진적 군집 순서(clustering ordering)를 계산한다.

이 순서는 데이터의 밀도 기반 군집화 구조를 나타내며, 광범위한 인자 설정으로부터 얻어진 밀도 기반 군집화에 해당하는 정보를 가지고 있다.[4]

4) 격자기반 기법(grid-based method)

격자기반 군집화 기법은 다해상도 격자 데이터 구조를 이용한다. 이 기법은 공간을 유한개의 셀들로 양자화 한다. 이 셀들은 격자 구조를 갖게 되고 군집화를 위한 모든 작업이 실행되는데, 이러한 접근의 주된 장점은 빠른 진행시간이다. 전형적으로 데이터 객체의 수에 무관하고, 단지 양자화된 공간인 각 차원의 셀의 수에 의존한다. 대표적인 기법으로 격자 셀에 저장되어 있는 통계적 정보를 탐지하는 STING, 웨이블릿(wavelet) 변환 기법을 사용하여 객체들을 군집화하는 WaveCluster 등이 있다.

(1) STING : 통계정보 격자(STatistical INformation Grid)

STING은 공간적 영역을 사각형의 셀로 나누는 격자기반 다해상도 군집분석 기법이다. 해상도의 단계에 따라 사각형 셀의 여러 수준이 있고, 그 셀들은 계층적 구조를 형성한다. 높은 수준의 각 셀은 보다 낮은 수준에서 몇 개의 셀을 형성하기 위해 분할된다. 각 격자 셀에 있는 속성들에 대한 평균, 최댓값, 최솟값 등의 통계적 정보는 미리 계산되고 저장된다.

이 통계적 인자는 하향식 격자기반 기법에 사용되며, 계층 구조 내의 층(layer)은 질의 응답과정이 시작되면서부터 결정된다. 이 층은 보통 작은 수의 셀을 포함하는데 현재 층의 각 셀에 대해 주어진 질의에 대한 셀의 적절성을 반영한 신뢰구간을 구한다.

적절하지 않은 셀은 다음번 고려할 때 제거되고 보다 낮은 수준의 공정에서는 남아있는 타당한 셀들만 점검한다. 이 공정은 가장 낮은 층에 도달할 때까지 반복된다. 그때, 질의 내용이 맞으면 질의를 만족하는 타당한 셀들의 영역이 답이 된다. 그렇지 않으면 타당한 셀에 떨어진 데이터는 제거되고, 질의의 요구사항과 맞을 때까지 계속 처리한다.

STING은 각 셀에 저장된 통계적 정보가 질의와 관계없이 격자 셀에 있는 데이터의 요약 정보를 나타내므로, 격자기반 계산은 질의 독립적(query-independent)이고, 격자 구조는 병렬 처리와 점진적 수정을 수월하게 한다. 그리고 셀들의 인자를 계산하기 우해 데이터베이스를 한번 거치므로 군집 생성의 시간복잡도는

$O(n)$ (n:데이터의 수)이고, 계층적 구조를 생성한 후 질의 공정 시간은 $O(g)$ (g: 가장 낮은 수준 격자 셀의 총 수)로 보통 n보다 현저하게 작으므로 효율성이 좋다는 장점이 있다.

STING은 군집분석 수행을 위해 다해상도 접근을 사용하므로 격자 구조의 가장 낮은 수준의 과립크기(granularity)에 의존한다. 과립크기가 매우 작으면 공정의 비용은 증가하고, 과립크기가 커지게 된다면 군집분석의 성능을 감소시키게 된다. 그리고 부모 셀을 구성할 때 자식과 이웃 셀들 간의 공간관계를 고려하지 않아 결과로 나온 군집의 형태는 모든 군집 경계가 수평적이거나 수직적이어서 대각선의 경계가 없는 단순띠(isothetic)이다. 이것은 기법의 빠른 공정시간에도 불구하고 군집의 정확성과 성능을 낮추게 되는 단점이 있다.[4]

(2) WaveCluster : 웨이블릿 변환을 이용한 군집화

WaveCluster는 데이터를 데이터 공간의 다차원 격자 구조에 배치한 후 요약하는 다해상도 군집분석 알고리즘이다. 그리고 원래의 특성 공간을 변환하기 위하여 웨이블릿 변환을 사용하여 변환된 공간에서 조밀한 영역을 찾는다.

웨이블릿 변환(wavelet transformation)은 신호를 다른 빈도 부분영역으로 분해하는 신호처리 기술이다. 웨이블릿 모델은 1차원의 웨이블릿 변환을 n번 적용함으로써 n차원 신호에 적용될 수 있다. 웨이블릿 변환 적용에서 다른 해상도의 수준에서 객체 간 상대적 거리가 보존되도록 데이터가 변환되는데 데이터의 자연스러운 군집 식별이 가능하도록 한다. 그 다음 군집들은 새로운 영역에서 조밀한 영역을 찾음으로서 식별된다.

웨이블릿 변환의 장점은 첫째로 무감독 군집화를 한다. 모자 모양 필터를 사용하여 점들이 군집되는 영역은 강조하고, 군집 경계 밖의 약한 정보는 억제한다. 따라서 본래의 특성 공간에서 조밀한 영역은 가까이 있는 점에 대해서는 흡입자로, 멀리 떨어져 있는 점에 대해서는 은폐자로 작용한다. 즉 데이터의 군집은 자동적으로 주위에 있는 영역을 나타내고 군집영역 이외의 부분은 제거한다는 의미로 이상치의 제거를 가져올 수 있다.

둘째로 웨이블릿 변환의 다해상도 성질은 다양한 정확도로 군집들을 찾는데 도움을 준다.

셋째로 웨이블릿 기반 군집분석의 시간복잡도는 $O(n)$ (n:데이터의 수)로 매우 빠르다. 그리고 알고리즘은 병렬로 구현이 가능하다.

WaveCluster는 격자기반이며 밀도기반 군집분석 기법으로 효과적으로 대용량 데이터를 다루고, 일정하지 않은 모양의 군집들을 발견하며, 이상치를 잘 다룬다. 그리고 입력의 순서에 둔감하고, 군집의 수 혹은 이웃의 반경 같은 입력 인자에 대한 명세를 필요로 하지 않는다. 또 20차원까지 데이터를 다룰 수 있다.[4]

5) 모델기반 기법(model-based method)

모델기반 군집화 기법은 주어진 데이터와 수학적 모델 사이에서 최적화 적합을 시도하는 방법으로 데이터에 내재하는 확률 분포의 혼합에 의해 생성된다는 가정에 기초한다. 모델기반 군집화 기법은 통계적 접근(statistical approach)이나 신경망(neural network) 접근방식이 있다.

(1) 통계적 접근(Statistical Approach)

개념적 군집화(Conceptual Clustering)는 기계 학습 군집화의 형태로 이름이 없는 객체의 집단이 주어졌을 때 객체들의 분류 구조를 만든다. 비슷한 객체의 그룹을 인식하는 보통의 군집화와는 달리, 개념 군집화는 개념이나 클래스를 나타내는 각 그룹에 대해 특징적 설명을 찾는다. 그러므로 개념적 군집화는 2단계 처리를 거친다. 군집화가 먼저 수행된 후 특성화가 뒤따른다. 군집화의 질을 결정하는 것은 개개의 객체들뿐만 아니라 유도된 개념 설명의 보편성과 단순성과 같은 요소들을 포함한다.

COBWEB은 단순하고 인기있는 점진적 개념적 군집분석 기법으로 입력 객체들은 범주적 속성 값의 쌍으로 표현되어진다. 그리고 분류 트리(classification tree) 모양의 계층적 군집화를 생성한다.

분류트리는 의사결정트리와는 다른 것으로 각 노드는 개념을 나타내고 노드 아래에 분류된 객체들을 요약하는 개념의 명세를 포함한다. 임의의 분류트리 레벨에서 형제 노드들은 분할(partition)이라 부르며, 분류 트리를 사용하여 객체들을 분류하기 위해 부분적 매칭 함수가 가장 좋은 매칭 노드의 경로를 따라 내려

가는 방법으로 사용된다.

COBWEB은 트리의 구성 범주 효용값이라 부르는 경험적 평가 측도를 사용하는데, 범주 효용값(Category utility, CU)은 다음과 같다.[4]

$$\frac{\sum_{k=1}^{n} P(C_k)[\sum_i \sum_j P(A_i = V_{ij}|C_k)^2 = \sum_i \sum_j P(A_i = V_{ij})^2]}{n}$$

$A_i = V_{ij}$: 속성-값의 쌍, C_k : 개념 클래스

(2) 신경망 접근(Neural Network Approach)

군집분석에 대한 신경만 접근은 각 군집을 견본(exemplar)으로 표현하는 경향이 있다. 견본은 군집의 기본형(prototype)으로 작용하고 특정 데이터 견본이나 객체에 대해 해당하지 않아도 된다. 새로운 객체들은 척도에 기반을 두어 그 견본이 가장 일치하는 군집에 분배 되는데 군집에 할당된 객체의 속성들은 군집 견본의 속성들로부터 예측될 수 있다.

신경망 접근에는 두 가지 유망한 방법으로 첫째는 경쟁 학습(competitive learning)이고, 두 번째는 자가 구성 특성지도(self-organizing feature maps)로 신경단위를 포함한다.

경쟁 학습(competitive learning)은 객체에 대한 승자완승(winner-takes-all)에서 경쟁하는 여러 단위영역들 또는 인공적인 뉴런(neurons)의 계층적 구조를 말한다. 주어진 층에서 단위영역은 다음 하위 레벨 안에서 모든 단위영역들로부터 입력을 받을 수 있으며, 활동적인 단위영역의 구성은 다음 상위 레벨의 입력 패턴을 나타낸다.

층 안의 연결들은 단지 주어진 군집 안에서만 활동적일 수 있기 위해 은둔적(inhibitory)이다. 선택된 단위영역(winning unit)은 군집 내 다른 군집과 연결시 현재와 비슷하거나 향후 객체들에 대해 더 강하게 응답할 수 있기 위해서 가중치를 수정한다.

군집분석의 마지막에 각 군집은 객체들에서 어떤 일관성을 탐지하여 새로운 특징(feature)으로 여길 수 있기에 최종 군집들은 하위 레벨 특성을 상위 레벨 특성으로 대응한 것으로 볼 수 있다.

자가 구성 특성 지도(Self-organizing feature maps, SOMs)에 의한 군집분석은 객체에 대해 경쟁하는 몇 개의 단위영역들을 가짐으로서 수행된다. 가중치 벡터가 현재 객체에 가장 가까운 단위영역은 선택된 혹은 활동 단위영역이 되는데, 입력 객체에 더 가까이에 가기 위해서 가장 근접한 이웃들의 가중치 뿐 아니라 선택된 단위영역의 가중치도 수정된다.

SOMs는 입력 객체에 어떤 위상이나 순서가 있고, 단위영역들은 공간에서 어떤 구조를 갖는다고 가정하며, 단위영역들의 구조는 특성 지도를 형성하게 된다. SOMs는 뇌에서 발생하는 처리와 유사하며, 2-D 또는 3-D 공간에서 고차원 데이터를 가시화 하는데 유용하다.[4]

Ⅲ. k-means 군집분석과 강화된 k-means 군집분석

일반적으로 가장 잘 알려지고 사용되는 분할기법은 MacQueen(1967)이 제안한 k-means 군집분석[1]이다. k-means 군집분석과 k-means 군집분석의 정확도 향상과 객체 재배정의 시간을 단축시킨 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]에 대해 알아보겠다.

1. k-means 군집분석 알고리즘

MacQueen(1967)이 제안한 k-means 군집분석[1]은 입력 값으로 k를 취하고, 군집 내 유사성은 높고 군집끼리 유사성은 낮게 되도록 n개 객체들의 집합을 k개의 군집으로 분해한다. 군집의 유사성은 군집의 무게중심으로 볼 수 있는 객체들의 평균값을 측정한다.

MacQueen(1967)이 제안한 k-means 군집분석은 다음과 같다.[4]

Algorithm : k-means
Input : 클러스터의 개수 k, n 개의 객체를 포함하고 있는 데이터베이스
Output : 제곱 오차 기준(square-error criterion)을 최소로 하는 k개의 클러스터
Method :
(1) k개의 객체를 임의로 선택해서 클러스터의 평균값(mean)으로 놓는다.
(2) 반복
(3) 각각의 객체를 k개의 클러스터 평균값(mean)과 비교하여
가장 가까이에 위치한 클러스터에 해당 객체를 (재)할당한다;
(4) 클러스터의 평균값을 갱신한다.
(5) 클러스터에 변화가 생기지 않을 때까지;

<그림 1> k-means 군집분석

MacQueen(1967)이 제안한 k-means 군집분석[1]은 다음과 같이 진행된다.

군집 평균이나 중심을 나타내는 값으로 객체들에서 k개를 임의적으로 선택한다. 남겨진 객체들은 객체와 군집 평균에 기초하여 가장 유사한 군집에 할당된다. 그리고 각 군집에 새로운 평균을 구한다. 이 과정을 기준함수가 수렴할 때까지 반복한다. 일반적으로 제곱오차(squared-error) 기준이 사용되며 다음과 같이 정의된다.[4]

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2$$

E는 데이터베이스에서 모든 객체들의 제곱오차를 합한 것이고, p는 주어진 객체를 나타내는 공간의 점이고, m_i는 군집, C_i의 평균이다.(p와 m_i는 다차원).

2. 강화된 k-means 군집분석

1) k-means 군집분석의 강약점

MacQueen(1967)이 제안한 k-means 군집분석[1]은 간단한 구조를 가지고 있고 많은 환경에서 빠르게 수렴한다. 그리고 비교사 학습(Unsupervised Learning) 알고리즘으로 사전 정보 없이 데이터의 내부 구조에 대해 의미 있는 자료구조를 얻을 수 있다. 또한 변수들에 대한 역할 정의가 필요 없기에 적용이 쉽고 객체들 사이의 유사성(또는 비유사성)의 정도를 수치로 잘 표현해 낼 수 있다면 다양한 형태의 데이터에 적용이 가능한 장점이 있다.

하지만 클러스터링 결과의 정확성은 보장할 수 없다. 즉 군집간의 거리는 가능한 멀고 군집 내 객체간의 거리는 가능한 작아야 하는데 임의로 선택한다면 바로 인접한 곳에 초기 클러스터링 평균값이 연이어 선정될 수 있다. 그렇게 된다면 클러스터링 결과의 정확도가 현저하게 떨어진다.

그리고 속성들의 형태가 다르거나 같은 형태의 속성이라도 속성 값의 범위가 다양할 경우 유사성에 대한 거리를 계산하는 기준과 가중치 결정에 대한 어려움이 따른다.

또한 초기 군집수 결정에 관한 것으로 군집수 k가 적합하지 않으면 좋은 결과를 기대하기는 어려울 뿐만 아니라 클러스터링의 결과가 최적이라는 보장을 하지 못한다.

객체 배정에 있어서는 초기 평균값의 임의 선정으로 인해 최종 클러스터링 평균값과 멀리 떨어지게 된 경우 객체 재배정의 반복 횟수가 늘어나 수행시간이 증가되는 단점이 있다.

2) 강화된 k-means 군집분석 알고리즘

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]의 초기치 선정과 객체의 배정 방법에서 기존 알고리즘과 차이를 두어, 초기치 선정의 정확도와 객체 재배정의 효율성 향상을 나타내었다.

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석은 다음과 같다.

Algorithm : 강화된 k-means

Input: n개의 객체를 포함하고 있는 데이터베이스, k : 클러스터의 개수

output: k개의 클러스터

Steps:

(1) 모든 객체간의 거리를 계산.

(2) 가장 가까운 거리를 갖는 한 쌍의 객체를 찾는다.

(3) 선택된 두 점과 가장 가까운 객체를 찾는다.
($0.75 \times n/k$ 개가 될 때까지, 소수점 이하는 올림, 선택된 객체 제외)

(4) 선택된 객체들의 평균값을 군집의 초기 평균값으로 한다.

(5) k개의 초기 평균값을 구할 때까지 (2)~(5)과정 반복

(6) 반복

(7) 평균값과 배정된 객체와의 거리를 측정한다.

(8) 갱신 후 평균값과 객체와의 거리가 전보다 커진 경우,
모든 클러스터의 평균값과 거리를 측정하여 가장 가까이에 위치한 클러스터에 (재)배정한다.

(9) 기준함수가 수렴할 때까지 반복

<그림 3> 강화된 k-means 군집분석

초기치 선정에서는 모든 객체간의 거리를 측정하고 가장 가까운 한 쌍의 객체를 선택한다. 이 두 객체와 가장 가까이 위치한 객체를 찾아 선택된 모든 객체가

$0.75\times(n/k)$[3] 개가 될 때까지 선택한 후, 선택된 객체들의 평균값을 초기 평균값으로 사용한다. 초기 평균값의 개수가 k개가 될 때까지 반복한다.(k : 클러스터의 수, n : 데이터의 수)

객체의 재배정에서는 객체와 이전 평균값과의 거리가 갱신된 평균값과의 거리보다 큰 경우의 객체에 대해서만 모든 클러스터의 평균값과 거리를 비교하여 가장 가까운 곳에 재배정한다.

기준함수는 제곱오차(squared-error) 기준이 사용되며 다음과 같이 정의된다.[4]

$$E=\sum_{i=1}^{k}\sum_{p\in c_i}|p-m_i|^2$$

E는 데이터베이스에서 모든 객체들의 제곱오차를 합한 것이고, p는 주어진 객체를 나타내는 공간의 점이고, m_i는 군집, C_i의 평균이다.(p와 m_i는 다차원).

Ⅳ. 제안하는 K-means 군집분석

1. 강화된 k-means 군집분석의 강약점

1) 강화된 k-means 군집분석의 강약점

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 객체 배정에서 갱신 전 후 평균값과의 거리를 비교하여 갱신 전보다 거리가 클 경우만 재배정을 해서 객체 배정의 실행시간을 줄이는 장점과 선택된 초기 평균값 간에 인접하지는 않아 정확도는 향상된다는 장점이 있다.

하지만 초기 클러스터링 평균값을 선택하기 위해 모든 객체간의 거리를 측정하고 시작해야 하기에, 초기 평균값 선택 실행시간이 MacQueen(1967)이 제안한 k-means 군집분석[1]보다 상당히 늘어나 알고리즘 전체 실행시간이 증가하게 된다.

그리고 군집들이 서로 인접해 있는 경우 객체가 소속된 군집의 갱신 전 평균값과의 거리보다 갱신 후의 평균값과의 거리가 늘어나야만 재배정이 되는 조건으로 인해, 재배정 조건은 해당하지 않으며 소속되어 있는 군집의 평균값과의 거리보다 인접한 군집의 갱신된 평균값의 거리가 더 가까워지게 된다면 재배정이 이루어져야할 객체들이 원래 소속되어야 할 군집에 포함되지 못하고 기존에 소속된 군집에 포함되어 정확도가 떨어지는 된다.

또한 데이터가 구(spherical) 형태의 군집을 이루지 않는 경우와 이상치(outlier)가 군집 밖 멀리 한 쌍으로 가까이 존재하는 경우에도 군집분석 결과의 정확도가 저하되는 단점이 있다.

2) 시간복잡도

MacQueen(1967)이 제안한 k-means 군집분석[1]은 초기 평균값의 임의 선정으로 인해 최종 클러스터링에 도달하는데 있어 객체 반복 재배정 횟수가 늘어나 $O(kln)$ (k:클러스터의 수, l:반복시행 수, n:데이터의 수, $k \leq n$) 의 시간복잡도로 많은 시간이 걸린다.

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 모든 객체간의 거리를 측정하여 가장 가까이에 위치한 두 점을 기준으로 평균값을 잡아가기에 초기 평균값의 최종 클러스터링의 평균값에 가까이 위치하게 만들고, 객체 배정에 있어 객체와 평균값 갱신 전·후 거리를 비교하여 갱신 전보다 거리가 늘어난 경우에만 배정하기에 객체 배정의 반복 횟수를 줄여 최대 $O(kn)$ (k:클러스터의 수, n:데이터의 수, $k \leq n$)의 시간복잡도를 갖는 장점이 있다.

하지만 초기 평균값 탐색 과정에 있어 모든 객체간의 거리를 계산해야 하기에 $O(n^2)$ (n:데이터의 수) 의 시간복잡도를 요구하는 단점을 동시에 갖고 있다.[2]

2. 제안하는 k-menas 군집분석

1) 제안하는 k-means 군집분석 알고리즘

제안하는 k-means 군집분석은 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]의 정확성은 유지하면서 실행시간을 단축할 수 있는 방법으로, MacQueen(1967)이 제안한 k-means 군집분석[1]의 빠른 초기평균값 선택의 임의추출(random sampling) 방식과 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]의 군집범위 $\alpha \times (n/k)$, $\alpha\,(0 \leq \alpha \leq 1)$ [3]와 객체배정 방식을 혼합한 k-means 군집분석을 제시한다.

제안하는 k-means 군집분석은 다음과 같다.

Algorithm : 제안하는 k-means

Input : n개의 객체를 포함하는 데이터, k : 클러스터의 수

output : k 개의 클러스터

Steps :

(1) 데이터에서 임의의 한 객체 선택한다. (선택된 객체 제외)

(2) 선택된 객체와 나머지 객체와의 거리를 구한다.

(3) 선택된 객체와 가까이에 위치한 객체를 $\alpha \times (n/k)$ 개 될 때까지 선택한다. ($\alpha\,(0 \le \alpha \le 1)$, 소수점 이하는 올림, 선택된 객체 제외)

(4) 선택된 객체들의 평균값을 군집의 초기 평균값으로 한다.

(5) k개의 클러스터가 될 때까지 (1)~(4) 과정을 반복한다.

(6) 반복

(7) 평균값과 배정된 객체와의 거리를 측정한다.

(8) 갱신 후 평균값과 객체와의 거리가 전보다 커진 경우, 모든 클러스터의 평균값과 거리를 측정하여 가장 가까이에 위치한 클러스터에 (재)배정한다.

(9) 기준함수가 수렴할 때까지 반복

<그림 4> 제안하는 k-means 군집분석

데이터에의 임의 한 객체를 선정하고, 선정된 객체와 나머지 객체와의 거리를 측정한 후, 가장 가까이에 위치한 한 객체를 $\alpha \times (n/k)$, ($\alpha\,(0 \le \alpha \le 1)$, 소수점 이하는 올림)[3] 수를 만족할 때까지 선택한다. 이렇게 선택된 객체들의 평균을 초기 클러스터의 평균값으로 하는데 클러스터의 수 k개가 될 때까지 반복한다.

그리고 객체 배정은 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]과 동일하게 갱신 전·후의 평균값과 객체의 거리를 비교하여 갱신 전보다 길이가 늘어난 경우에만 객체의 재배정을 실행한다.

기준함수는 제곱오차(squared-error) 기준이 사용되며 다음과 같이 정의된다.[4]

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2$$

E는 데이터베이스에서 모든 객체들의 제곱오차를 합한 것이고, p는 주어진 객체를 나타내는 공간의 점이고, m_i는 군집, C_i의 평균이다.(p와 m_i는 다차원).

2) 시간복잡도

제안하는 k-means 군집분석의 초기치 선정에 있어서의 시간복잡도는 데이터에서 임의의 한 객체를 선택하고, 선택된 객체와 나머지 객체와의 거리를 비교하는데 클러스터의 수 k개를 만족할 때까지 반복하므로 $O(kn)$ (k:클러스터의 수, n:데이터의 수, $k \leq n$)가 된다.

Ⅴ. 실험 결과 및 분석

본 논문에서 제시하는 k-means 군집분석의 성능 평가를 위하여 MacQueen(1967)이 제안한 k-means 군집분석[1]과 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]을 비교한 실험 및 결과를 알아본다.

1. 실험환경

구현 환경은 Microsoft Window XP Professional, Intel(R) Pentium(R) D CPU 2.8GHz, 2GB RAM이며, 프로그램은 R version 2.13.1 사용하였다.

2. 실험 데이터

본 논문에서는 UCI Machine Learning Repository[5]의 IRIS data set, Image Segmentation data set을 실험에 사용하였다.

1) IRIS data set

IRIS data set은 150개의 객체, 5개의 속성을 가지고 있으며 속성은 다음과 같다.

Sepal Length (꽃받침의 길이), Sepal Width (꽃받침의 너비), Petal Length (꽃잎의 길이), Petal Width (꽃잎의 너비), Species (꽃의 종류 - Setosa / Versicolor / Virginica 의 3종류로 구분되며 각 50개씩이다.)

2) Image Segmentation data set

Image Segmentation data set은 2100개의 객체, 20개의 속성을 가지고 있으며 속성은 다음과 같다.

class(7종류의 그림으로 구분되며 각 300개씩이다.), region-centroid-col(영역의 중심 픽셀 열), region-centroid-row(영역의 중심 픽셀 행), region-pixel-count (영역의 픽셀 수), short-line-density-5(길이가 5인 라인의 개수를 찾는 추출 알고리즘 결과), short-line-density-2(길이가 2인 라인의 개수를 찾는 추출 알고리즘 결과), vedge-mean(영역의 가로 인접 픽셀의 대비를 측정), vegde-sd, hedge-mean(수직 인접 픽셀의 대비 측정), hedge-sd, intensity-mean((R + G + B)/3 의 영역 전체의 평균), rawred-mean(R의 영역 전체 평균), rawblue-mean(B의 영역 전체 평균), raw green-mean(G의 영역 전체 평균), exred-mean(빨강을 초과한 값 측정 (2R - (G + B))), exblue-mean(파랑을 초과한 값 측정 (2B - (G + R))), exgreen-mean(초록을 초과한 값 측정 (2G - (R + B))), value-mean(RGB의 3-d 비선형 정보), saturatoin-mean

3. 실험 결과

1) IRIS data set 실험 결과

(1) k-means 군집분석 실행 결과

k-means 군집분석 실행 결과				
	초기평균값 케이스번호	군집수	정확도	실행시간 (단위:s)
1	34, 48, 62	32:22:96	57.33%	1.10
2	35, 100, 88	50:61:39	90.00%	0.89
3	45, 75, 78	50:61:39	90.00%	0.80
4	16, 52, 128	50:61:39	90.00%	0.91
5	96, 85, 122	50:61:39	90.00%	0.70
6	72, 75, 120	50:62:38	90.67%	0.48
7	49, 72, 101	50:62:38	90.67%	0.30
8	83, 93, 86	50:61:39	90.00%	0.99
9	27, 90, 103	50:61:39	90.00%	0.41
10	22, 70, 135	50:61:39	90.00%	0.89
11	33, 38, 101	32:22:96	57.33%	0.97
12	94, 114, 132	50:62:38	89.33%	0.69
13	36, 28, 104	21:32:97	49.33%	0.30
14	32, 39, 99	32:22:96	58.00%	0.80
15	7, 96, 102	50:61:39	88.67%	0.80
평균			80.76%	0.74

<표 2> IRIS data set k-means 군집분석 실행 결과

<표 2>의 MacQueen(1967)이 제안한 k-means 군집분석[1]의 실행 결과에서 보면 초기 평균값 선택 시에 인접해 있는 두 객체가 선정될 경우에는 정확도 많이 떨어지는 것을 알 수 있다.

MacQueen(1967)이 제안한 k-means 군집분석은 초기 평균값 선택의 시간복잡도와 객체 배정의 시간복잡도가 각각 $O(k)$, $O(kln)$ (k:클러스터의 수, l:반복시행수, n:데이터의 수, $k \leq n$)로서 실행시간은 아주 빠른 것으로 나타난다.

(2) 강화된 k-means 군집분석 실행 결과

강화된 k-means 군집분석 실행 결과			
초기 평균값 케이스 번호	군집수	정확도	실행시간 (단위:s)
102,143,114,122,150,84,128,139,115,127, 124,147,71,112,134,120,73,64,104,79,135, 148,67,129,92,138,133,56,74,111,117,85, 116,57, 149,55,69,78	50:61:39	88.67%	6.00
8,40,50,1,28,29,18,27,5,12,41,38,35,10,36,22,49,2 1,32,30,31,24,3,20,26,47,2,44,7,11,25,48,37,13,46, 4,45,6			
58,94,61,99,82,81,80,60,70,54,90,65,83,93,95,68,1 00,63,89,91,97,96,107,72,62,98,88,75,86,76,52,42, 59,9,66,39,43,19			

<표 3> IRIS data set 강화된 k-means 군집분석 실행 결과

<표 3>의 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2] 실행 결과에서 초기 평균값 케이스 번호 중 진하게 된 숫자는 가장 가까운 한 쌍의 객체로 선정된 객체이며, 나머지 케이스 번호는 선택된 한 쌍과 가까이에 위치한 객체들이다.

정확도는 MacQueen(1967)이 제안한 k-means 군집분석[1]에 비해 늘어났으나, 초기치 선택의 시간복잡도 $O(n^2)$ (n:데이터의 수)로 인해서 실행시간에 있어서는 MacQueen(1967)이 제안한 k-means 군집분석에 비해 실행시간은 크게 늘어났다.

(3) 제안하는 k-means 군집분석 실행 결과

제안하는 k-means 군집분석 실행 결과									
	α =0.5			α =0.75			α =0.9		
	군집수	정확도 (%)	실행 시간 (단위:s)	군집수	정확도 (%)	실행 시간 (단위:s)	군집수	정확도 (%)	실행 시간 (단위:s)
1	50:62:38	89.33	0.84	50:62:38	89.33	0.50	50:61:39	88.67	0.82
2	22:32:96	64.00	0.54	50:61:39	88.67	1.28	50:62:38	89.33	0.55
3	50:61:39	88.67	0.54	50:61:39	88.67	0.49	50:62:38	89.33	0.54
4	50:61:39	88.67	0.92	50:61:39	88.67	0.56	50:61:39	88.67	0.69
5	50:61:39	88.67	1.13	50:61:39	88.67	1.00	50:61:39	88.67	0.53
6	50:61:39	88.67	1.02	50:61:39	88.67	1.41	50:61:39	88.67	0.73
7	50:61:39	88.67	1.03	50:61:39	88.67	0.96	50:61:39	88.67	1.13
8	50:62:38	89.33	0.42	50:61:39	88.67	1.42	50:61:39	88.67	1.27
9	22:32:96	42.66	0.85	50:61:39	88.67	1.18	50:61:39	88.67	0.79
10	22:32:96	42.66	0.85	50:61:39	88.67	0.70	50:62:38	89.33	0.51
11	50:62:38	89.33	0.77	50:61:39	88.67	0.48	50:61:39	88.67	0.81
12	50:62:38	89.33	0.75	50:61:39	88.67	0.56	50:61:39	88.67	0.42
13	50:62:38	89.33	0.44	50:61:39	88.67	0.68	50:61:39	88.67	0.93
14	50:62:38	89.33	0.72	50:61:39	88.67	1.30	50:61:39	88.67	0.83
15	50:62:38	89.33	0.42	50:62:38	89.33	0.61	50:61:39	88.67	1.03
평균		81.20	0.75		88.76	0.88		88.80	0.77

<표 4> IRIS data set 제안하는 k-means 군집분석 실행 결과

<표 4>의 제안하는 k-means 실행 결과를 보면 $O(kn)$ (k:클러스터의 수, n:데이터의 수. $k \leq n$)의 시간복잡도를 갖는 초기 평균값 선택 알고리즘으로 인해 실행속도가 MacQueen(1967)이 제안한 k-means 군집분석[1]과 가까울 정도로 향상되었다. 또한 정확도는 유지되고 있다.

(4) 분석 결과

	정확도(%)	시간복잡도		실행시간(s)
		초기평균값 선택	객체배정	
k-means	80.76	$O(k)$	$O(kln)$	0.74
강화된 k-means	88.67	$O(n^2)$	$O(kn)$	6.00
제안한 k-means	88.76	$O(kn)$	$O(kn)$	0.88

(k:클러스터의 수, l:반복시행 수, n:데이터의 수, $k \leq n$)

<표 5> IRIS data set 통합 결과 분석표 (α=0.75)

<표 5>의 IRIS data set 통합 결과 분석표를 보면, MacQueen(1967)이 제안한 k-means 군집분석[1]은 초기 평균값 선택의 시간복잡도가 작기에 당연히 실행시간에서 월등히 빠른 실행시간이 나타남을 알 수 있다.

그리고 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 MacQueen(1967)이 제안한 k-means 군집분석에 비해 정확도는 향상 되었지만, 초기 평균값 선택의 시간복잡도가 기존 k-means 군집분석에 비해 많이 크기 때문에 실행시간은 MacQueen(1967)이 제안한 k-means 군집분석보다 훨씬 더 늘어났다.

마지막으로 제안하는 k-means 군집분석은 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 시간복잡도 보다 작게 구현되어졌기에, 실행시간을 MacQueen(1967)이 제안한 k-means 군집분석의 실행시간에 가까게 나타나고 있으며, Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 높은 정확성도 유지시키고 있다.

2) Image Segmentation data set 실험 결과

(1) k-means 군집분석 실행 결과

k-means 군집분석 실행 결과		
	정확도(%)	실행시간 (단위:s)
1	46.90	90.94
2	53.52	35.67
3	51.57	53.73
4	60.10	101.90
5	54.38	84.86
평균	53.29	73.42

<표 6> Image Segmentation data set k-means 군집분석 실행 결과

<표 6>의 Image Segmentation data set k-means 군집분석 실행 결과를 보면, MacQueen(1967)이 제안한 k-means 군집분석[1]은 데이터의 수가 2000개가 넘어가도 초기 평균값 선택의 시간복잡도가 작기에 당연히 실행시간에서 빠른 실행시간이 나타남을 알 수 있다.

정확도는 IRIS data set k-means 군집분석 실행 결과와 비교해서는 떨어짐을 보이는데 이는 군집의 형태가 구(spherical)형태가 아니거나 군집을 뚜렷이 나누는 유효한 속성만이 아닌 속성 전부를 사용해서 정확도의 저하를 가져온 것으로 볼 수 있다.

(2) 강화된 k-means 군집분석 실행 결과

강화된 k-means 군집분석 실행 결과	
정확도(%)	실행시간 (단위:s)
46.67	1406.83

<표 7> Image Segmentation data set 강화된 k-means 군집분석 실행 결과

<표 7>의 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2] 실행 결과에서 보면 초기치 선택의 시간복잡도 $O(n^2)$ (n:데이터의 수)로 인해서 실행시간에 있어서는 MacQueen(1967)이 제안한 k-means 군집분석[1]에 비해 실행시간은 크게 늘어났다.

즉 데이터의 수가 2000개가 넘어가자 높은 시간복잡도로 인해 전체 실행시간에서 월등히 증가됨이 나타나고 있다.

정확도는 MacQueen(1967)이 제안한 k-means 군집분석에 비해 낮게 나타나고 있는데 초기 평균값 선택에서 선정된 가장 가까운 한 쌍의 객체들이 확실히 떨어져 있지 않음으로 인해 정확도가 더 낮게 나타난 것으로 볼 수 있다.

(3) 제안하는 k-means 군집분석 실행 결과

제안하는 k-means 군집분석 실행 결과						
	α =0.5		α =0.75		α =0.9	
	정확도(%)	실행시간 (단위:s)	정확도(%)	실행시간 (단위:s)	정확도(%)	실행시간 (단위:s)
1	50.24	61.22	46.14	101.75	48.24	159.99
2	53.29	145.41	49.48	83.98	58.14	94.61
3	57.00	122.30	56.10	132.95	51.76	163.72
4	43.90	172.14	59.24	126.58	53.71	170.73
5	50.00	76.36	57.05	156.54	60.52	130.46
평균	50.89	115.486	53.60	120.36	54.47	143.902

<표 8> Image Segmentation data set 제안하는 k-means 군집분석 실행 결과

<표 8>의 제안하는 k-means 실행 결과를 보면 $O(kn)$ (k:클러스터의 수, n:데이터의 수. $k \leq n$)의 시간복잡도를 갖는 초기 평균값 선택 알고리즘으로 인해 데이터의 수가 2000개 넘자 MacQueen(1967)이 제안한 k-means 군집분석[1]보다 증가하기 시작한다.

그리고 α의 값이 커질수록 제안하는 k-means 군집분석의 전체 실행시간이 증가함을 보이는데, 각 군집의 초기 평균값 선정을 위해 선택되는 객체의 수가 늘어나기 때문이다.(α=0.5이면 150개, α=0.75이면 225개, α=0.9이면 270개)

또한 α의 값이 증가할 때마다 정확도도 향상되고 있다.

(4) 분석 결과

	정확도(%)	시간복잡도		실행시간(s)
		초기평균값 선택	객체배정	
k-means	53.29	$O(k)$	$O(kln)$	73.42
강화된 k-means	46.67	$O(n^2)$	$O(kn)$	1406.83
제안한 k-means	53.60	$O(kn)$	$O(kn)$	120.36

(k:클러스터의 수, l:반복시행 수, n:데이터의 수, $k \leq n$)

<표 9> Image Segmentation data set 통합 결과 분석표 (α=0.75)

<표 9>의 Image Segmentation data set 통합 결과 분석표를 보면, MacQueen(1967)이 제안한 k-means 군집분석[1]은 초기 평균값 선택의 시간복잡도가 작기에 당연히 실행시간에서 월등히 빠른 실행시간을 나타냄을 알 수 있다.

그리고 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 MacQueen(1967)이 제안한 k-means 군집분석 보다 정확도가 하락 되었다.

데이터의 객체 수가 2000개가 넘어가자 초기 평균값 선택의 시간복잡도가 큰 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석은 MacQueen(1967)이 제안한 k-means 군집분석 보다 실행시간이 크게 증가했다.

마지막으로 제안하는 k-means 군집분석은 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 시간복잡도 보다 작게 구현되어졌기에, 실행시간을 단축시켰으며, Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석 보다 높은 정확성을 나타내고 있다.

Ⅵ. 결론

본 논문에서는 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]의 초기치 선정 방법에서 시간복잡도를 향상시켜 높은 정확도를 유지하면서 실행시간은 단축시키는 알고리즘을 구현하고 성능을 평가하였다.

MacQueen(1967)이 제안한 k-means 군집분석[1]은 초기 평균값 선택의 시간복잡도가 다른 두 알고리즘에 비해 월등히 낮기에 빠른 실행시간을 나타냈다.

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석은 정확도의 향상은 나타났으나, 초기 평균값 선택 시의 큰 시간복잡도로 인해 실행시간이 MacQueen(1967)이 제안한 k-means 군집분석에 비해 큰 차이를 보이며 높게 나타났다.

제안한 k-means 군집분석은 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 시간복잡도를 향상시켜 실행시간이 IRIS 데이터의 경우 MacQueen(1967)이 제안한 k-means 군집분석의 실행시간과 비슷하게 나타났으며, Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 정확도는 그대로 유지시켰다.

Image Segmentation 데이터의 경우는 데이터 객체수가 2000개를 넘어가자 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석이 MacQueen(1967)이 제안한 k-means 군집분석 보다 실행시간이 엄청나게 증가했다. 그리고 정확도도 하락했다.

하지만 제안한 k-means 군집분석은 MacQueen(1967)이 제안한 k-means 군집분석의 정확도 보다 조금 향상되었고, Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 실행시간 보다 단축된 실행시간을 나타냈다.

제안한 k-means 군집분석은 간단한 구조를 가지면서 빠르게 실행되고 최대한 정확도를 높이기에 수치형 데이터 분석에 사용이 쉬운 알고리즘이다. 그리고 이상치(outlier) 탐색에도 효율적이다.

초기 평균값 선택 과정에서 임의 선택된 객체와 군집범위 만큼 가까이에 위치

한 다른 객체를 포함해 가는데, 가장 가까운 객체와의 거리가 지나치게 큰 경우 이상치(outlier)라 판단하고 클러스터링에서 제외시킬 수 있다. 또한 임의 선택된 첫 번째 객체가 이상치(outlier)가 아닌 경우, 군집범위 외의 객체들을 이상치(outlier)로 판단이 가능하다.

하지만 데이터의 수 n이 무한대로 커지거나 지나치게 커질 경우는 MacQueen(1967)이 제안한 k-means 군집분석[1]과 비교해서 실행시간이 크게 차이 날 수 있기에 여기서는 데이터의 수가 무한대로 커지거나 지나치게 큰 경우는 제외한다.

향후 구(spherical) 형태의 군집을 갖지 않거나, 군집간 객체 수의 차이가 큰 경우, 군집이 서로 인접해 있는 경우 군집 경계 부분에서의 객체배정의 보완이 필요하다.

Ⅶ. 참고문헌

[1] J. MacQueen "Some methods for classification and analysis of multivariate observations." In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics, L. M. Le Cam and J. Neyman (Eds.). University of California Press, 1967.

[2] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the K-means Clustering Algorithm" In Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[3] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids." Proc. of the 3rd International Conference on Machine Learning and Cybernetics, August 2004.

[4] Data Minig Concepts and Techniques. Jiawei Han and Micheline Kamber, p.28, pp.42-52, pp.413-428

[5] E.Keogh C. Blake and C.j. Merz. UCI repository of machine learning databases, 1998

[6] “고객관계관리를 위한 데이터 마이닝 방법론”. 강현철 외 6명 자유아카데미. pp.271-276

[7] Data Mining for Business Intelligence. GALIT SHMUELI, 2009

[8] Bradley, P.S., Fayyad, U.M., 1998. "Refining initial points for K-means clustering." Proc. 15th Internat. Conf. on Machine Learning (ICML'98).

[9] INTRODUCTION TO DATA MINING. PANG-NING TAN, MICHAEL STEINBACH, VIPIN KUMAR

[10] S. Deelers, and S. Auwatanamongkol, 2007. "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance" World Academy of Science, Engineering and Technology 35 2007

Abstract

방대한 양의 데이터가 넘쳐 나오는 상황에서 데이터의 유용한 정보나 패턴을 추출해내야 하는 필요성의 기인해 데이터 마이닝은 주목을 받고 있다.

군집분석은 데이터 마이닝의 중요한 기법으로, 군집 내 데이터의 유사성을 최대로 하는 반면 군집 간 비유사성을 최대로 데이터를 군집화 시키는 방법이다.

군집분석은 연구의 활발한 주체로서 효과적이고 효율적인 군집분석을 할 수 있는 방법을 찾고 있다.

군집분석의 여러 가지 기법 중 분할기법을 사용하는 MacQueen(1967)이 제안한 k-means 군집분석[1]은 가장 유명하면서도 많이 사용되고 있다. k-means 군집분석은 간단하면서도 다양한 데이터 형태에 적용될 수 있다.

그러나 k-means 군집분석은 초기 평균값의 의존도가 너무 높아 초기 평균값 임의 선택 시에 인접한 객체들이 선택된다면 클러스터링 정확도는 저하될 뿐 아니라 객체의 재배정에도 시간을 많이 할애하게 된다.

MacQueen(1967)이 제안한 k-means 군집분석의 초기 평균값 선택과 객체의 재배정에서의 단점을 보안하여 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석[2]은 클러스터링 결과의 정확도를 크게 향상 시켰으며, 객체의 재배정 시간을 단축 시켰다.

Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석 역시 정확도의 향상과 객체 재배정 과정의 실행시간을 단축 시켰지만, MacQueen(1967)이 제안한 k-means 군집분석 보다 알고리즘 전체 실행시간이 길어진다는 단점이 있다.

본 논문에서는 Abdul Nazeer와 Sebastian(2009)이 제안한 강화된 k-means 군집분석의 초기 평균값 선택 시간복잡도를 향상시키는 방법을 제시한다.