데이터 마이닝에서의 범주형데이터 군집분석을 위한 초기치 선정방법

大學院

吳　秀　岷

2014年　2月

Initial points selection method for the clustering of categorical data in data mining

Su-min Oh
(Supervised by professor Chul Soo Kim)

A thesis submitted in partial fulfillment of the requirement
for the degree of Doctor of Philosophy

2014 . 2.

Department of Computer Science and Statistics

GRADUATE SCHOOL

JEJU NATIONAL UNIVERSITY

목 차

List of Table

List of Figure

Abstract

Data mining is the process of analyzing data from different perspectives and generates useful information. Technically, data mining finds correlations or patterns out of dozens of fields in large data.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group, called a cluster, are more similar to each other than to those included in other groups. It is not only a main task of exploratory data mining, but also a common technique for statistical data analysis, used in many fields, including information retrieval.

The k-means clustering method is usually based on vector quantization that is popular for cluster analysis in data mining. It also aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. As a segmentation method for cluster analysis schemes, the k-means algorithm is based on Euclidean distance, and is best suited for the analysis of numerical data set. However, in reality, there exist so many types of categorical data and more attention has been put on clustering such categorical data. After all, many researches need the cluster-based analysis mechanism especially for categorical data.

A study of Huang Z. 1997 has presented the k-modes algorithm which is an expansion of the original k-means clustering algorithm. The k-modes algorithm extends the existing k-means paradigm to cluster categorical data using the similarity criteria through the comparison

between attributes of the data by selecting representative data mode. However, the k-modes algorithm has a drawback in that it must decide the number of modes in advance of the iterative optimization process. If too many or less options are chosen, the accuracy of clustering will be severely deteriorated. In this paper, built on top of the well-known k-modes algorithm, how to select the number of modes and how to analyze the cluster set have been presented. The number of initial clusters decided by this way, makes it possible to efficiently estimate the final number of clusters. This step involves creating, merging and updating a great number of feasible initial modes of high similarity.

Experimental results, discover that the accuracy of the proposed algorithm yields almost equal results, compared with the case beginning from the optimized initial points. After all, the suggested algorithm is more effective than existing ways in the analysis of unknown data.

The rest of this thesis in organized as follows: Legacy segmentation strategies of cluster analysis techniques have been reviewed in Chapter II. Then, Chapter III describes our Chain k-modes algorithm in detail. This idea focuses on how to select the initial values for the case that they are not explicitly specified. In addition, how to analyze the clustered sets created by our scheme and the limited bound has been explained. Chapter IV has demonstrated the performance measurement results of the proposed algorithm, obtained from extensive simulation using R(3.0.1). Finally, Chapter V summarizes and concludes this thesis.

요 약

데이터에서 의미 있는 정보를 찾아내는 것은 매우 중요한 일이며, 이러한 과정을 데이터 마이닝(Data Mining)이라고 한다. “데이터 마이닝”이라는 용어는 1995년 지식발견 및 데이터 마이닝 국제학술대회를 개최한 이후 본격적으로 사용되어지고 있으며, 다양하게 정의되고 있다. [Berry and Linoff, 1997]는 “의미 있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반자동화된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정”이라고 정의하였으며, [Hand et al, 2001]는 “대량의 데이터로부터 유용한 정보를 추출하는 것”으로 정의하기도 한다.

즉, 데이터 마이닝은 미지의 데이터로부터 의미 있는 임의의 정보를 얻는 일련의 모든 방법이라고 할 수 있으며, 많은 기법들이 제시되고 있다. 최근 다양한 분야에서 빅데이터로 정의되는 정형, 비정형의 데이터에 대한 분석을 요구하고 있으며, 이러한 데이터를 수집, 분석하는 방법으로 데이터 마이닝 기법을 활용하여 실현시키고 있으며, 그 중요성은 날로 부각되고 있다. 데이터 마이닝 기법 중 군집분석기법은 다양한 분야의 수많은 데이터들로 부터 의미 있는 정보를 추출하기 위한 중요한 기법으로 인식되어 왔으며, 현재까지 군집분석기법에 대한 연구는 계속되고 있다.

군집분석기법 중에서 분할기법에 해당하는 k-means 알고리즘[Macqueen, 1967]은 유클리드 거리를 기반으로 하며, 수치형데이터의 분석에 매우 효율적이다. 하지만, 현실에서는 수치형데이터 뿐만 아니라 수많은 범주형데이터가 존재하며 중요시 되고 있다. 결국, 범주형데이터의 분석이 가능한 군집분석기법을 필요로 하였으며, [Huang Z. 1997a]의 연구에서는 기존 k-means 알고리즘을 수치형데이터에서 범주형데이터로 확장한 k-modes 알고리즘을 제안하였다. k-modes 알고리즘은 데이터의 속성간의 비교를 통한 유사도측정을 기반으로 하여 성질이 비슷한 데이터들에 대하여 대표 mode를 선정하는 방식이다. 하지

만, k-modes 알고리즘은 mode의 개수를 미리 정해야 하는 문제점이 있으며, mode의 개수를 너무 많이 설정하거나 적게 설정하면 군집의 정확도가 낮아지는 문제점을 지니고 있다.

본 연구에서는 k-modes 알고리즘을 기반으로 mode의 개수를 선정하는 방법과 이를 기반으로 하는 군집분석기법을 제안하였다. 초기치 mode들은 임의로 선택된 표본으로부터 추정하였다. 이를 기반으로 유사도가 높은 다수의 초기치 mode를 생성, 병합, 갱신하는 과정을 통해 최종 군집의 개수를 추정할 수 있었다. 실험결과 제안 알고리즘의 정확도는 최적의 초기치를 기반으로 군집분석을 수행한 결과와 유사한 결과를 보였다. 이는 제안 알고리즘이 초기치의 정보를 알 수 없는 미지의 데이터에 대한 분석에서 기존의 다양한 방법들보다 매우 효과적이라고 할 수 있다.

본 논문의 구성은 다음과 같다. II장에서는 분할기법 군집분석에 대한 설명을 하고 하였으며, III장에서는 제안 알고리즘에 대한 전반적인 과정을 설명하였다. 초기치가 주어지지 않았을 경우, 초기치를 선정하는 방법과 한계기준 및 제안 알고리즘을 이용한 군집분석 기법을 설명하였다. IV장에서는 제안 알고리즘의 실험결과이며, 실험은 R(3.0.1)을 활용하였다. V장에서는 논문에 대한 결론을 정리하였다.

1 데이터 마이닝

1.1 데이터 마이닝

최근 컴퓨터의 하드웨어 지속적인 발전과 더불어 강력해진 컴퓨터기술을 기반으로, 데이터의 수집, 저장 및 공급이 가능하게 됨으로서 정보기술 분야는 비약적인 발전을 이루었으며, 데이터는 다양해지고 대용량화되고 있다. 하지만 이러한 데이터들은 인간의 이해력을 넘어서고 있으며 보다 강력한 분석 도구를 필요로 하고 있다. 데이터 마이닝은 기계학습, 통계학, 데이터베이스, 고성능 컴퓨팅, 신경망, 정보 검색, 패턴인식, 공간 데이터 분석 및 다양한 학문에서 사용할 수 있는 기법을 포괄하고 있다. 즉, 양적으로 팽창되고 다양해진 데이터로부터 유용한 정보를 추출해야 할 필요성이 요구되고 있다. 데이터(data)는 연구나 조사등을 기반으로 관찰이나 측정을 통해 수집되는 숫자, 문자, 영상, 음성 등의 형태로서, 컴퓨터에 저장되는 자료(資料, data)이다. 또한, 이러한 데이터를 다양한 분석 및 처리를 통해 의미 있는 결과로 얻어진 결과를 정보(情報, information)라고 한다. 데이터 마이닝(Data mining)은 이러한 데이터로부터 유용한 정보, 지식, 규칙, 패턴 및 특성 등을 추출하고, 이를 이해하기 쉬운 형태로 변형하는 일련의 과정을 의미한다. 데이터 마이닝은 사용자에게 최적의 의사결정을 제공하지만, 데이터 마이닝 용어자체가 하나의 기법은 아니다. [Fayyad, Piatetsky-Shapiro, Smyth, 1996]는 우리가 필요로 하는 정보를 얻는 모든 행위를 의미하며, “데이터베이스에서의 지식발견(KDD: Knowledge Discovery in Database)을 위한 일련의 과정과 기법을 통칭하는 표현”이라고 하였다. 하지만 데이터마이닝이 KDD와 같은 의미를 의미하지는 않는다. 데이터 마이닝은 보다 적극적으로 사용자와의 상호작용을 강조하고 있다. 데이터로부터 추출된 정보로부터 유용한 정보를 추출한 후 이를 기반으로 합리적인 의사결정을 제안하는 도구로서 KDD에서의 한 과정이면서 KDD를 전체를 포함하는 과정을 의미한다. 결국, 데이터 마이닝은 대용량의 데이터로부터 흥미로운 지식을 발견하기 위한 다

양한 기법을 개발하기 위한 효율적이고 확장 가능한 기법이 필요하다.

Figure 1. 지식 정보 추출과정

Figure 1은 [Han & Kamber, 2006]가 설명하는 데이터 마이닝에서의 지식 정보 추출을 위한 필수적인 단계이다. 먼저, 각종 데이터로부터 잡음과 불일치 데이터를 제거하는 과정을 통해 데이터를 수집한다. 다음 단계로, 다수의 데이터에 대하여 연산, 정리, 결합 등의 과정을 거치면서 분석 작업과 관련된 데이터들을 검색한다. 이에 대하여 데이터 패턴을 추출하기 위한 데이터 마이닝 기법을 적용하는 가장 중요한 과정을 거치게 된다. 추출된 패턴에 대하여 몇 가지 척도를 이용하여 측정 평가하는 과정을 거치면서 사용자에게 의미 있는 지식 정보를 제공한다. 즉, 데이터 마이닝은 변환된 데이터를 가지고 일정한 패턴을 찾는 과정으로, 지식 정보 추출과정의 하나의 중요한 단계로서 표현이 되고 있다. 정보는 시시각각으로 빠르고 다양하게 변화한다. 이러한 정보는 불확실성이 커지게 되며, 그 가치가 적어질 수 있으므로 의미 있는 데이터를 추출하는 방법에 대한 연구는 매우 중요한 일이다.

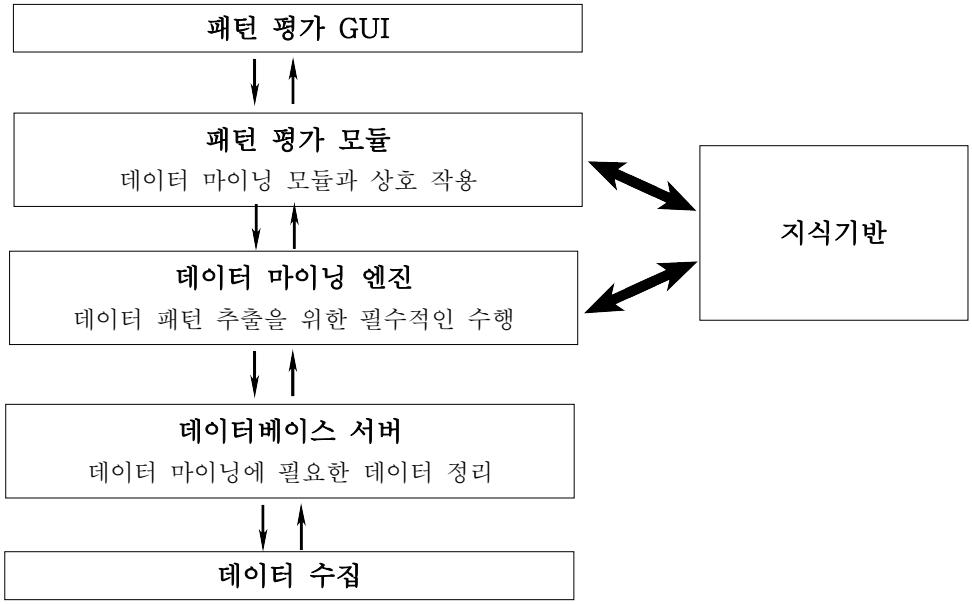


Figure 2. 개선된 데이터 마이닝 개념

Figure 2는 데이터 마이닝은 기존의 KDD 보다 진보된 시스템을 보여주고 있다. 데이터 마이닝의 가장 중요한 기본 요소는 상황에 맞는 올바른 방법을 선택하고, 이를 기반으로 정확한 정보를 추출하는 것이다. 데이터 마이닝은 지식정보 및 다른 요소들과 상호 연관을 통해 효율적인 처리를 가능하게 한다. 결국, 데이터 마이닝은 대용량 데이터베이스를 비롯한 다른 여러 가지의 다양한 데이터들로부터 상호작용을 통해 흥미로운 패턴을 찾는 과정으로서 표현된다. 이는 데이터베이스 시스템, 데이터 웨어하우스, 통계, 기계학습, 정보검색 및 신경망, 패턴인식, 공간데이터분석, 영상데이터베이스, 신호처리, 비즈니스, 경제학, 생명정보과학 등의 많은 응용분야에서 활용된다.

1.2 데이터 마이닝에서의 군집분석

데이터 마이닝에서 군집분석(Cluster Analysis)은 비슷한 속성을 가지고 있는 데이터를 합치면서 의미 있는 군집을 형성하는 과정이며, 다양한 기법들이 개발되고 있다. 군집분석기법은 데이터 마이닝 기법 중 의사결정트리(Decision Tree), 신경망분류(Neural Network), 분류(If-Then)규칙과 같은 분류(Classification)와 예측(Prediction)하는 기법들과 유사하다. 하지만, 군집분석은 데이터의 클래스레이블이 주어지지 않으며, 군집은 군집내에서는 강한 유사성을 보장해야 하며, 군집간에 대해서는 상이한 특성을 보이게 된다. 통계학에서 군집분석은 k-means 알고리즘이 가장 보편적으로 활용되고 있다. k-means 알고리즘은 데이터 레코드간 거리를 기반으로 하고 있으며, SPSS, SAS, R, S-Plus 등의 소프트웨어에서 잘 구현되어 있다.

군집분석은 데이터 마이닝 기법 중 매우 효율적이고 범용의 방법으로서 효율적인 군집분석 방법을 찾는 연구가 진행 중에 있다. 연구의 주제로는 데이터의 확장성, 속성이 다른 데이터 타입을 다루는 능력, 임의의 형태에서 군집 발견, 군집의 개수를 결정하는 방법, 이상치를 다루는 능력, 고차원, 입력순서, 상호작동성 등에 대하여 연구가 진행 중에 있다.

데이터의 확장성은 대용량의 데이터를 다루는 능력을 의미한다. 대용량의 데이터는 수백만의 데이터 레코드를 가지고 있으며, 다양한 크기의 데이터에 대하여 편향되지 않는 방법을 요구하고 있다.

속성(attributes)이 다른 타입을 다루는 능력은 군집분석에 있어 매우 중요한 연구주제이다. 많은 알고리즘이 수치형 속성을 기반으로 하고 있다. 하지만 실제 데이터들은 범주형 속성(Categorical attributes)및 혼합형 속성(Mixed attributes)등으로 구성되어 있으며, 이러한 다양한 속성에 대한 분석방법이 필요하다.

임의의 형태에서 군집 발견은 유클리드 거리를 기반으로 하는 군집분석의 단점을 보완하기 위한 방법이다. 유클리드 거리는 수치형 속성에 대한 분석이므로

일반적으로 구형태의 군집을 취하게 되며, 이는 특정한 형태의 군집에 대해서는 군집결과가 좋지 않게 된다. 결국 임의의 형태의 군집을 찾을 수 있는 알고리즘도 개발해야 한다.

군집의 개수를 결정하는 방법은 군집분석의 가장 큰 단점인 초기치 선정문제이다. 대부분의 군집분석방법은 적절한 군집의 수를 지정해야한다. 이는 군집의 분류 정확도 및 분석 결과에 영향을 주므로 매우 중요한 연구주제이다.

이상치를 다루는 능력은 데이터의 군집의 질에 관련된 문제이다. 대부분의 데이터는 이상치, 결측치 및 오류데이터를 포함하고 있다. 이러한 데이터를 찾아내고 이를 해결하는 문제는 군집의 품질에 직접적인 영향을 주며, 다양한 분야에서 적용할 수 있으므로 많은 연구가 진행 중이다.

몇몇 알고리즘은 데이터 레코드의 입력 순서에 따라 알고리즘의 결과가 달라지기도 한다. 입력 순서는 데이터 군집분석에 영향을 주어선 곤란하다. 알고리즘의 개발에 있어서도 입력 순서에 둔감하도록 설계해야 한다.

대다수의 군집 분석 알고리즘은 저차원 데이터들에 대하여 특정화 되어있다. 하지만 실제 데이터는 고차원의 데이터로 이루어진 경우가 많으므로, 다양한 속성의 고차원 데이터를 다룰 수 있어야 한다.

상호작동성은 군집분석 결과가 사용자가 이해하기 쉬어야 한다는 점이다. 즉, 복잡한 데이터로부터 의미 있는 정보를 추출하고 이를 이해하기 쉽게 표현할 수 있느냐는 것은 중요한 연구주제이다.

본 연구에서는 이러한 연구조건에 최대한 부합되도록 알고리즘을 설계하였다. 대용량의 데이터에서 수행가능하고, 범주형 속성 데이터를 다루었으며, 거리기반이지만 구형태가 아니어도 군집이 가능하도록 하였다. 특히, 군집의 개수를 결정하기 위하여 데이터의 속성을 활용하는 방법을 제안하였다.

1.3 군집분석의 방법

군집분석의 방법은 데이터의 타입과 분석의 목적과 이에 대한 응용에 따라 달라지며, 필요에 따라 같은 데이터에 대하여 여러 기법을 사용하여 분석 할 수 있다. 군집분석 방법은 다양하게 분류할 수 있다. 분할기법(Partitioning Method), 계층적 기법(Hierarchical Method), 밀도기반 기법(Density-Based Method), 격자기반 기법(Grid-Based Method) 및 모델기반 기법(Model-Based Method)등으로 구분할 수 있다.

· 분할기법(Partitioning Methods)

분할기법은 개의 데이터에 대하여 k개의 데이터 분할을 만드는 기법이다. 분할기법 알고리즘은 군집들이 내부적으로는 강한 유사도를 보이며 군집간에는 상이도가 크게 된다. 즉, 군집간의 차이가 크면 잘 된 군집분석이라 할 수 있다. 일반적으로 분할기법은 초기 분할k에 대하여 반복적인 재배정기법(Iterative Relocation Technique)을 사용하며, 좋은 결과를 얻기 위해서는 모든 군집들은 최소한 하나 이상의 데이터를 포함해야하며, 모든 데이터는 k개의 군집으로 중복없이 군집되어야 한다.

· 계층기법(Hierarchical Method)

계층기법은 주어진 데이터들을 계층적으로 분할하며, 계층의 분할 형태에 따라 상향식 접근방식과 하향식 접근방식으로 구분할 수 있다. 이러한 방식은 일반적으로 트리형태의 구조를 이루게 된다. 계층기법은 조건이 종료될 때까지 반복하면서 주변의 데이터들을 합치는 과정을 수행하며, 데이터가 하나의 군집으로 되거나 혹은 더 작은 군집으로 분할될 때 까지 수행된다. 계층기법 군집분석의 효율을 높이는 방법으로는 트리에 대한 각각의 다른 군집분석 기법을 적용하는 방법이 있다. 트리구조의 계층을 이용한 반복적인 감소를 통한 군집분석 기법으로 BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies)기

법, 구 형태를 중심으로 하는 CURE(Clustering Using REpresentatives)기법 및 동적인 모델을 이용한 계층 군집분석 알고리즘인 Chameleon기법 등이 있다.

・밀도기반기법(Density-Based Method)

밀도기반기법은 데이터속성의 밀도 값을 기반으로 군집분석 하게 된다. 이 방식은 데이터들의 이상치 값을 배재한 형태의 군집을 생성하게 된다. DBSCAN (Density-Based Method Spatial Clustering of Application with Noise)기법은 데이터들의 밀도를 가지고 분석하는 기법으로서 밀도가 높은 지역을 중심으로 군집분석하며 이상치 값을 갖는 공간에 대해서도 임의의 형태의 군집을 구현할 수 있다. OPTICS(Ordering points to identity the clustering structure) 기법은 DBSCAN 기법의 단점을 보완하기 위하여 제안된 기법이다. 밀도를 계산하기 위해 사용되어지는 인자들의 상호작용을 통해 광범위한 인자들로부터 점진적인 군집분석 순서를 계산하여 군집분석을 수행한다. DENCLUE (Density-Based Clustering)기법은 밀도분포함수를 기반으로 하는 군집분석 기법으로서 데이터간의 연관성을 영향력 함수(Influence function)를 사용하여 모델화한다. 데이터의 밀도는 데이터의 영향력 함수들의 합으로 표현이 되며, 이 함수의 극대값을 군집화에 활용하는 기법이다. 안정적인 수학적 기반의 수행 및 고차원 데이터에 대한 효과적인 수행을 보이며, 분할기법, 계층기법등의 다양한 기법에 활용할 수 있다. 또한 군집분석 과정에서 발생하는 이상치 값들에 대한 효율적인 군집분석에 효과가 있다.

・격자기반기법(Grid-Based Method)

격자기반기법은 데이터공간을 격자구조로 이루어진 유한개공간으로 나누고 군집분석은 이러한 격자 내에서 격자들을 군집분석하게 된다. STING (Statistical Information Grid)은 데이터 공간구조를 사각의 격자기반으로 표현하여 군집분석을 수행한다. 높은 수준의 셀들은 보다 낮은 수준의 셀로 분할되기 위해 계층구조내의 질의 응답과정 반복을 통하여 계산되어지게 되고, 질의에 적합한 셀들이 결과로서 구해진다.

이 기법의 장점은 각 셀의 정보가 독립성을 유지할 수 있으며, 병렬처리 및 점

진적 수정을 수월하게 함으로서 효율성을 높인다. 하지만 군집의 형태가 수직수평으로만 관계를 함으로써 수행시간에 있어 효율적일 수는 있으나 군집의 정확성은 높지 않다. Wavecluster는 데이터 공간의 다차원 격자 구조를 기반으로 작동하는 밀도기반 알고리즘이다. 군집의 형태가 일반적으로 모자형의 군집으로 표현이 가능하여 군집화 되는 영역과 군집화 되지 않는 영역간의 정보를 보다 효율적으로 표현한다. 결국 군집의 형태가 일정하지 않은 데이터에 대해서 효율적으로 작동하며, 기존의 기법들(BIRCH, CLARANS, DBSCAN)보다 우수한 성능을 보여준다. CLIQUE(Clustering In Quest)기법은 밀도기반기법과 격자기반 알고리즘을 혼합시킨 방법으로 고차원데이터에 대해 효율적인 기법이다. 데이터 공간을 사각형태의 격자형태로 구분하여 밀도에 따른 사각형의 조밀도를 계산하게 된다. 이 기법은 데이터의 순서에 영향을 받지 않으며 특정한 분포를 따르지 않는다. 또한, 입력된 데이터에 따른 선형변화가 가능하다. 하지만 두 가지 기법의 단순혼합방식으로 인해 군집분석 결과의 정확도가 낮아질 수 있다.

· 모델기반

모델기반(Model based method) 기법은 각 군집에 모델을 가정하고 주어진 모델에서 가장 잘 맞는 데이터를 찾는 방식이다. 통계적 분석결과를 기반으로 하여 군집의 수를 결정하게 되며 이상치 값을 고려한 군집분석을 수행한다.

본 연구에서는 분할기법의 장점과 계층적 기법의 상향식(bottom-up)기법을 활용하였다. 분할기법의 장점은 데이터를 k개의 mode로 구분하여, 반복, 재배정 과정을 거치면서 최적의 k개의 군집을 만들 수 있으며, 계층적 기법에서 상향식(bottom-up)기법은 여러 개의 데이터를 병합하면서 최종적으로 최적의 군집으로 병합한다. 본 연구에서는 이러한 장점들을 기반으로, 먼저 데이터를 많은 mode로 분할시키며, 상향식 기법에 따라 mode들을 병합, 갱신, 재배정 과정을 거치면서 최적의 k개로 mode들을 군집화 하였다.

2 분할기법 군집분석

군집분석에 대한 연구는 활발하게 진행되고 있다. 최근 연구의 주제로는 대용량 데이터에서의 군집분석기법의 확장성, 속성이 다른 타입을 다루는 능력, 임의의 형태에서의 군집 발견, 최적의 군집결과를 위한 최소의 인자를 선정하는 초기치 선정 문제, 잡음(Outlier) 데이터를 다루는 기법, 실시간으로 입력되는 데이터에서 발생하는 편의현상, 고차원에서의 군집분석, 제약 조건에 따른 군집분석 기법 및 사용자간의 상호의사소통의 효용성 등이 있다.

본 연구에서는 다양한 군집화 기법 중에서 분할기법을 기반으로 하는 알고리즘을 다루고 있다. 일반적으로, 데이터의 속성은 수치속성으로 나타나는 수치형 데이터와 범주속성들로 이루어진 범주형데이터 및 수치형데이터와 범주형데이터가 혼합되어 존재하는 혼합형데이터로 구분한다. 수치형데이터를 기반으로 하는 군집분석 기법으로는 k-means 알고리즘과 [Kaufman, 1987]이 제안한 k-medoids 알고리즘이 잘 알려져 있으며, 범주형데이터에 대해서는 k-means 알고리즘을 변형한 k-modes 알고리즘이 있다. 수치형데이터와 범주형데이터가 혼합되어있는 혼합형데이터를 군집분석 할 수 있는 기법으로는 [Ahmad, 2003]이 제안한 k-prototype 알고리즘이 알려져 있다.

2.1 데이터에서의 레코드와 속성

데이터(data)에서의 레코드는 데이터베이스에서의 개체(entity)를 의미하며, 레코드는 현실에서 생각할 수 있는 개념이나 정보의 단위로서 서로 구별되는 자료들이다. 단독으로 존재할 수 있는 레코드들은 스스로 정보로서의 역할을 할 수 있으며, 하나의 레코드에는 하나 이상의 속성(attributes)으로 구성되어있다. 또한, 레코드내의 각 속성은 그 개체의 특성이나 상태를 설명한다.

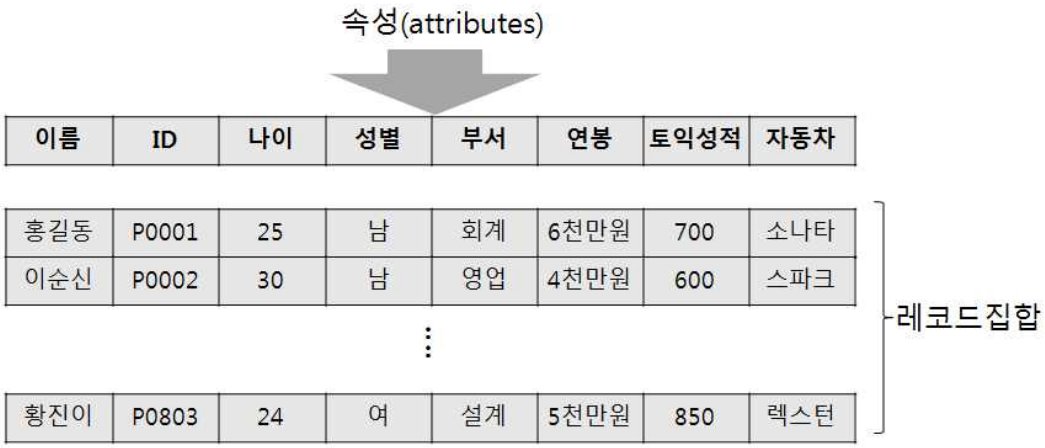


Figure 3. 데이터에서의 레코드 및 속성

예를 들어, Figure 3에서는 회사원들의 이름, ID, 나이, 성별, 부서, 연봉, 토익성적, 자동차의 8가지 속성으로 구성되어있는 레코드집합을 보여주고 있다. 여기에서 이름, ID, 나이와 같은 속성들은 회사원 레코드의 특성을 나타내고 있다. 또한, 이러한 레코드집합은 데이터베이스로 저장된다. 레코드의 속성은 크게 수치형 속성과 범주형 속성으로 구분할 수 있다. 속성 중 나이, 연봉 및 토익성적은 수치형 속성이며, 이름, ID, 성별, 부서 및 자동차는 범주형 속성이 된다. 군집분석에서는 데이터의 속성에 따라 적절한 기법을 적용할 때 사용자에게 최적의 분석 결과를 제공 할 수 있다.

2.1 수치형데이터의 군집분석 알고리즘

2.1.1 수치형데이터의 속성

· 구간척도변수

연속형 척도로서 수치적으로 측정할 수 있는 속성(attributes)에 대한 표현을 구간척도변수(Interval scaled variable)라고 한다. 예를 들어, 몸무게, 키, 위도, 경도 등으로 표현된 척도들이 대표적인 구간척도변수가 된다. 이러한 척도들에 대한 군집분석은 결국 측정된 수치의 단위에 따라 군집화에 영향을 받게 되며, 이에 대해서 측정된 수치의 표준화 과정을 수행할 수도 있다. 그리고 데이터의 레코드 및 군집들 간의 유사도 측정 및 상이도 측정은 유클리드 거리를 기반으로 한다.

임의의 개의 데이터에 대하여 레코드 , $j(i,j \in N)$가 p개의 속성을 포함할 때, 레코드는 p개의 데이터 속성에 대한 p차원으로 표현이 가능하다.

즉, 거리 데이터의 레코드간 거리 $d(i,j)$는

$$d(i,j) = \quad |x_1 - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \quad + |x_{ip} - x_{jp}|^2$$

으로 표현이 가능하다.

또한, 구간척도변수를 속성별로 가중치를 적용하면, 가중치(w)는 속성p에 대해서 다음과 같이 표현할 수 있다.

$$d(i,j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \quad + w_p|x_{ip} - x_{jp}|^2}$$

이러한 가중치적용방식은 범주형데이터 속성의 레코드의 경우에도 유사하게 적용할 수 있다.

2.1.2 k-means 알고리즘

[Macqueen, 1967]이 제안한 k-means 알고리즘은 초기치 k를 기반으로, 군집내 유사성은 높고 군집간 유사성은 낮게 되도록 개의 데이터를 k개의 군집으로 군집화하는 기법이다. 이 때, 각 means들은 각 군집의 평균값을 취한다.

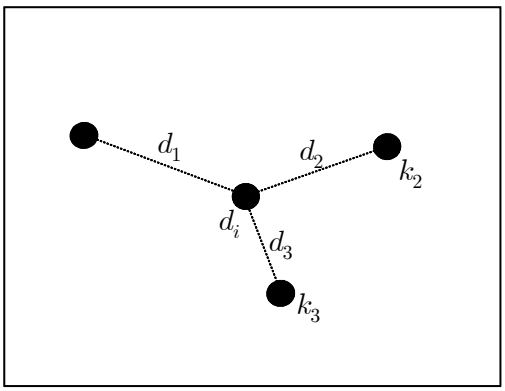


Figure 4. 2차원 평면상의 데이터와 군집간의 거리

Figure 4에서 d_i는 2차원 평면상의 한 점으로 표현된 데이터들이다. 는 먼저 선정된 초기 군집의 중심 k_1, k_2, k_3와 각각의 거리를 계산하게 된다. Figure 4에서는 k_1과 가장 가깝게 나타나므로 d_i는 k_1의 군집에 포함이 된다. 마찬가지 방식으로 반복과정을 통해 모든 데이터를 k_1, k_2, k_3와 거리를 계산해서 포함관계를 찾아낸다. k_1, k_2, k_3들이 포함하는 각각의 데이터들에 대하여 데이터들의 중심을 계산하게 되어 k_j k'_j로 군집의 중심을 구한다. 알고리즘의 반복을 통해 β번 반복 후 $k_j^{\beta-1}$ k_j^{β}의 결과가 변화가 없다면 군집분석을 종료한다.

· 유사도측정방법

임의로 선택된 레코드에 대하여, 다른 군집과의 평균(*means*)거리를 기반으로 군집분석을 진행한다.

$$d = \sum_{=1p\in}^{k} |p - m_i|^2$$

- p : 주어진 레코드의 위치
- m_i : 군집의 중심위치

일반적으로 레코드간 거리 d는 데이터에서의 레코드와 군집 간의 제곱오차의 합으로 표현한다.

---

k−means algorithm

→ 입력 : n개의 수치형데이터베이스, 초기치 선정k의 개수

(1) 입력된 데이터에 대한 초기치 k개의 평균값(*means*) 지정

(2) k와 데이터간의 유클리드거리 기반의 유사도 측정

(3) 유사도 값에 따른 소속 군집 할당

(4) 군집의 *means* 갱신

(5) 군집의 변화가 없을 때까지 (2)~(4)단계 반복수행

→ 출력 : 최적 배정 k 군집

Figure 5. k−means 알고리즘

k−means 알고리즘은 수치형데이터에 최적화 되어있다. 군집내 및 군집간 유사도를 유클리드 거리를 기반으로 설명하고 있으며, 레코드와 군집간의 거리가 작으면 유사하다고 판단하여 해당 군집에 할당하며, 거리가 크면 상이하다고 판단하여 다른 군집으로 할당하는 알고리즘이다. k−means 알고리즘은 수치형데이터에서 분석 속도가 빠르고 좋은 결과를 보여준다. 하지만 거리 계산방식이 유클리드거리를 기반으로 하기 때문에 범주형데이터 환경에서는 분석 할 수 없다. 또한, k−means 알고리즘은 사용자가 군집의 초기치 개수 k를 미리 정해야 한다.

이는 잘 선정된 초기치는 최종 군집의 분류 정확도 및 결과에 영향을 주게 되며, 초기치가 부적절하게 선정될 경우 군집분석의 결과가 매우 나쁘게 나타나는 문제를 가지고 있다. 또한, 군집의 선정은 k개의 군집 중심을 임의로 선정함으로써 정확도가 낮아지는 문제점이 나타나기도 한다.

2.1.3 EM 알고리즘

[Dempster, Laird, Rubin, 1977]이 제안한 EM(Expectation-Maximization) 알고리즘은 불완전 정보(missing data)에 대한 사후분포의 해석적 표현을 얻을 수 있으며, 확률적인 모델을 기반으로 하고 있다. 완전 정보 데이터(complete data)의 최대우도추정(Maximum Likelihood Estimation)의 해석적 표현이 가능할 때는 언제나 강력한 계산도구로 사용할 수 있으며, 추정치(estimate)가 안정적이다. EM 알고리즘은 보이지 않는 잠재 변수(latent variable)에 의존하는 확률모델에서 모수들의 최대우도추정치(Maximum Likelihood Estimates of parameters)를 찾고자하는 알고리즘이다.

완전 정보의 확률변수 와 불완전 정보의 확률변수 Z에 대하여 모수 벡터 가 존재할 때, $L(\theta)$를 최대로 하는 θ의 값을 찾는다. 이때, $P(X|\theta)$는 θ의 likelihood가 된다. 이때 log를 적용하면, $L(\theta) = \ln P(X|\theta)$이 된다.

(X,Z)에 대한 확률분포는

$$L(\theta;X,Z) = P(X,Z|\theta)$$

가 되며, 우도 함수는

$$L(\theta;X) = P(X|\theta) = \quad P(X,Z|\theta)$$

로 정의하게 되며 최대우도추정치를 구한다. EM 알고리즘은 임의의 모수 $\theta^{(t)}$을 이용하여 새로운 모수 $\theta^{(t+1)}$을 E(Expectation)단계와 M(Maximization)단계로

나누어 찾게 된다.

E단계에서는 불완전한 잠재 변수의 기대치를 계산하기 위하여 주어진 정보를 이용하여 추정치를 구하고 불완전 정보에 대한 기댓값을 추정하게 되며,

$$\theta|\theta^{t)} = E_{|X,\theta^{t)}}[\log L(\theta|\theta^{(t)})]$$

을 구한다.

M단계에서는 주어진 데이터와 기대치가 부여된 잠재 변수를 이용하여 모수들의 최대우도추정치를 계산한다. E단계에서 계산된 $Q(\theta|\theta^{(t)})$에 대하여 불완전 정보의 기대 추정치를 최대우도추정치로 하는

$$\theta^{(t+1)} = \arg max\ Q(\theta|\theta^{(t)})$$

을 찾는다.

EM 알고리즘은 완전정보 모델의 최대우도 추정이 쉬울 때 특히 유용하다. 이 알고리즘은 k-means 알고리즘을 확장한 개념으로서, 데이터를 특정한 군집에 할당하는 방법이 아닌, 해당 군집에 속할 확률을 나타내는 가중치를 기반으로 군집에 할당하게 된다. EM 알고리즘도 수치형데이터에 최적화 되어있으며, 초기치 k를 임의로 선정한다

EM algorithm

→ input : With any initial values of the parameter

E-step : $Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}}[\log L(\theta|\theta^{(t)})]$

M-step : $\theta^{(t+1)} = \arg max\ Q(\theta|\theta^{(t)})$

→ output : when the current parameter values approximately coincide with the previous ones

Figure 6. EM 알고리즘

2.1.4 k-medoids 알고리즘

[Kaufman, Rousseeuw, 1987]이 제안한 k-medoids 알고리즘은 k-means 알고리즘이 수치적 이상치에 민감하다는 점을 고려하여 이를 개선한 방법이다. 군집에서 레코드의 중심을 평균값으로 취하는 대신에 군집 중심에서 가장 가까운 곳에 위치한 레코드를 medoids로 지정하고, 다른 레코드와 연관된 참조점 간의 거리를 기반으로 하여 상이도 합을 최소화한다. k-medoids 알고리즘은 AM(Partitioning Around Medoids)이라고 알려져 있으며, k-medoids 알고리즘은 개의 초기 medoids를 임의로 선정한 후 N개의 데이터를 k개로 분할한다. 데이터의 레코드들은 medoids간의 제곱오차가 가장 낮은 medoids로 소속되며, 더 나은 medoids로 만들기 위해 반복적으로 medoids를 갱신한다. 군집의 중심이 평균을 이용하지 않으므로, 이상치가 존재할 때 k-medoids 방식은 이상치의 영향을 덜 받기 때문에 k-means 알고리즘 보다 효율적이다. 하지만 k개를 정해야 하는 문제점은 여전히 해결해야하는 문제점이다.

K-medoids algorithm (PAM)

→ input : Initial setting. Choose the number of clusters, K,

1. Select K entries $c, c_2, ..., c \in I$

1.1 Assume initial cluster list S_k is empty

2. Clusters update. Given K medoids $c_k \in I$

2.1 Determine clusters S'_k(k=1, ..., K)

with the Minimum distance rule applied to dissimilarity $d(i,j), i,j \in I$

3. Stop-condition. Check whether $S' = S$. If yes, end clustering

S=$\{S_k\}$, c= (c_k). Otherwise, change S for S'

→ output : Medoids update. Given cluster S_k,

determine their medoids c_k (k=1,...k) and go to step 2

Figure 7. k-medoids 알고리즘(PAM)

2.1.5 Fuzzy C-means 알고리즘

[J.C.Bezdek,1974]이 제안한 Fuzzy C-means 알고리즘은 각 데이터의 레코드들이 특정 군집에 속하는 정도를 계산하고, 이를 군집간의 거리로 나타내는 기법이다. 군집과 레코드간의 유사성은 군집의 중심과 레코드간의 유클리드 거리에 비례하게 되고, 이를 레코드와 군집간의 연관정도로 측정하게 된다.

개의 레코드를 X x , x_2, ..., $x_n\}$라고 하고, 퍼지 군집의 중심 벡터를 $V=\{v_1,\ v_2,\ ....,\ v_c\}$라고 하면, 각 군집과 데이터간 거리는 다음과 같이 표현된다.

$$J_m = \sum_{=1,J=1}^{n,c} u_{ij}^m |x_i - c_j|^2, \quad 1 \quad m <$$

-. c_j : j 번째 퍼지중심

-. u_{ij} : i, j 데이터의 연관정도

-. m : 각 데이터의 연관정도에 대한 퍼지값(1보다 큰 값 사용)

Fuzzy C-means algorithm

→ input : 군집의 개수(c)와 퍼지값(m) 계산

1. 데이터의 연관정도 계산

   $\sum_{i=1}^{c} u_{ij} = 1$을 만족하는 x_j에 대한 u_{ij} 갱신

2. 군집의 중심 계산

   $v_i = \sum u_{ij}^m x / \sum u_{ij}^m$, $i = 1,\ 2,\ ...,\ c$

3. 군집에 대한 소속 행렬 U 계산

$$u_{ij} = \left(\frac{1}{d^2(x_j, v_i)} \right)^{\frac{1}{m-1}} / \sum_{k=1}^{c} \left(\frac{1}{d^2(x_j, v_k)} \right)^{\frac{1}{m-1}}$$

→ 종료조건 : $|\{J_m^{(l)} - J_m^{(l-1)}\}| \quad \epsilon$

종료조건이 만족할 때까지 2,3과정 l회 반복

Figure 8. Fuzzy C-means 알고리즘

2.2 범주형데이터의 군집분석 알고리즘

2.2.1 범주형데이터의 속성

· 명목변수

명목변수(Nominal Variable)는 둘 이상의 상태를 가지고 있는 이항 변수의 일반형으로서, 색상, 국가, 이름과 같이 구분할 수 있는 속성(attributes)이며, 속성의 개수( )는 1, 2, 3, ..., n 으로 이루어진 정수로 표현할 수 있다.

명목변수의 상이도 측정은 변수간의 단순비교를 통하여 계산한다. 그리고 레코드간의 유사도 및 상이도 계산에 필요한 거리( )는 다음과 같이 표현된다.

$$d(i,j) = \frac{N_n - N_m}{}$$

- i, j : 전체 데이터내의 임의의 데이터,
- N_n : 전체 속성의 개수,
- N_m : 전체 속성 중 일치한 속성 개수

즉, 단순 비교를 통하여 레코드간의 거리(d)가 크면 상이도는 증가하며 유사도는 낮게 나타난다.

· 이항변수

0 또는 1로 표현되는 이항변수(Binary Variable)는 수치형으로 표현할 수 있다. 하지만 이항변수에 대한 유클리드기반의 거리 측정 방식은 군집분석 결과에 오차가 발생하게 된다. 그래서 이항변수에 대해서는 다음의 유사도 측정방식을 사용한다.

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- $i,j=1$,
- r $i=1,j=0$,
- s $i=0,j=1$,
- t $i,j=0$

· 서열변수

이산서열변수(Discrete Ordinal Variable)는 명목변수와 유사하지만 서열 값들이 의미를 가지고 순서화 되어있다. 예를 들어, 스포츠경기의 “금”, “은”, “동”은 명목변수이지만, 실제 변수가 포함하고 있는 가치는 다르며, 서열변수의 f개 항목에 대하여 1, …, 로 정의한다. 그리고 서열변수의 상이도 측정은 구간 척도형 변수와 매우 유사하게 계산이 되며, 각 서열에 대한 가중치를 적용함으로서 계산되어진다. i번째 데이터의 순위를 r_{if} 라고 할 때, i번째 레코드의 z_{if}는 다음과 같이 표현할 수 있다.

$$z_{\mathrm{if}} = \frac{r_{\mathrm{if}} - 1}{M_{\mathrm{if}} - 1}$$

결국, 서열변수에서는 각 변수의 범위를 [0.0, 1.0]으로 표시함으로써 각 변수들에게 일정한 가중치를 적용하여 상이도를 측정할 수 있다.

2.2.2 k-modes 알고리즘

[Huang, 1997]이 제안한 k-modes 알고리즘은 범주형데이터를 군집분석하기 위하여 k-means 알고리즘을 확장한 기법이다. 유클리드 거리를 기반으로 하는 k-means 알고리즘의 거리계산 방법을 확장하여 범주형 속성의 동일여부를 레코드간의 거리로 판단하게 된다. 개의 레코드 집합으로 이루어진 데이터에서 번째 레코드 $N,(i \in N)$는 $N_i = C_1, ..., C_p\}$으로 p개의 범주형 속성을 포함하고 있다. k-modes 알고리즘은 x번째 레코드의 $C_r(r \in p)$번째 속성의 속성값과 y번째 레코드의 $C_r(r \in p)$번째 속성의 속성값을 비교하여 유사도 측정에 필요한 거리($d_{x,y}$)를 계산한다.

	자동차	취미	직업	집	수익
$record_1$	티코	테니스	학생	아파트	100만원 이하
$record_2$	티코	인라인	학생	단독주택	100만원 이하
$record_3$	벤츠	테니스	의사	단독주택	100만원 이상
$record_4$	소나타	골프	변호사	아파트	100만원 이상

Figure 9. 범주형데이터에 대한 속성 값 표현

Figure 9에서 초기치로 $\{record_1, record_4\}$가 선택될 경우, $\{record_1, record_4\}$는 다른 데이터들과 거리를 비교하게 된다. 이미 언급했듯이 범주형데이터의 속성은 유클리드 거리 개념을 사용할 수 없으므로 확장된 방식으로 레코드를 비교하게 된다. 범주형 속성은 비교하는 레코드의 속성 값이 같으면 0, 같지 않으면 1을 취하며, 이에 대한 합을 유사도 거리로 정의 된다. $record_1$은 $record_2, record_3$ 및 $record_4$에 대하여 (2, 5, 4)의 거리가 발생한다. 결국, $record_1$은 거리가 가장 가까운 $record_2$와 같은 mode를 구성하게 된다. 마찬가지 방법으로 $record_4$는 $record_3$와 mode를 구성한다.

k-modes 알고리즘의 거리계산은 $,y\} \in$, 일 때,

$$d_{,y} = \sum_{i=1}^{p} \delta(x_i,\ y_i), \quad where \quad \delta(x_i,\ y_i) = \begin{matrix} 0 & (x_i = y_i) \\ 1 & (x_i \quad y_i) \end{matrix}, \ i \in p$$

로 구한다.

즉, 비교하는 데이터베이스의 레코드들에 대하여 각각의 속성 값이 같을 경우는 0, 같지 않을 경우는 1로 계산하게 되며, 비교하는 레코드의 거리는 이들에 대한 합으로 계산된다. 결국, 거리($d_{x,y}$)는 두 레코드간의 상이도 값이 된다. 이때 거리($d_{x,y}$)가 증가하면 비교하는 레코드는 속성값이 다른 것이 많다는 것을 의미하며, 거리($d_{x,y}$)가 낮아지면 속성이 유사한 것이 많다는 것을 의미한다. 또한, 알고리즘의 반복을 통해 레코드는 유사도가 높은 군집으로 할당된다.

---

k−modes algorithm

→ 입력 : N개의 범주형 속성 데이터, 초기치 k의 개수 지정

(1) 입력된 데이터에 대한 초기치 k들에 대한 $mode$선정

(2) $mode$와 레코드간의 상이도기반의 유사도거리 측정

(3) 유사도 값에 따른 군집 할당

(4) 군집의 $mode$ 갱신

(5) 군집의 변화가 없을 때까지 (2)~(4)단계 반복수행

→ 출력 : 최적 배정 k 군집

Figure 10. k-modes 알고리즘

k-modes 알고리즘도 초기치 개수 k개를 임의로 선정하는 문제점을 가지고 있다. 또한 속성의 값의 동일여부만을 판단하므로 속성의 중요도를 판단할 수 없다는 문제점을 가지고 있다.

2.2.3 개선 k-modes 알고리즘

개선 k-modes[오수민, 김철수, 2006] 알고리즘은 초기치의 선정을 유클리드 거리를 기반으로 선정하고, 이에 대한 유사도계산은 가중치를 적용하여 군집분석을 수행한다. 또한, 기존 방식은 단순 비교를 통한 거리기반방식으로서 데이터 속성 성질을 고려하지 않는다. 하지만, 개선 알고리즘에서는 특정한 속성에 가중치를 적용함으로서 군집들 간의 상이도가 커지는 효과가 있으며, 좀 더 효율적인 군집분석을 할 수 있다. 또한, 가중치는 초기 웨이트( )값 설정 시 사용함으로서 알고리즘 수행에 영향이 적도록 하였다.

개선 k-modes 알고리즘은 기존 k-modes 알고리즘을 변형한 형태로서 다음과 같이 표현된다.

$$(x,\ y_i) = \begin{matrix} 0 & (x_i = y_i) \\ 1 & (x_i \quad y_i) \end{matrix} \quad \begin{cases} 1+w_i & (x_i = y_i) \\ 0 & (x_i \quad y_i) \end{cases}, \quad i \in m, \quad x, y \in$$

개선 $k-modes\ \ algorithm$

→ 입력 : 개의 범주형데이터베이스, 초기치 선정k의 개수

Step 1.

1.1. 데이터에 대한 가중치 w계산 및 속성의 영향력(T)설정

1.2. 초기치k 선정

Step 2.

2.1 T 따른 k와 전체 객체간의 유사도 계산

2.2 유사한 군집에 데이터 할당

Step 3.

3.1 군집의 변화가 없을 때까지 (Step 2) 반복수행

3.2 변화가 없을 경우 정확도 계산 및 출력

→ 출력 : 최적 배정 k 군집

Figure 11. 개선 k-modes 알고리즘

2.2.4 Initial points refining algorithm

[Bradly P, 1998]이 제안한 초기치 선정방법을 개선한 *itial points refining algorithm*은 전체 데이터를 *subset*으로 요약하고, *subset*에 대한 군집분석을 수행하게 된다. *subset*은 *Bradly et al.*을 기반으로 군집분석을 수행하게 된다. 결국 *subset*의 선정에 따라 군집분석결과가 달라지므로, *subset* 선정이 매우 중요하다.

Initial points refining algorithm

step 1 : // sub-sampling

1.0 $CM = 0$

1.1 For $i = 1,...,J$

1.1.1 Let S be a small random sub-sample set of Data

1.1.2 Let SP_i be a randomly selected K sample from S_i

1.1.3 $CM_i = Clustering(SP, S_i, K)$

1.1.4 $CM = CM \cup CM_i$

step 2 : // Refinement

2.1 $FMS = 0$

2.2 For $i = 1,...,J$

2.2.1. Let $FM_i = Clustering(CM_i, CM, K)$

2.2.2. Let $FMS = FMS \cup FM_i$

step 3 : // Selection

3.1. $Let FM = ArgMin_{M}\ Distortion(FM_i, CM)\}$

3.2 Return (FM)

Figure 12. Initial points refining algorithm

선택되어진 *subset*에 대하여 반복수행을 통해 군집들을 병합, 삭제의 과정을 통해 최적의 군집분석 결과를 보여준다.

2.2.5 k-representative algorithm

[Thu-Hien Thi Nguyen, Van-Nam Huynh]이 제안한 k-representative algorithm은 군집의 센터를 정하는 방법을 개선한 알고리즘으로서 군집의 센터를 군집 중심의 빈도를 이용하여 계산하는 방법이다. 군집 가 Q q, ..., q_m로 정의 되었을 때, $q_j = (c_j, f_c) | c_j \in D_j \}$로 표현하게 된다.

---

k-representative algorithm

1. Initialize a k-partition of D randomly
2. Calculate k-representatives, one for each cluster.
3. For each X_i calculate the dissimilarities

   $d(X_i, Q_l), \quad l = 1, \ ..., \ k$

   Reassign X_i to cluster C_i

   Update both $Q_l, Q_{l'}$
4. Repeat Step 3 until no objet has changed clusters after a full cycle test of the whole data set.

Figure 13. k-representative 알고리즘

2.3 혼합형 데이터의 군집분석 알고리즘

2.3.1 k-prototype 알고리즘

[P.A. Vijaya, M. Narasimha, D.K. Subramanian, 2004]은 일반적으로 데이터는 수치형 속성과 범주형 속성으로 구분할 수 있으며, k-prototype 알고리즘은 수치형 속성에 대해서는 k-means 알고리즘을 적용하며, 범주형 속성에 대해서는 k-modes 알고리즘을 적용하고, 이 두 알고리즘을 결합하여 표현하였다. 또한, 두 알고리즘을 병합하는 방식으로 기존에 가지고 있는 초기치 개수 k를 선정하여 군집분석을 수행한다.

N, ..., N_n }개의 레코드들은 각각 A_1, ..., A_p, A_{p+1}, ..., A_m의 속성을 가질 때, p개의 수치형 속성과 $m-p$개의 범주형 속성을 포함한다. 이때 두 속성을 구분하여 연산을 수행하게 되며 거리계산은 다음과 같다.

$$d(x,y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=1}^{m-p} \delta(x_j, y_j)$$

---

k−prototype algorithm

→ 입력 : n개의 혼합형 데이터베이스, 초기치 선정 k

(1) 입력된 데이터에 대한 초기치 k개를 지정

(2) k와 데이터간의 속성별 구분(수치형 속성, 범주형 속성) 및 웨이트(γ) 적용, 유사도거리 측정

(3) 유사도 값에 따른 군집 할당

(4) 군집 갱신

(5) 군집의 변화가 없을 때까지 (2)~(4)단계 반복수행

→ 출력 : 최적 배정 k 군집

Figure 14. k-prototype 알고리즘

−*prototype* 알고리즘은 수치형 속성과 범주형 속성이 혼합된 데이터에 대해서 효과적으로 군집분석을 할 수 있는 장점을 가지고 있다. 하지만 기존 k−means 알고리즘과 k−modes 알고리즘에 대하여 단순히 혼합한 방식이므로 기존에 가지고 있는 초기치 선정 문제를 가지고 있다. 또한, 범주형 속성과 수치형 속성 간의 웨이트(γ)적용 방식에 대한 연구가 필요하다.

3 초기치 선정 방법

수치형데이터에 대한 분석기법은 k-means 알고리즘을 기반으로 다양한 기법이 개발되고 있으며 다양한 분야에서 활용되고 있다. 이들 알고리즘은 k개로 선정된 초기치를 기반으로 군집분석을 수행하게 되며, 초기치와 레코드간의 유사도를 계산하는 다양한 연구가 진행되어져 왔다. 하지만 초기치의 개수 k개를 결정하는 문제는 아직도 많은 연구가 필요하다. 즉, 임의의 데이터가 4개의 군집으로 이루어져있다면 초기치의 개수 k를 4로 정하고, 이 4개의 초기치를 어떻게 선택하느냐에 대한 연구는 다양하게 이루어지고 있다. 하지만 4개의 군집이라는 것을 모르는 경우, 초기치를 왜 4개로 정해야하는지에 대한 논의는 부족한 편이다. 특히, 컴퓨터를 기반으로 하는 기계 학습의 경우 임의의 대용량 데이터나 시계열 데이터처럼 실시간으로 발생하는 데이터에서는 최적의 k에 대한 결정은 기존의 방법으로는 한계를 가지고 있다. 본 논문에서는 최적의 k를 결정하는 방법과 데이터와 군집간의 유사도를 계산하는 기법을 제안하는 새로운 방법으로 크게 두 가지를 고려하였다. 첫째, 범주형데이터의 군집분석을 위한 새로운 초기치 선정을 위해 연구자의 자의적 결정을 최소화하면서 초기치를 선정하는 방법과, 초기치들 간의 유사도 계산 시 데이터의 속성을 고려하면서 유사도를 계산하는 방법을 제안한다. 이 두 가지 방법을 알고리즘에 적용하기 위해 몇 가지 고려해야할 점이 있다. 군집분석기법과 계층적 분류기법의 차이점, 기계학습, 범주형데이터의 수치적 표현 및 대용량 데이터에서의 샘플링 등이다.

· 군집분석기법과 계층적 분류기법의 차이점

군집분석과 계층적 분류기법은 둘 다 데이터를 구분한다는 점에서는 같은 목적을 가지게 된다. 분류기법은 의사결정에 있어 주어진 범주형 레이블을 기반으로 지도학습(supervised learning)으로 분석된다. 즉, 이미 결정된 결과를 목적으로 하여 최적의 결과를 도출하는 방식이며, 본 논문에서 다루고 있는 군집분석

은 자율학습(unsupervised learning)을 지향한다. 또한, 군집분석에서는 이미 분류된 데이터도 군집의 변화가 발생 할 경우 데이터의 정보를 갱신한다.

· 데이터 샘플링

데이터를 분석하는데 있어 표본분석은 모집단을 설명하는 가장 대표적인 방법이다. 일반적으로, 모집단이 매우 크며, 조사가 용의치 않은 경우 샘플조사를 하게 되며 이는 샘플의 반복추출을 통해 보다 정확히 모집단의 정보를 예측할 수 있다. 특히, 중심극한정리는 이러한 반복추출로 인해 발생하는 샘플의 정보가 정규분포에 가까워지고 있으며 이는 모집단의 정보를 대표한다는 점에서 매우 유용하다. 본 논문에서는 데이터의 레코드수가 적은 경우는 모든 데이터를 전수 분석하였으며, 레코드의 수가 많은 경우는 레코드를 100개씩 샘플을 수집하여 분석하였다.

· 범주형데이터의 수치화

일반적으로 범주형데이터의 수치화는 명목척도로서의 수치화표현이다. 즉, 범주형데이터를 수치화하는 것은 의미가 없어진다. 하지만, 속성 내에서 범주 값 빈도들의 분포는 그 속성을 대표한다고 할 수 있으며, 이러한 분포는 다항분포가 된다. 아울러 군집간의 유의성은 −test와 Fisher's exact probability test를 이용하여 분석하였다.

· 기계학습

기계학습은 훈련데이터를 통해 주어진 속성을 기반으로 결과를 예측하는 기법을 의미하며, 정해져있는 예측이 아닌 임의의 결과를 발견할 수 있으며, 일반화할 수 있다. 기계학습의 대표적인 예로서 신경망, 결정트리, 유전알고리즘, 선형분별분석, k−Nearest Neighbor, 퍼셉트론, SVM 및 EM 알고리즘 등이 있다.

· mode와 군집

mode는 하나의 데이터로도 표현할 수 있을 정도의 강한 유사성을 내포하고 있는 데이터들의 집합이다. 군집은 이러한 mode들의 병합과 갱신을 통해 생성

되는 mode들의 집합이 된다.

・Fisher's exact probability test

Fisher's exact probability test는 범주형데이터에 대한 교차분석에서 유용하게 활용되고 있다. A요인과 B요인이 독립이고 기대빈도가 5보다 작은 셀이 존재할 경우 일반적으로 Fisher's exact probability test을 활용하게 된다.

	B1	B2	
A1	a	b	a+b
A2	c	d	c+d
	a+c	b+d	N

$$= \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!\ \ !}$$

본 논문에서 다루고 있는 범주형데이터에 대한 분석에서 군집분석의 주요한 논의 중 하나는 이 Fisher's exact probability test을 이용하여 mode간의 유의성을 검정하고, 그 결과를 유사도 값으로 정의하였다.

3.1 제안 알고리즘

범주형 속성으로 이루어진 개 레코드 집합 D $[x$, x_2, x_3, ..., x $]^T$에 대하여 $d_i = x_{i1}, x_{i2}, x_{i3}, ..., x_{ip}\}$는 $i(i \in N)$번째 레코드이며, i번째 레코드는 p개의 범주형 속성$\{C_1, C_2, C_3, ..., C_p\}$으로 구성되어 있다. p개의 범주형 속성에 대하여 $C_r,(r \in p)$는 다양한 범주형 속성 값을 가지고 있다.

$$D = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}$$

$$C = [C_1 \ C_2 \dots \ C_p]$$

$$d_i = [x_{i1} \ x_{i2} \dots \ x_{ip}]$$

$$\text{성값} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \lfloor & & & \end{bmatrix}$$

c_{ij} : $i(i \in N)$번째 레코드의 $j(j \in p)$번째 범주형 속성의 값

임의의 레코드 d_i는 p개의 속성에 대하여 다양한 속성 값으로 구성되어있다. 예를 들어, c_{i1}은 레코드의 첫 번째 범주형 속성(C_1)의 속성 값들 중 i번째 범주형 속성 값을 의미한다.

군집분석에서 연구자는 초기치 k를 임의로 정하게 된다. 알고리즘에 따라 군집의 수는 적절히 조절되기도 하지만, 결국은 연구자가 정한 k의 영향을 받게 된다. 이러한 방식은 기존 데이터의 정보가 제공되는 경우에는 매우 유용하다. 하지만, k개의 초기치가 어떻게 선정되느냐에 따라 분석결과가 크게 달라질 수 있음을 의미한다. 또한, 대용량의 데이터 및 실시간으로 생성되는 데이터에 대해서 k의 개수를 미리 정하는 방식은 효율적이라고 할 수 없으며, 새로운 정보가 발생했을 때 유연하게 대처할 수 없다.

본 연구에서는 대용량 데이터 및 실시간 데이터 등에서 데이터에 대한 정보를

전혀 알 수 없는 경우를 고려한 범주형데이터로 하였다. 만약, 분석하는 모든 데이터가 하나의 군집으로 요약될 경우, 초기치 값은 한 개만 있으면 된다. 임의로 선정된 한 개의 초기치는 어떤 레코드가 선정되더라도 군집에 영향을 주지 않는다. 군집이 하나 이상으로 이루어진 데이터일 경우 초기치를 새롭게 생성시켜 새로운 군집의 초기치로 정한다. 따라서 본 연구에서는 연구자가 임의로 선정하는 최초의 초기치의 개수를 1개로 하였으며, 초기치를 기반으로 여러 개의 초기치 mode들을 생성시키고, 이를 병합, 갱신하면서 군집의 결과를 최적화함으로써, 군집의 개수를 추정하였다.

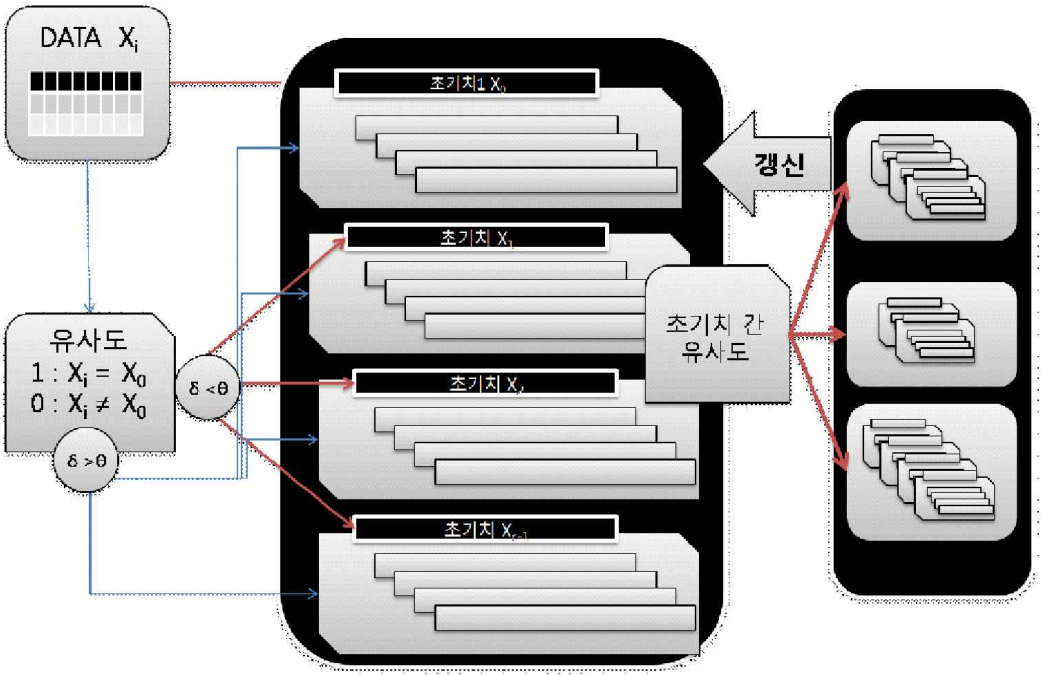


Figure 15. 제안 알고리즘의 전체과정

Figure 15는 제안 알고리즘의 전체적인 과정을 보여주고 있다. 데이터 에서 첫 번째 초기치 레코드 $_1(k_1 = d_i \in D)$을 임의로 선정하고 $d_j(d_j \in D)$와 k_1의 유사도( )를 비교한다. 유사도 값이 한계기준 θ보다 크다면 레코드 d_j 는 k_1과 동일한 mode를 구성하며, θ보다 작으면 새로운 mode를 생성시킨다. 데이터의 모든 레코드들은 $(m \in K)$개의 mode들과 비교하여 유사하면 해당 mode에 소속되

며, 모든 과 비교하여 유사도가 낮으면 k_{m+1}번째 mode를 생성한다.

k-means 알고리즘에서는 군집의 중심을 레코드와 군집간 거리의 평균을 이용하지만, 제안 알고리즘에서는 해당 속성의 최빈값의 범주형 속성 값을 각 속성의 대푯값으로 하여 mode를 선정한다. 이후 과정에서 한계기준 θ는 군집의 유사도를 조절할 수 있는 기준으로 이용되며, 한계기준 θ를 1에 근접한 값으로 설정하면 초기 군집들은 내부적으로 강한 유사도를 지니게 되며, 한계기준 θ를 0에 가깝게 설정하면 각각의 초기 군집의 크기는 커지고, 약한 유사성을 보이게 되지만, 초기치 군집의 개수는 적어진다.

3.2 한계기준

3.2.1 한계기준 의 선정

일반적으로 모집단이 매우 클 경우 데이터를 전부 조사하는 방법은 비효율적이거나 불가능한 경우가 많다. 이에 대하여 본 연구에서도 주어진 데이터를 전부 사용하지 않고, 임의의 표본을 추출하여 이를 기반으로 군집분석의 초기치 개수의 추정과 군집분석을 수행하였다. 표본은 균등분포(U(0,1))를 기반으로 m개의 데이터를 추출하였다. 이때 m개의 데이터에 대하여 두 개의 데이터를 임의로 추출하고, 이에 대한 k-modes 알고리즘의 유사도 δ를 구하였다. 이를 반복하여 유사도계산 결과 집합 $\delta = \delta_1, \delta_2, \quad , \delta_i\}$를 구하고, 이에 대한 평균유사도 δ_{mean}을 계산한다. δ_{mean}은 초기 mode를 정하는 한계기준 θ가 된다. 또한, δ_{mean}은 분석하는 데이터의 전반적인 유사도를 알 수 있는 정보가 된다. 예를 들어, 속성이 22개인 mushroom 데이터의 8124개의 데이터들 중 300개를 표본으로 샘플링하고, 표본 개수의 30%정도에 대한 유사도를 계산한 경우 δ_{mean}=18.5이면, 한계기준은 19로 정한다. 결국, 비교하는 데이터와 mode에서 19개의 속성이 동일하면 같은 mode에 소속되며, 이보다 낮으면 새로운 mode를 발생하게 된다. 즉, 군집의 개수를 결정하는데 주요한 역할을 하는 한계기준 θ를 연구자가 자의적으로 정하지 않고 데이터 자체적으로 발생하는 값을 이용함으로서 분석하는 데이터에 대한 상대적이고 합리적인 기준이 된다. 이후 한계기준 θ는 제안 알고리즘의 유사도 계산에 활용된다.

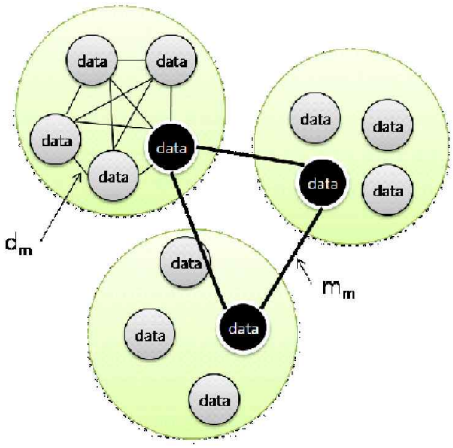


Figure 16. 군집내 유사도( )과 군집간 유사도(m_m)

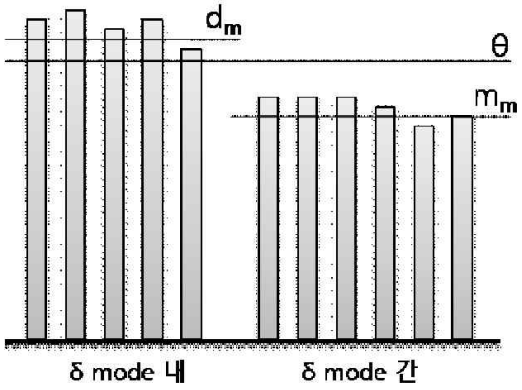


Figure 17. d_m과m_m에 따른 θ의 상대적 위치

Figure 16에서는 mode의 구조를 보여주고 있다. 데이터의 레코드는 같은 mode내에서는 강한 유사도를 보이며, 다른 mode의 레코드들과는 약한 유사도가 나타나야 한다. 이때, mode내의 레코드를 샘플링 하여 평균 유사도 을 구할 수 있으며, 이때, 각 mode별로 d_m값은 일반적으로 비슷하게 나타난다. d_m을 결정하는 속성의 종류가 각 mode별로 차이가 있으나 전체적인 유사도는 비슷하게 나타난다. m_m은 mode간의 유사도 평균이다. 각 mode들은 대표 mode를 가지게 된다. 속성의 값이 수치형 값이 아니므로 속성의 평균을 할 수 없기 때문에 속성 값의 빈도가 가장 높은 속성 값을 그 속성의 대표속성으로 하며 모든 속성별로 대표속성을 모아서 해당 mode의 대표 mode가 되며, 이 때 각 대표 mode가 임의의 레코드가 되지는 않는다.

한계기준 θ는 초기 생성되는 mode들의 유사도 및 초기치 mode의 개수를 결정하는 한계기준의 기능을 수행한다. 많은 연구에서 이러한 θ의 값을 경험적 수치로서 정의하고 있다. 제안 k-modes 알고리즘에서 유사도 계산은 속성값이 동일한지에 따라 1과 0으로 구분한다. 두 속성의 값이 같다면 1, 다르면 0으로 하므로, 두 레코드의 유사도는 최소값은 0에서 최대값 22를 가지게 된다. 또한, 만약 두 레코드가 유사하다면 유사도 값이 증가하게 된다. 한계기준 θ의 값을 높게 설정하면 레코드간의 mode의 유사도는 강하게 군집되며, 한계기준 θ의 값을 낮게 설정하면 mode의 유사도는 낮아지게 된다. 본 연구에서는 이러한 한계기준

의 값을 임의의 값이 아닌 데이터의 속성들 간의 평균유사도를 기반으로 설정하는 방법을 제안한다. 대표 mode들 간의 k-modes 유사도 결과 값은 상이도가 증가하므로 유사도 값은 낮아지게 되며, 평균 유사도가 m 이 된다. 일반적으로 m_m은 d_m 보다 낮게 나타난다. 군집의 갱신과정에서는 한계기준이 필요하다. 즉, 주어진 한계기준에 대하여 분석을 더 수행해야 하는지 여부를 결정해 줄 기준 값이 필요하다. 기존의 한계기준은 연구자가 임의로 정하였으나, 본 논문에서는 한계기준을 θ로 정의하고 이를 d_m과 m_m을 이용하여 정의함으로서 연구자의 자의적 결정을 배제할 수 있다.

한계기준 θ는

$$\theta = m_m + (d_m - m_m) \times \frac{d_m}{m_m}$$

로 계산한다.

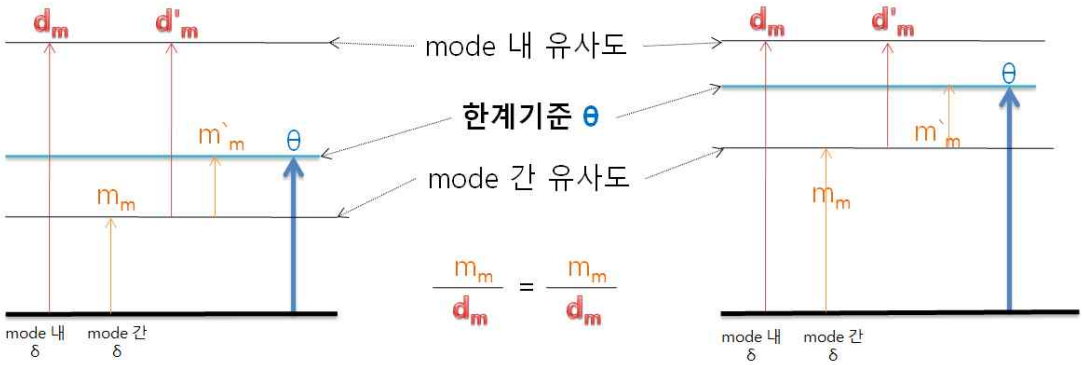


Figure 18. 군집간의 유사도 따른 의 상대적 위치

Figure 18은 d_m과 m_m에 대한 상대적 θ의 상대적 위치를 보여주고 있다. 만약 군집내 레코드의 평균 유사도(d_m)가 높고 군집간 유사도(m_m)가 낮다면, 한계기준 θ의 값도 낮아지게 된다. 또한, 군집내 레코드의 평균 유사도(d_m)가 높고 군집간 유사도(m_m)도 높으면 한계기준 θ의 값은 증가하게 된다. 즉, 레코드들의 유사도가 높으면 θ도 자연스럽게 높은 값의 한계기준을 정하게 됨으로서, 한계기준 θ는 d_m과,m_m의 값의 특성에 따라 상대적 위치로 결정된다.

3.2.2 mode의 병합 및 갱신

한 레코드와 다른 레코드간의 유사도를 비교하면, 그 유사도 값들은 다양하게 나타날 것이며, 이는 상황에 따라 다양한 분포를 보인다. Figure 19에서는 k-modes 알고리즘의 유사도계산 결과에 대한 한계기준을 보이고 있다. θ 분위수는 레코드와 레코드간의 유사도 값에서 유동적으로 변하며 해당 레코드와 θ 분위수 이상 유사할 경우 같은 mode가 되며, 그 이하의 유사도에서는 새로운 mode로 정의된다.

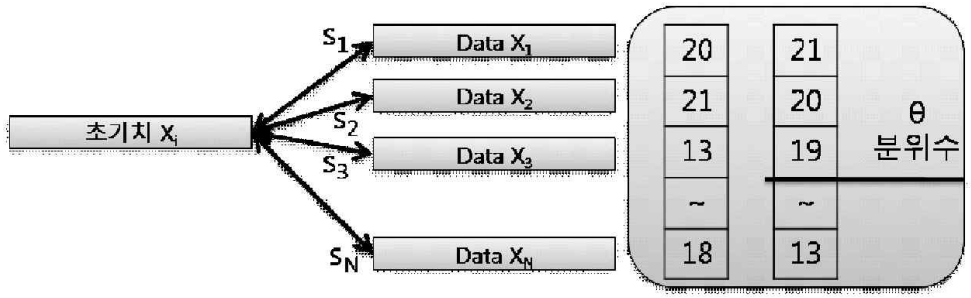


Figure 19. 한계기준 와 레코드 간 유사도비교

Figure 19에서는 초기치 군집과 레코드간의 유사도에 대하여 θ분위수의 한계를 보여주고 있다. 레코드들은 해당 군집과 유사도를 비교한 후 θ분위수보다 유사도 값이 크다면, 이는 같은 군집으로 병합하게 된다. 즉, 한계기준 θ보다 유사도가 크다는 것은 비교하는 두 레코드가 유사한 점이 많음을 의미한다. 이는 개선 k-modes 알고리즘을 보다 효율적으로 개선시킨 방법으로 제안 알고리즘의 기반이 된다.

	=98% 정확도	θ=95% 정확도	θ=90% 정확도	$\theta(d\ ,s_m)$ 정확도	θ=80% 정확도
θ=98% mode개수	-0.322				
θ=95% mode개수		0.227			
θ=90% mode개수			-0.402		
$\theta($ $,s_m)$ mode개수				0.387	
θ=80% mode개수					0.378

Table 1. 한계기준 에 따른 초기치 정확도와 mode의 개수와의 상관관계

Table 1에서는 한계기준 θ값에 따른 초기 mode의 분류 정확도와 초기치 mode 개수와의 관계를 보여주고 있다. 초기 mode의 분류 정확도는 초기치가 많을수록 정확도는 높아지게 되지만 적절한 군집으로 분류되었을 때도 분류의 정확도가 유지되어야 하므로, 적절한 초기치의 선정이 매우 중요하다. 본 논문에서는 d_m과 m_m을 이용한 θ분위수를 가장 최적화 시킬 수 있는 기준으로 하였다. 두 레코드가 θ이상 동일하다면 이는 최종적으로는 동일 군집으로 판단한다. $(1-\theta)\%$의 유의성이 존재할 수 있으며, 레코드들은 mode의 갱신과정에서 좀 더 최적의 mode가 있다면 갱신되어진다.

3.3 초기 mode의 유사도 및 갱신

3.3.1 기존 k-modes 알고리즘의 개선

초기 mode들은 개의 최적의 군집으로 병합, 갱신 하는 과정이 필요하다. 일반적으로 초기 mode들은 $(s \in p)$번째 속성에 대하여 C개의 속성 값을 가지고 있으며, 이것은 C_s개의 속성 값이 가지는 확률 $p_1, p_2, \ldots, p$ 로 나타나는 다항분포이다. 즉, 각각의 mode들은 p개의 다항분포를 포함하게 된다. 각각의 초기 mode들은 속성별로 다항분포를 취하게 되며 초기 mode간의 유사도는 각 속성별 다항분포의 유의성의 합으로 정의한다. 비교하는 두 초기 mode의 s번째 속성에서 C_s개의 속성 값들을 포함하므로, 각 속성 값들에 대한 비교가 가능하다. 만약, 두 초기 mode의 $s(s \in p)$번째 속성 값에서 유의한 차이가 없다면 유사한 mode로 판단할 수 있으며 차이가 크다면 상이한 mode로 판단한다. 본 논문에서는 속성의 빈도에 대한 비교는 Fisher's exact probability test를 수행하였으며 mode간의 유사도는

$$\delta_{ij} = \begin{matrix} 1, & p_{ij} > \alpha(5\%) \\ 0, & p_{ij} < \alpha(5\%) \end{matrix}, \quad i,j \in K$$

이다.

실제 두 가지 속성을 비교할 경우 제안 알고리즘의 유사도 계산은 다음과 같다.

$$(x,y) = \sum_{=1}^{2}, \quad \delta(x,y)_{jj'} = \begin{cases} 1 & , \ \chi^2_{(x_j,x_{j'})=(y_j,y_{j'})} \\ 0 & , \ \chi^2_{(x_j,x_{j'}) \quad (y_j,y_{j'})} \end{cases}$$

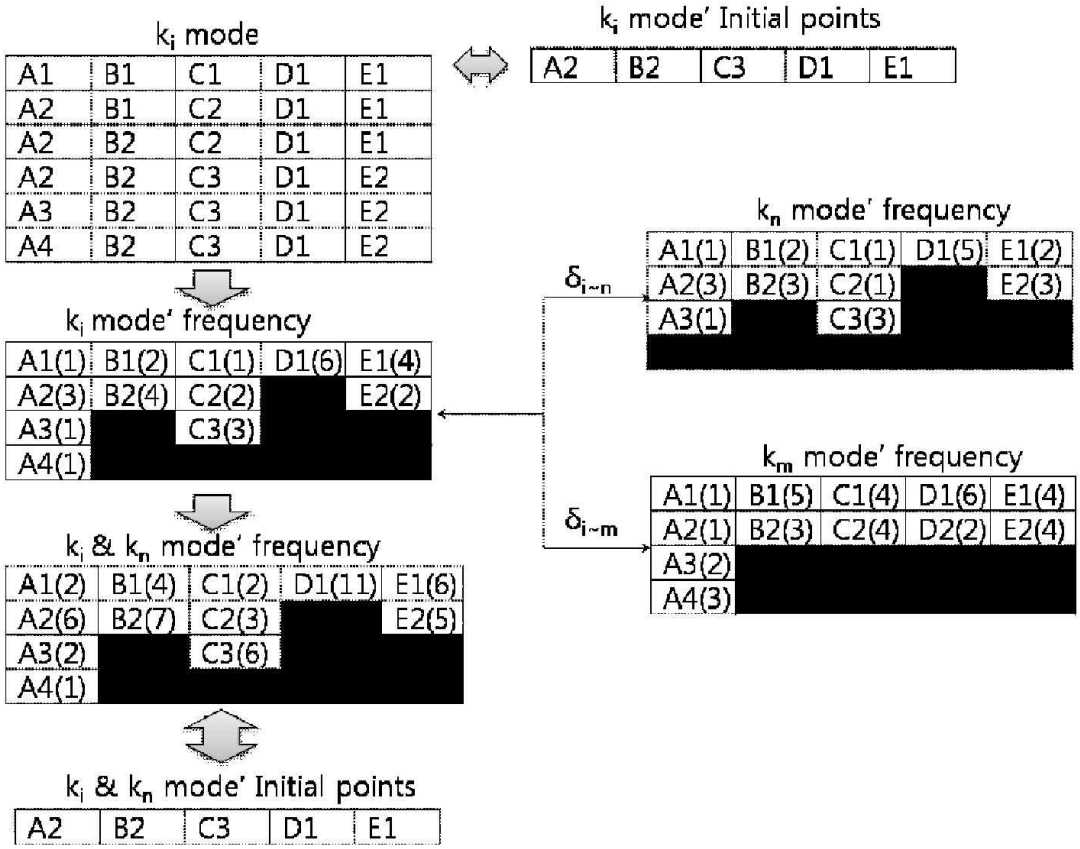


Figure 20. mode의 속성 값과 mode의 비교 및 병합

Figure 20에서 $,k_j(i,j\in$)를 기반으로 하는 초기 mode k_i와 k_n, k_m $(n,m\in K)$간의 비교과정 보여주고 있다. k_i~k_n에서는 모든 속성에서의 빈도들이 비슷한 분포를 보이므로 k_i와 k_n은 유사하다는 결론을 내리고 병합하게 된다. 병합된 k_i와 k_n은 새로운 mode가 되며 $\{A2,\ B2,\ C3,\ D1,\ E1\}$의 속성값으로 이루어져 있다.

3.3.2 Chain k-modes 알고리즘

지금까지는 각 mode에 대하여 각 속성별 비교만을 수행하였다. 하지만 범주형데이터의 특성상 범주형데이터의 연관성은 군집의 특성을 대표하므로 매우 중요하다. 즉, 자동차라는 범주에 대하여 고가의 자동차를 구입하는 사람들이 집이라는 범주에서는 고가의 집을 보유하고 있을 확률은 당연히 높아진다. 이러한 범주들이 수치형 속성일 경우에는 각 속성에 대한 상관분석 및 다중회귀분석 혹은 로지스틱 분석 등 다양한 방법으로 분석할 수 있다. 하지만 데이터가 범주형으로 이루어져 있으며, 속성이 다양하게 나타나고 군집에 대한 정보가 부족할 경우 기존의 방법으로는 해결할 수가 없었으나, [오수민, 송준모, 김철수, 2012]에서는 범주형 속성을 Chain으로 연결하여 다양한 범주형 속성들을 군집분석 할 수 있었다. 본 연구에서는 범주형 속성간의 Chain을 구성하여 비교함으로서 k-modes 알고리즘의 군집분석 방법을 개선한 Chain k-modes 알고리즘을 제안한다.

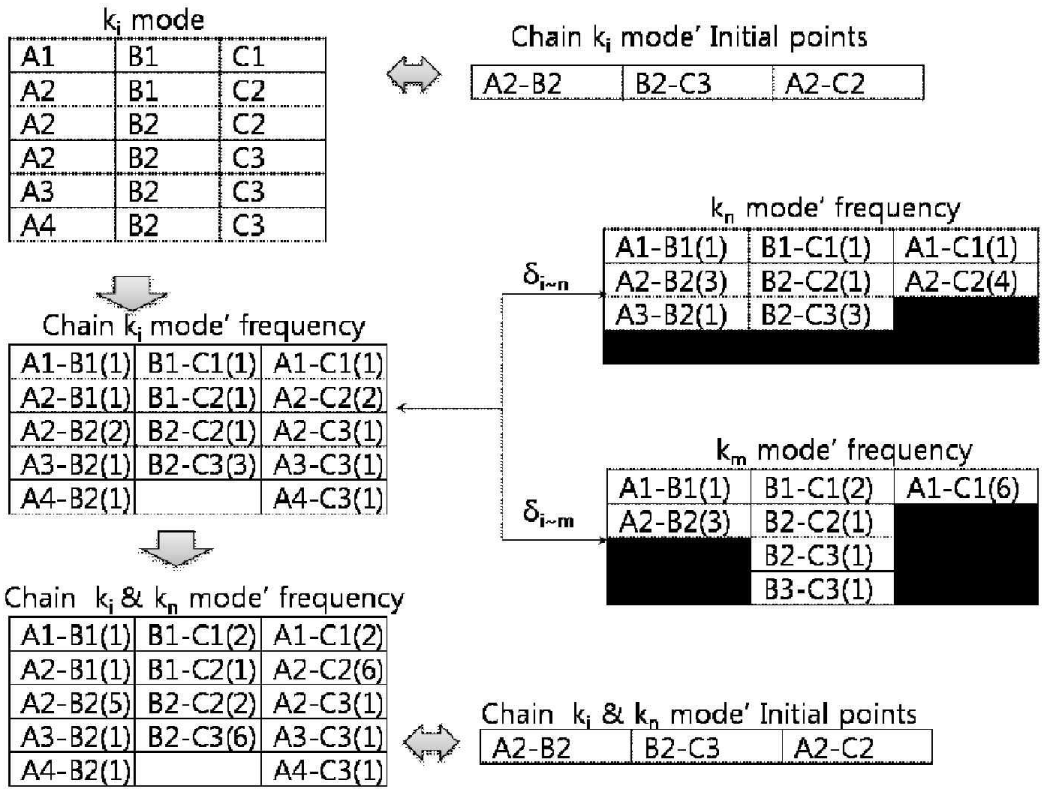


Figure 21. Chain mode의 속성 값과 mode의 비교 및 병합

Figure 21은 Chain mode의 속성 값과 mode의 비교 및 병합을 보여주고 있다. Chain k-modes 알고리즘은 속성들을 Chain으로 연결하여 각 연결에 대한 빈도를 알 수 있다. 즉, 기존의 방법이 자동차 범주와 집이라는 범주를 서로 독립적인 관계로 보고 있다면, Chain 방법은 자동차와 집의 두 범주를 동시에 확인할 수 있으며, 이러한 체인간의 빈도 역시 다항분포를 취하게 된다. 또한, 이를 응용하면 3개 이상의 Chain도 구성이 가능하다. 하지만 Chain을 구성할 경우 기존에는 데이터가 p개의 속성을 포함할 경우 p회의 유사도 계산과정이 필요하다. 또한 비교하는 초기 mode의 개수가 개일 경우 최종적으로 np회의 계산과정이 필요하다. 하지만 Chain의 경우 각 데이터 당 $_2 = p!/2!(p-2)!$개 Chain이 발생한다.

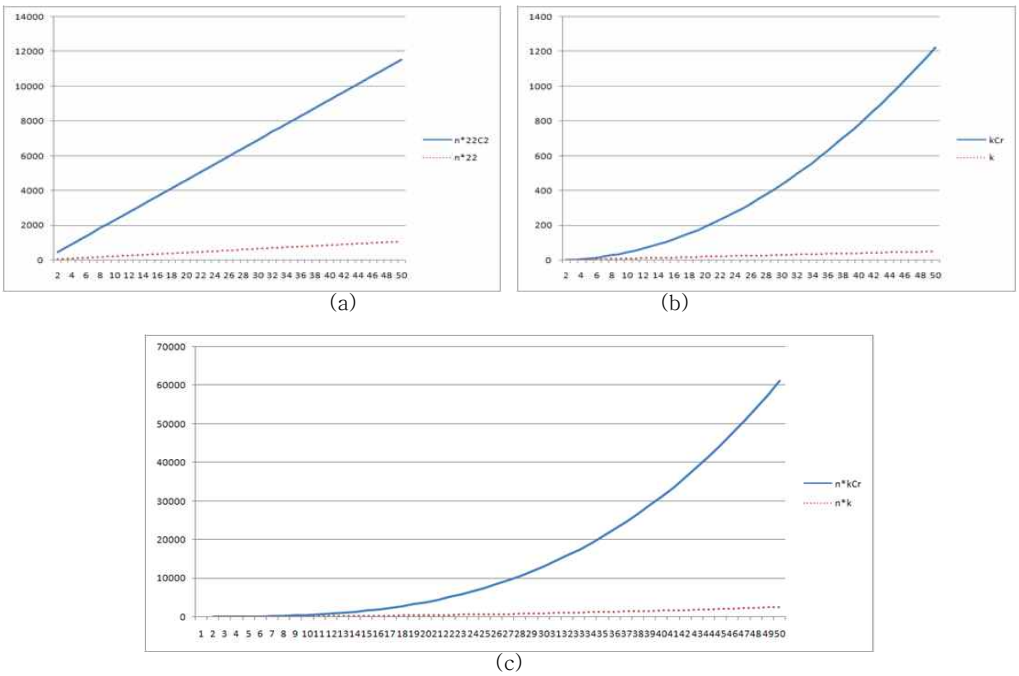


Figure 22. 기존 알고리즘과 제안 알고리즘의 수행횟수 비교

Figure 22는 기존 알고리즘과 제안 알고리즘의 수행횟수를 비교한 결과이다. 제안 알고리즘은 p개의 속성에 대하여 비교하는 속성이 2개일 경우 발생하는 체인은 기존 알고리즘에서 비교해야하는 회수 보다 상대적으로 많은 비교를 요구

한다. 이는 단순 반복횟수의 증가분이므로 병렬처리 시스템 등을 이용할 경우 해결할 수 있다. 또한, 반복횟수의 조절은 실제 모든 Chain을 전부 확인할 필요는 없으므로 유용한 속성에 대한 Chain만을 비교하였을 경우에도 결과값에는 영향을 주지 않는다.

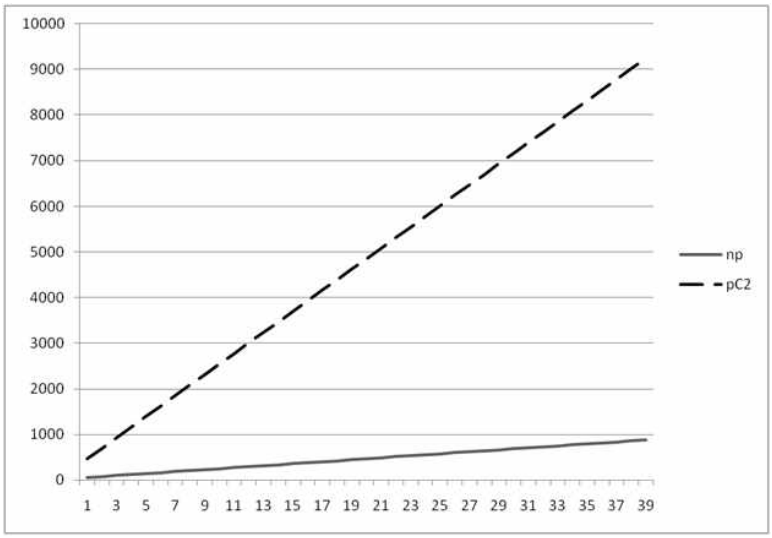


Figure 23. mode개수에 따른 기존 k-modes와 Chain k-modes의 반복회수 비교

Figure 23는 Mushroom data set에서 속성의 개수 22개에 대한 초기 mode의 개수에 따른 반복회수를 보여주고 있다. 기존 k-modes 알고리즘 방식으로는 속성의 연관성을 고려하지 못하는 반면 mode의 개수가 증가할수록 수행속도는 완만하게 증가하고 있으며, 속성의 연관성을 고려한 Chain k-modes는 비교하는 mode의 개수가 증가할수록 상대적으로 많은 계산량을 요구한다. 본 논문에서는 이러한 성질을 이용하여 두 가지 방법을 모두 고려하였다.

원본 데이터에서 초기치 mode가 다량으로 생성될 경우에는 기존 k-modes 알고리즘을 수행하고, 초기치 mode가 병합 갱신하는 과정에서 mode의 개수가 10개 이하의 mode로 갱신될 경우부터는 Chain k-modes 알고리즘을 활용하여 보다 엄밀히 군집분석을 수행한다. 또한 많은 Chain을 효율적으로 비교하기 위하여 유사도 계산을 -test가 아닌 Fisher's exact probability test로 하였다. 결국, 제안하는 알고리즘의 유사도 계산방식은 p개의 속성에 대하여 2개의 속성

의 빈도를 동시에 비교하는 방식이며 이에 대한 유의성 검정을 통해 유사도를 계산한다.

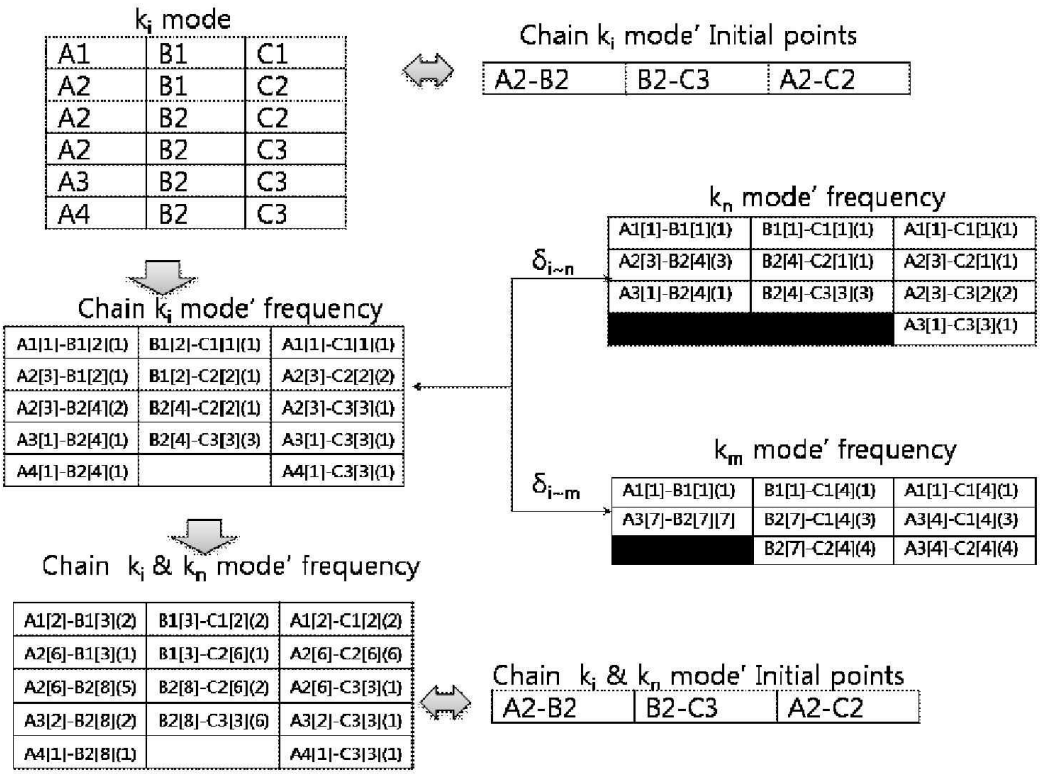


Figure 24. Chain k-modes 알고리즘에서의 빈도 비교

Figure 24에서는 두 개의 연결된 속성에 대하여 발생하는 Chain들을 모두 비교하는 것이 아니라 각 체인에서 발생하는 Chain 중 빈도가 가장 높은 Chain의 빈도를 이용하여 Fisher's exact probability test을 한다. 예를 들어, 의 첫 번째 항목의 A2[3]-B2[4](2)는 A속성 중 A2요인이 3개 있으며, B2요인이 4개 있는 항목과 연결된 경우가 2가지로 나타나므로, k_i의 mode는 속성 A, B에 대하여 A2[3]~B2[4]가 연관성이 높은 Chain이라는 것을 의미한다. 이에 대하여 k_n, k_m에서 A2와 B2의 빈도는 k_n은 3과 3이고 k_m은 0과 7이다.

와 k_n, k_m에 대하여 A2, B2 속성에 대한 Fisher's exact probability test 검정결과는 다음과 같다.

	k_i	k_n
A2	3	3
B2	4	3

$p-value = 1.000$

	k_i	k_m
A2	3	0
B2	4	7

$p-value = 0.045$

즉, A2와 B2로 이루어진 Chain에 대하여 k_i는 k_n에 대한 유사성이 더 높다는 것을 알 수 있다. 이러한 방식으로 모든 Chain들에 대하여 유사도를 계산하고, 알고리즘의 반복과 갱신을 통해 각 레코드의 소속 mode의 변화가 없을 때까지 알고리즘을 수행하면, 해당 mode를 최적화 된 군집으로 판단하고 알고리즘을 종료하게 된다.

3.4 초기치 개수 선정 및 종료조건

3.4.1 초기치 개수의 선정

초기치의 개수는 군집분석에 있어서 가장 중요하다. 초기치의 개수를 몇 개로 하느냐에 따라 군집분석의 결과가 완전히 변하거나 군집의 정확도가 큰 폭으로 변하게 된다. 많은 연구에서 선정된 초기치를 최적화 하는 방법을 제안하고 있으나, 몇 개의 군집으로 수렴되는지에 대한 논의는 미흡하다. 본 연구에서는 다양한 초기치에 대하여 새로운 비교방법인 Chain k-modes 알고리즘, 한계기준와 Fisher's exact probability test를 기반으로 mode의 병합 및 갱신과정을 통해 최적의 군집으로 수렴되도록 하였다.

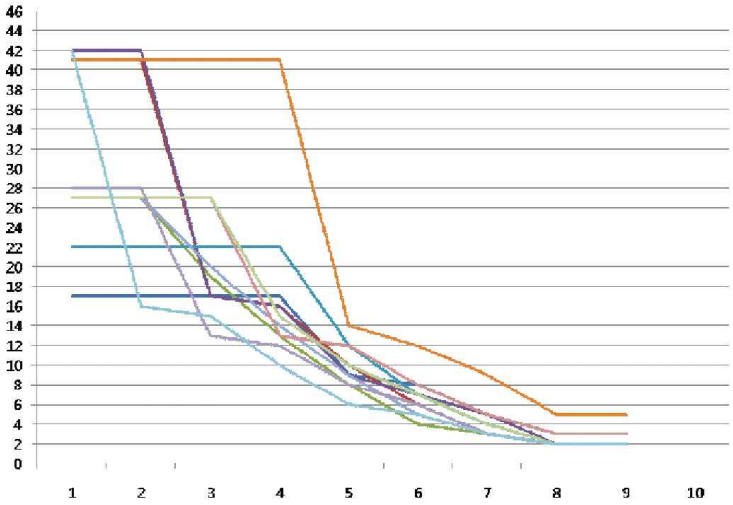


Figure 25. 제안 알고리즘의 초기치 수렴

Figure 25에서는 제안 알고리즘을 Mushroom data에 적용하여 초기치 mode들이 생성, 병합 및 갱신되는 과정을 보여주고 있다. 대부분의 결과에서 mode들은 2개로 병합되고 있으며, 몇몇 실험에서는 군집이 3개 혹은 5개로 군집화 되

는 경우를 볼 수 있다. 하지만 대부분의 실험에서 2개의 군집으로 되는 경우가 많으므로, 우리는 데이터가 2개의 군집으로 이루어져 있으며, 초기치의 개수는 2개로 결정 할 수 있었다. 즉, Mushroom data가 2개의 군집으로 이루어져 있을 확률이 높음을 의미하며, 실제 Mushroom data는 독이 있고, 없음의 2가지로 구분된다.

3.4.2 종료조건

제안 Chain k-mode 알고리즘의 종료조건은 두 가지 조건으로 발생한다. 첫째로 고전적인 조건으로서, 기존 k-mode 알고리즘의 종료조건과 동일하다. 즉, 레코드의 소속군집이 반복에 따른 변화가 없을 경우, 군집분석기법은 분석이 완료 되었다고 판단한다. 본 연구에서도 종료조건은 동일하지만, 추가적인 조건이 포함되게 된다. 둘째 조건은 군집분석과정에서 발생하는 과 m_m으로 판단하는 것이다. d_m은 군집 내 유사도 값으로서 군집내 유사도의 정도를 파악할 수 있으며, m_m은 군집간 유사도로서 군집간 상이도를 알 수 있다. 일반적으로 d_m은 m_m보다 항상 높게 나타나야 한다. 왜냐하면, 군집분석의 기본 조건이 군집내 레코드들 간에는 높은 유사도가 나타나야 하고, 서로 다른 군집간 레코드에서는 낮은 유사도(높은 상이도)가 나타나야 하기 때문이다.

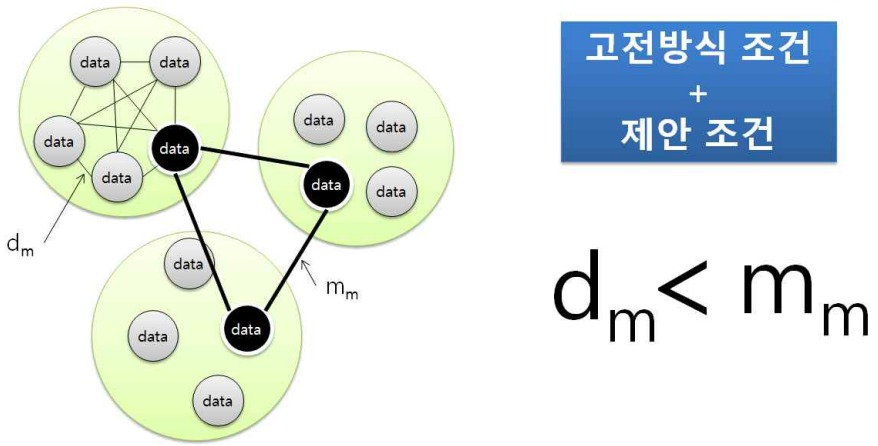


Figure 26. 제안 chain k-modes 알고리즘의 종료조건

제안 알고리즘은 데이터를 여러 개의 mode들로 분리하고 다시 병합 갱신하는 과정을 하면서 최적의 군집을 추정하게 되며, 이때 너무 많은 군집을 병합하다보면 상이한 군집도 병합할 수 있다. 만약, 병합된 군집들의 군집내 유사도들과 군집간 유사도를 비교하여 군집간 유사도가 군집내 유사도보다 높게 나타나면 군집분석기법은 수행과정을 종료하게 된다.

3.5 Chain k-modes 알고리즘

ain k-modes algorithm

STEP I.

1. 랜덤 초기치 mode k 생성
2. $k_1 \sim d_i (d_i \in data)$간의 k-modes 알고리즘을 이용한 유사도(δ) 계산
    - 2.1 유사도 한계기준 θ정의
    - 2.2 샘플 데이터의 유사도 δ 측정
        - 2.2.1 δ θ : 기존 mode에 포함
        - 2.2.2 $\delta < \theta$: 새로운 mode로 정의
3. m개의 군집 k_1, k_2, ..., k_m 생성
    - 3.1 초기 mode 병합
    - 3.2 초기 mode 갱신

STEP II.

4. 초기 mode $k_i \sim k_j$의 유사도(δ_{ij}) 계산
    - 4.1 각 mode를 기반으로 속성의 빈도
    - 4.2 속성의 Chain을 적용한 유사도 계산
    - 4.2 Fisher's exact probability test
    - 4.3. K개의 군집으로 갱신
    - 4.4 종료기준 : 군집내 유사도 평균 < 군집간 유사도 평균
5. 최적의 군집화 완료

Figure 27. 제안 Chain k-modes 알고리즘

4 실험결과

본 논문에서는 Chain k-modes 알고리즘의 성능을 확인하기 위하여 실험 데이터를 생성하여 정확도를 확인하였으며, UCI Machine Learning Repository에서 제공하는 Mushroom dataset과 Small Soybean dataset을 대상으로 실험하였다. 분석도구는 R-program(3.0.1)을 활용하였다. 실험 데이터의 생성은 군집의 개수가 2개와 3개인 경우 및 4개인 경우에 대하여 군집의 혼합된 정도를 조절하면서 확인하였다. Mushroom dataset과 Soybean data에 대해서는 기존 k-modes 알고리즘과 개선 k-modes 알고리즘, 제안 알고리즘을 비교하였다. 기존 k-modes 알고리즘은 초기치 mode의 개수를 알고 있는 경우 임의로 선정된 초기 mode를 기반으로 군집분석을 수행하며, 개선된 k-modes 알고리즘은 초기치 mode의 정보를 알고 있는 경우, 좀 더 최적화된 mode를 선정하여 분석하는 방법이다. 본 논문에서 제안하는 방법은 군집의 개수가 알려지지 않았을 경우를 한 개의 초기치를 기반으로 최적화된 군집으로 군집하는 방법이다.

4.1 Simulation

시뮬레이션은 R-program을 사용하였다. 실험은 임의로 k개의 다차원 정규분포를 발생시킨 후, 이를 제안 Chain k-modes 알고리즘으로 군집분석을 하여 실제로 k개의 군집으로 잘 군집하는지를 확인하였으며, 군집의 개수가 다른 경우 이것이 의미하는 것이 무엇인지 분석하였다.

먼저, 다차원 정규분포를 발생시키기 위하여 R에서 추가 패키지 중 "mvtnorm(On Multivariate t and Gauss Probability in R)"을 이용하였다. 본 연구는 범주형 속성에 대한 군집분석을 목표로 하므로 수치형으로 발생하는 정규분포를 범주형데이터로 변환해야한다. 이에 대해서는 정규형으로 발생된 모든 데이터 값을 가우스 함수(Gaussian function)를 처리하여 정수로 표현 함으로서, 이를 범주형데이터로 정의하고 분석하였으며, 데이터의 속성은 정규분포의

차원이 된다.

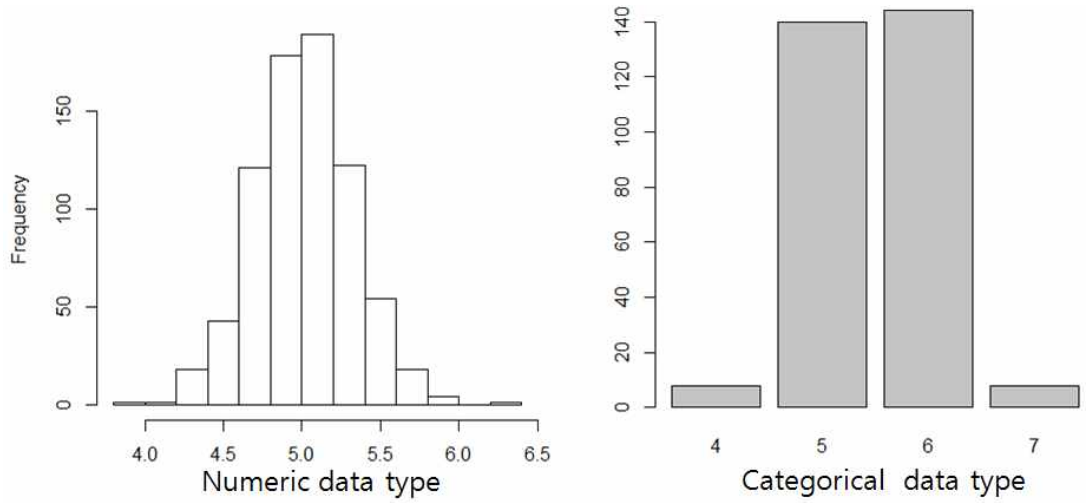


Figure 28. 수치형데이터와 범주형데이터의 비교

Figure 28은 수치형데이터를 범주형데이터로 변환한 결과를 비교하고 있다. 먼저, R 소프트웨어의 mvtnorm 패키지를 이용하여 평균이 5이고 표준편차가 0.3인 정규난수 300개를 발생시킨다. 난수들은 3~7사이의 값이 발생하며, 이는 연속형 값으로 표현된다. 발생된 난수에 대해서는 수치적 계산이 가능하며, 범주적 개념은 아직 적용할 수 없다. 이때, 난수들에 대하여 가우스함수를 적용하면 모든 난수들은 4~7사이의 정수값으로 표현된다. 물론 가우스 함수에 의해서 변형된 난수들도 수치형이다. 하지만 이를 범주형 자료에 대한 명목형 변수와 동일하다. 만약, 4=A, 5=B, 6=C, 7=D라고 하면, 결국, 5,0.3)을 따르는 정규분포의 수치형 데이터의 속성은 속성값이 주로 B와 C의 빈도가 많은 범주형 데이터의 속성으로 변환된다. 이러한 방법으로 다변수 정규분포로부터 여러 개의 범주형 속성으로 구성된 범주형 데이터를 발생시킬 수 있다.

N	1차원	2차원	3차원	4차원	5차원	6차원	7차원	8차원	9차원	10차원	소속군집
1	3.4	5.7	9	8.8	9.3	11.1	10	10.9	13.7	13.8	1
2	4.1	5.3	6.8	8.1	8.6	11.3	11.3	11.3	12.5	14.9	1
3	5.7	5.7	7.2	7.7	9.2	9	10.3	13.3	12.8	15	1
4	4.8	6.1	8.1	7.8	10.4	9.2	11.4	13.1	13.5	14.4	1
5	6.5	6.3	6.6	9.3	9.3	10.7	11.9	11.7	13.9	13.4	1
6	5	6.8	7.6	8.5	10.2	9.1	10.8	12.6	13.5	13.9	1
7	4.8	4.5	7.4	9.3	8.3	10.9	11.3	12.3	12.8	14.5	1
8	4.6	5.4	7.5	7.7	7.3	9.7	10.6	12.3	13.4	14.1	1
9	3.1	6.4	7.8	6.2	8.6	10	9.6	11.4	13	13.5	1
10	5.8	5.5	6.1	7.1	9.6	9.2	9.6	11.4	13	14.1	1
11	4.1	6	6	8.1	10.5	9.9	10.8	12.4	13.3	15.1	1
12	5.4	6.9	9.2	7.4	9.7	10.2	11.3	11.9	12.3	14.2	1
13	3.8	6.2	7.6	7.2	8.9	10.3	9.6	11.3	12.7	15.2	1
14	5.1	5.7	6.7	8.2	8.8	10.5	10.7	11.8	12.9	14.5	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	6.1	7.2	8.5	8.9	9.6	11.2	12.7	13.3	14.5	15.2	2
2	6.5	7.4	8.3	9.3	10.1	11.1	12.1	13.3	14.4	15.2	2
3	6	6.9	7.9	8.7	9.8	10.8	11.9	13.2	13.8	15	2
4	6.2	7.2	8	8.6	9.6	11.1	11.5	13.1	13.6	14.7	2
5	6.2	7	8.3	9.1	10.2	11.5	12	12.7	14.6	15	2
6	5.9	7.4	8.2	9.2	10.2	10.6	12.1	12.9	13.7	14.9	2
7	5.7	7.1	8.1	9.1	10.3	11.1	11.8	12.7	13.7	15.3	2
8	5.9	7	7.8	8.8	9.5	10.4	11.8	12.7	13.7	14.9	2
9	6.3	7	8.5	8.8	10.2	10.9	11.8	12.8	13.9	15.6	2
10	5.9	6.7	8.5	9.4	10.2	11.4	12.5	13.1	14	15.4	2
11	5.8	7.1	7.6	9.2	10.1	10.9	11.9	12.9	14.4	15	2
12	6	6.8	7.8	9.2	9.7	10.9	11.6	12.6	13.9	14.9	2
13	6.1	7.2	8	9.4	10.2	11	12.2	12.8	13.6	15	2
14	5.7	7	7.8	9	9.5	11	12.2	13.5	13.9	14.6	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2. 두 개의 10차원 정규분포의 데이터

N	속성1	속성2	속성3	속성4	속성5	속성6	속성7	속성8	속성9	속성10	소속군집
1	3	6	9	9	9	11	10	11	14	14	1
2	4	5	7	8	9	11	11	11	13	15	1
3	6	6	7	8	9	9	10	13	13	15	1
4	5	6	8	8	10	9	11	13	14	14	1
5	7	6	7	9	9	11	12	12	14	13	1
6	5	7	8	9	10	9	11	13	14	14	1
7	5	5	7	9	8	11	11	12	13	15	1
8	5	5	8	8	7	10	11	12	13	14	1
9	3	6	8	6	9	10	10	11	13	14	1
10	6	6	6	7	10	9	10	11	13	14	1
11	4	6	6	8	11	10	11	12	13	15	1
12	5	7	9	7	10	10	11	12	12	14	1
13	4	6	8	7	9	10	10	11	13	15	1
14	5	6	7	8	9	11	11	12	13	15	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	6	7	9	9	10	11	13	13	15	15	2
2	7	7	8	9	10	11	12	13	14	15	2
3	6	7	8	9	10	11	12	13	14	15	2
4	6	7	8	9	10	11	12	13	14	15	2
5	6	7	8	9	10	12	12	13	15	15	2
6	6	7	8	9	10	11	12	13	14	15	2
7	6	7	8	9	10	11	12	13	14	15	2
8	6	7	8	9	10	10	12	13	14	15	2
9	6	7	9	9	10	11	12	13	14	16	2
10	6	7	9	9	10	11	13	13	14	15	2
11	6	7	8	9	10	11	12	13	14	15	2
12	6	7	8	9	10	11	12	13	14	15	2
13	6	7	8	9	10	11	12	13	14	15	2
14	6	7	8	9	10	11	12	14	14	15	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3. 정규분포 데이터의 범주형 변환

예를 들어, Table 2는 평균이 5와 6을 따르는 두 개의 10차원 정규분포를 발생시킨 결과이며, 공분산행렬은 차원별로 다양한 값이 나타나도록 하였다. Table 3의 데이터는 발생된 정규분포 데이터에 대하여 정수로 반올림한 후 차원을 데

이터의 속성으로 하는 데이터로 변화 시킨 것을 의미한다. 즉, 1번 속성이 자동차 범주이고, 1:티코, 2:모닝, 3:소나타, 4:아반떼, 6:싼타페, 7:렉스턴, 8:그랜져라고 하면, 소속군집이 1인 군집에서는 1번 속성의 값은 {3, 4, 5, 6}의 범주값으로 구성되어있으며 이는 1번 군집의 자동차에 대한 범주들은{소나타, 아반떼, 싼타페, 렉스턴}을 포함하는 레코드임을 의미하며, 소속군집이 2인 군집에서는 {6, 7}로 구성되어있고, 이는 {싼타페, 렉스턴}이 포함된 레코드라는 것을 의미한다. 실험에서는 제안 Chain k-modes 알고리즘의 성능을 확인하기 위하여 다양한 개수(K=2, 3, 4)의 군집으로 실험하였으며, 특히, K=2일 경우 군집의 혼잡한 정도를 다양하게 조절하면서 알고리즘을 적용하였다.

4.1.1 군집의 개수(K=2)에 따른 실험 Type I

본 실험에서는 평균이 5와 6인 두 개의 정규난수를 발생을 시켜 분석하였다. 군집에 대하여 각각 750개의 난수 데이터를 발생시켰으며 총 1500개의 데이터를 분석하였다.

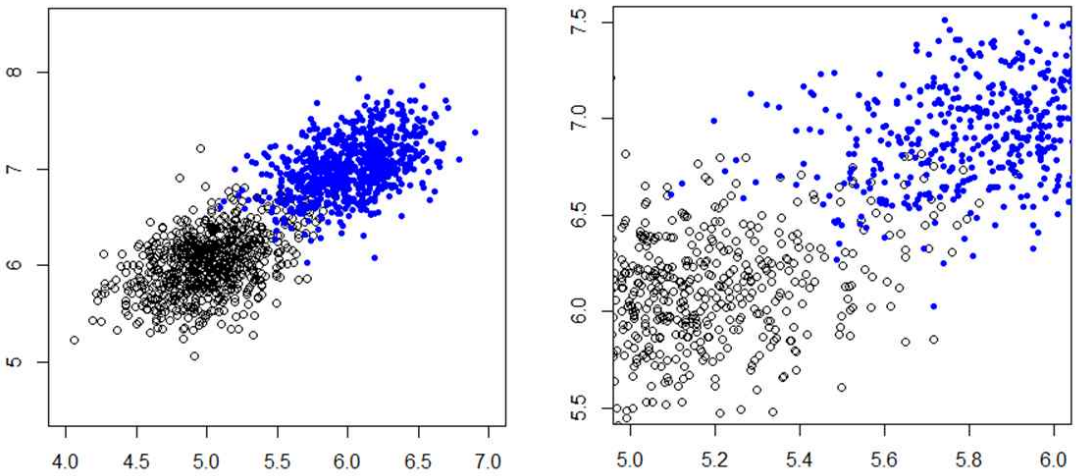
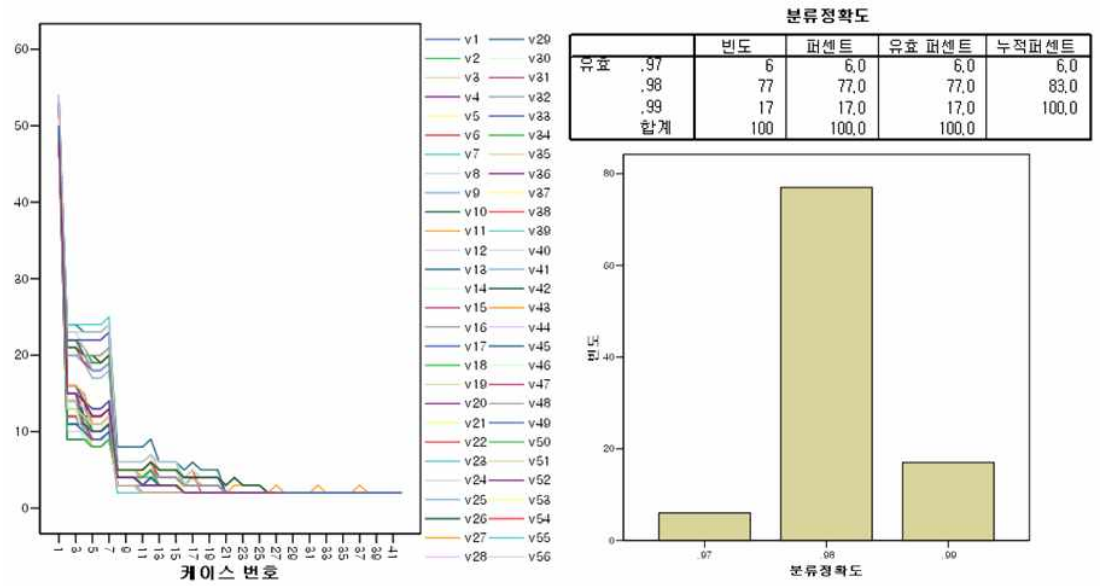


Figure 29. 비교적 덜 혼잡한 두 개의 10차원 정규분포

Figure 29는 발생된 난수의 plot과 두 군집의 중복 지점을 보여주고 있다. 두 개의 군집은 비교적 잘 구분되어지고 있다. 제안 Chain k-modes 알고리즘을 수행하기 위하여 알고리즘에는 이 데이터가 2개의 군집으로 구성되어있음을 제공하지 않고 분석하였다. 분석결과 1500개의 데이터에 대하여 군집은 항상 2개의 군집으로 형성되었다. Figure 30는 알고리즘을 100회 반복하는 동안 데이터가 최적의 군집의 개수로 수렴하는 과정을 보여주고 있다. 먼저, 1500개의 데이터는 첫 알고리즘의 수행으로 인해 약 50여개의 mode로 구분되고 알고리즘의 반복을 통해 2개의 군집으로 수렴되었음을 알 수 있다. 또한, 분류의 정확도는 100번 중 17번이 99%이상의 분류 정확도를 보이고 있으며, 대부분 98%이상의 정확도는 77%로 나타나고 있다. 또한, 100회 반복수행에서 대부분이 97%이상의 높은 분류 정확도를 보이고 있다. 제안하는 Chain k-modes 알고리즘의 기본적인 성능은 만족스런 결과라고 할 수 있다. 하지만 발생된 데이터가 중복이 많지 않으므로 자연스러운 결과라고 할 수도 있다.



분류정확도

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	.97	6	6.0	6.0	6.0
	.98	77	77.0	77.0	83.0
	.99	17	17.0	17.0	100.0
	합계	100	100.0	100.0	

Figure 30. 실험(4.1.1)에 대한 Chain k-modes 알고리즘 분석결과

4.1.2 군집의 개수(K=2)에 따른 실험 Type II

본 실험에서는 평균이 5와 6인 두 개의 정규난수를 발생시켜 분석하였으며, 군집의 확실한 구분를 위하여 공분산 행렬을 적절히 조절하여 구분이 뚜렷한 1500개의 데이터를 발생시켰다.

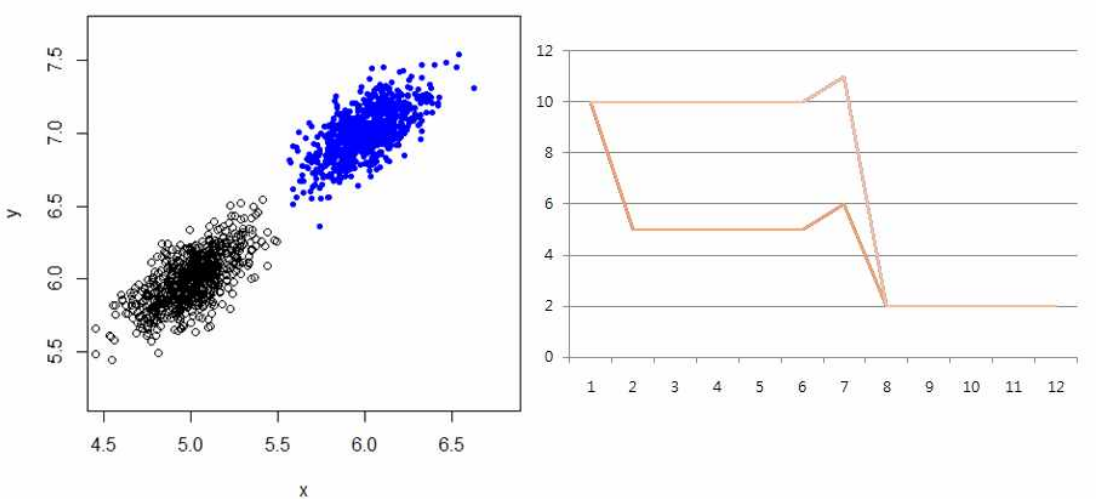


Figure 31. 구분이 분명한 군집에 대한 Chain k-mode 알고리즘 분석결과

Figure 31은 구분이 분명한 데이터에 대한 알고리즘의 분석결과로서, 100회 반복수행에서 100%의 분류 정확도를 보여주고 있으며, 2개의 군집으로 잘 구분되었다.

4.1.3 군집의 개수(K=2)에 따른 실험 Type III

본 실험에서는 평균이 5와 6인 두 개의 정규난수를 발생을 시켜 실험하였다. 군집에 대하여 각각 750개의 난수 데이터를 발생시켰으며 총 1500개의 데이터를 분석하였다. 또한 군집의 구분을 모호하게 하기 위하여 공분산행렬을 적절히 조절하였으며, Figure 34에서와 같이 중복되는 지점에서는 데이터들이 혼잡하게 섞여 있음을 알 수 있다.

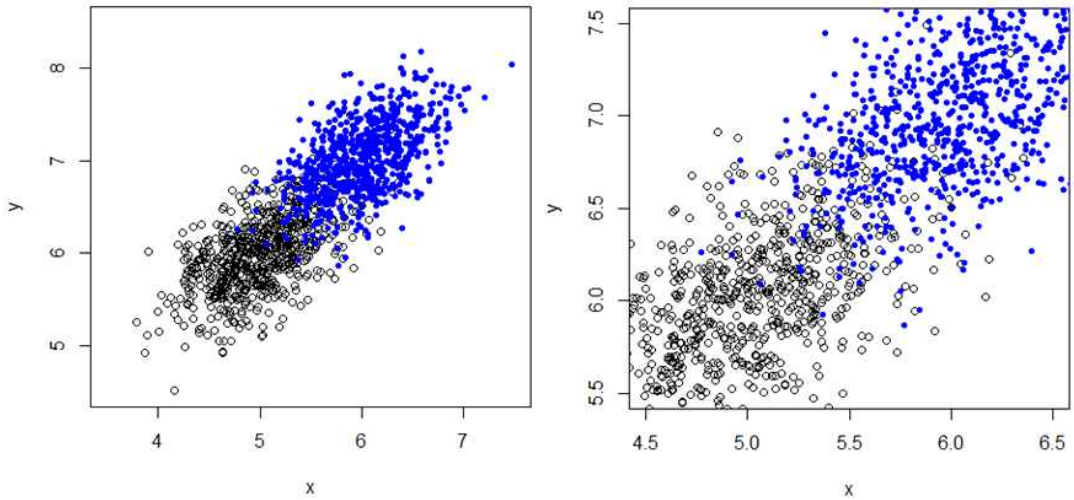


Figure 32. 혼잡한 두 개의 10차원 정규분포 군집

분류정확도

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	.50	4	4.0	4.0	4.0
	.51	3	3.0	3.0	7.0
	.74	1	1.0	1.0	8.0
	.76	1	1.0	1.0	9.0
	.80	2	2.0	2.0	11.0
	.81	1	1.0	1.0	12.0
	.83	2	2.0	2.0	14.0
	.86	2	2.0	2.0	16.0
	.87	1	1.0	1.0	17.0
	.89	6	6.0	6.0	23.0
	.92	1	1.0	1.0	24.0
	.93	2	2.0	2.0	26.0
	.94	3	3.0	3.0	29.0
	.95	62	62.0	62.0	91.0
	.96	9	9.0	9.0	100.0
	합계	100	100.0	100.0	

군집의개수

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	2	82	82.0	82.0	82.0
	3	17	17.0	17.0	99.0
	4	1	1.0	1.0	100.0
	합계	100	100.0	100.0	

상관계수

		분류정확도	군집의개수
분류정확도	Pearson 상관계수	1	.189
	유의확률(양쪽)		.059
	N	100	100
군집의개수	Pearson 상관계수	.189	1
	유의확률(양쪽)	.059	
	N	100	100

Table 4. 실험(4.1.3)에 대한 Chain k-modes 알고리즘의 분류 정확도 및 개수

Figure 32는 발생된 난수의 plot과 두 군집의 중복 지점을 보여주고 있다. 두 개 군집의 중복지점은 구분하기 모호하게 섞여있다. Table 4는 Chain

k-modes 알고리즘의 분석결과이다. 군집의 개수는 군집의 정보를 미리 제공하지 않았으나 100번 중 82번에서 2개의 군집으로 분류하였으며, 17번은 3개의 군집으로 분류하였으며, 4개의 군집으로 분류한 경우도 1회 나타났다. 분류의 정확도는 100번 중 71번이 95%이상의 분류 정확도를 보이고 있으며, 다소 낮은 분류 정확도가 나타났다. 이는 제안 하는 Chain k-modes 알고리즘의 강력한 장점이라고 할 수 있다. 결과에서 17회의 군집의 개수가 3개로 분류된 경우, 이는 당연히 분류정확도는 낮다고 판단된다. 하지만 이는 데이터를 2개의 군집으로 제한하였을 때 분류의 정확도를 의미한다. 하지만, 제안 하는 Chain k-modes 알고리즘에서는 두 개의 성질이 다른 순수한 두개의 군집으로 분류하고, 이에 대하여 두 개의 군집의 성질이 섞여있는 새로운 군집을 구분할 수 있다는 것이다. 즉, 미지의 데이터에 대한 분석에서 사전 정보가 없어도 Chain k-modes 알고리즘은 순수한 군집과 혼합된 군집을 새로운 군집으로 구분할 수 있다.

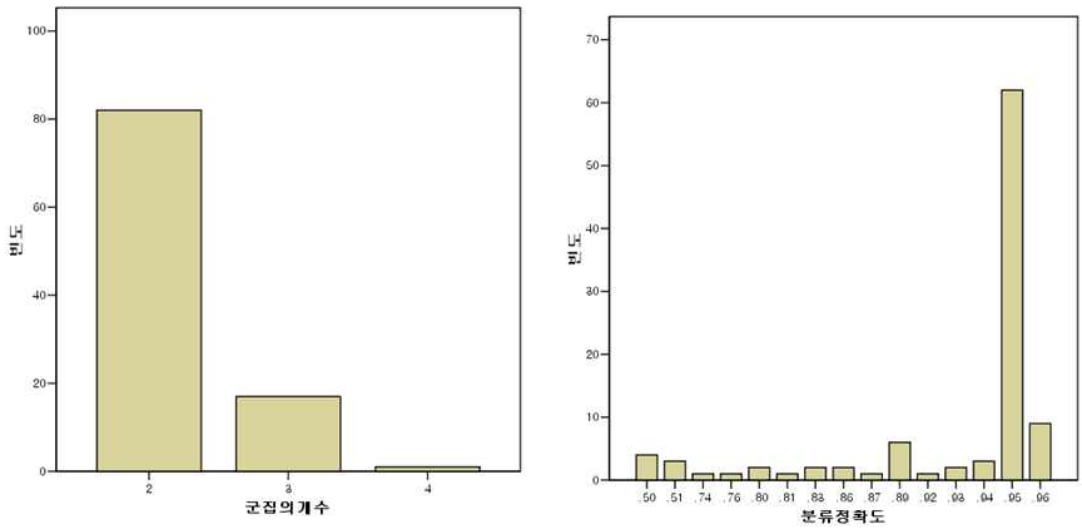


Figure 33. 실험(4.1.3)에 대한 Chain k-modes 알고리즘의 군집의 개수 및 분류정확도 그래프

4.1.4 군집의 개수(K=2)에 따른 실험 Type IV

본 실험에서는 평균이 5와 6인 두 개의 정규난수를 발생을 시켜 분석하였다. 군집에 대하여 각각 750개의 난수 데이터를 발생시켰으며 총 1500개의 데이터를 분석하였다. 또한 군집의 구분을 모호하게 하기 위하여 공분산행렬을 적절히 조절하였으며, Figure 34에서와 같이 데이터들이 중복되는 지점에서는 매우 혼잡하게 섞여 있음을 알 수 있다.

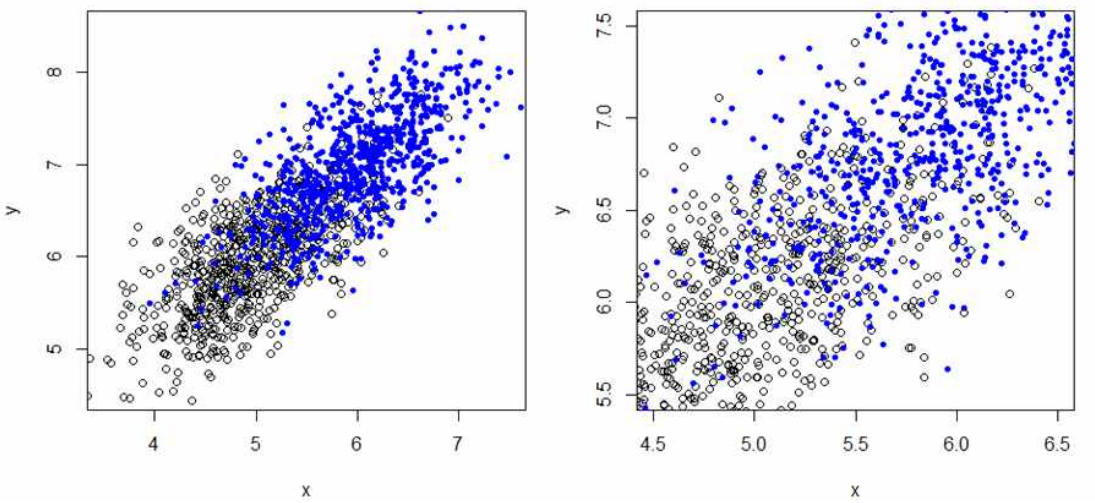


Figure 34. 매우 혼잡한 두 개의 10차원 정규분포

Figure 34는 발생된 난수의 plot과 두 군집의 중복 지점이 매우 혼잡하게 뒤섞여 있음을 보여주고 있다. 분석결과 1500개의 데이터에 대하여 군집은 2개, 3개, 4개 및 5개의 군집으로 구분되는 경우가 나타났으며, 이는 매우 의미 있는 결과이다. 분류의 정확도는 Chain k-modes 알고리즘이 3개 이상 분류되었을 때 의미가 없어지며, 군집의 개수가 매우 중요한 관심이 된다. 100회 반복결과 군집의 개수는 100번 중 68번이 2개로 구분되었으며, 29번 정도는 3개의 군집으로 구분되었다. 이는 Chain k-modes 알고리즘의 군집분석 과정에서 두 군집의 속성이 섞여있는 데이터를 새로운 군집으로 판단한다는 것을 의미하므로, 군집이 섞여있는 데이터일수록 Chain k-modes 알고리즘은 더욱 우수한 성능을 보여주고 있다.

분류정확도

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	.51	6	6.0	6.0	6.0
	.52	24	24.0	24.0	30.0
	.70	1	1.0	1.0	31.0
	.76	7	7.0	7.0	38.0
	.77	1	1.0	1.0	39.0
	.78	3	3.0	3.0	42.0
	.79	16	16.0	16.0	58.0
	.80	5	5.0	5.0	63.0
	.81	31	31.0	31.0	94.0
	.82	3	3.0	3.0	97.0
	.84	3	3.0	3.0	100.0
	합계	100	100.0	100.0	

군집의개수

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	2	68	68.0	68.0	68.0
	3	29	29.0	29.0	97.0
	4	2	2.0	2.0	99.0
	5	1	1.0	1.0	100.0
	합계	100	100.0	100.0	

상관계수

		분류정확도	군집의개수
분류정확도	Pearson 상관계수	1	.420**
	유의확률 (양쪽)		.000
	N	100	100
군집의개수	Pearson 상관계수	.420**	1
	유의확률 (양쪽)	.000	
	N	100	100

**. 상관계수는 0.01 수준(양쪽)에서 유의합니다.

Table 5. 실험(4.1.4)에 대한 Chain k-modes 알고리즘의 분류정확도 및 군집의 개수

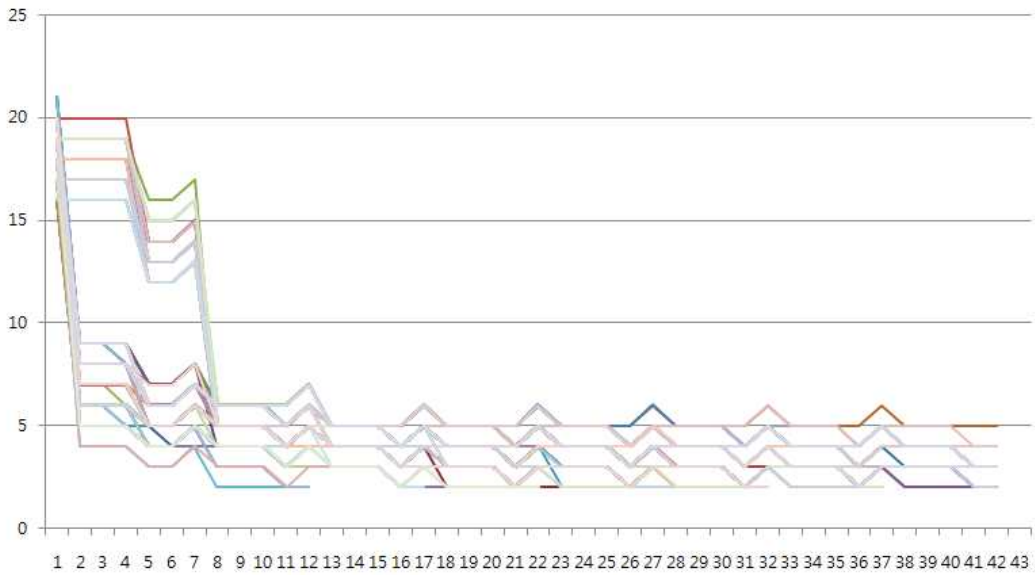


Figure 35. 실험(4.1.4)의 Chain k-modes 알고리즘의 반복회수

Figure 35에서는 Chain k-modes 알고리즘의 반복 회수를 보여주고 있다. 반복회수가 많아질수록 데이터들은 2개, 3개, 4개 및 5개의 군집으로 수렴하고 있음을 알 수 있다.

4.1.5 군집의 개수(K=3)에 따른 실험

본 실험에서는 평균이 5, 6 및 7인 세 개의 정규난수를 발생시킨 결과이다. 군집에 대하여 각각 300개의 난수 데이터를 발생시켰으며 총 900개의 데이터를 분석하였다. 또한 군집의 구분을 모호하게 하기 위하여 공분산행렬을 적절히 조절하였으며, Figure 36에서와 같이 중복되는 지점에서는 비교적 구분하기 양호하도록 설정하였다. 하지만 군집 구분(k=3)에 대한 정보를 제공하지 않을 경우 데이터의 경계는 매우 모호하게 된다.

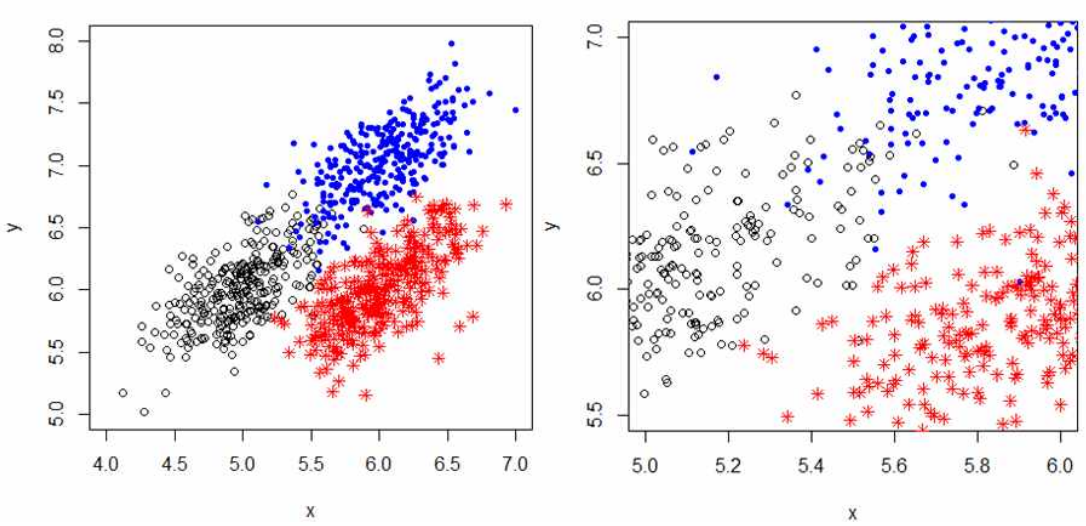


Figure 36. 비교적 덜 혼잡한 3개의 10차원 정규분포

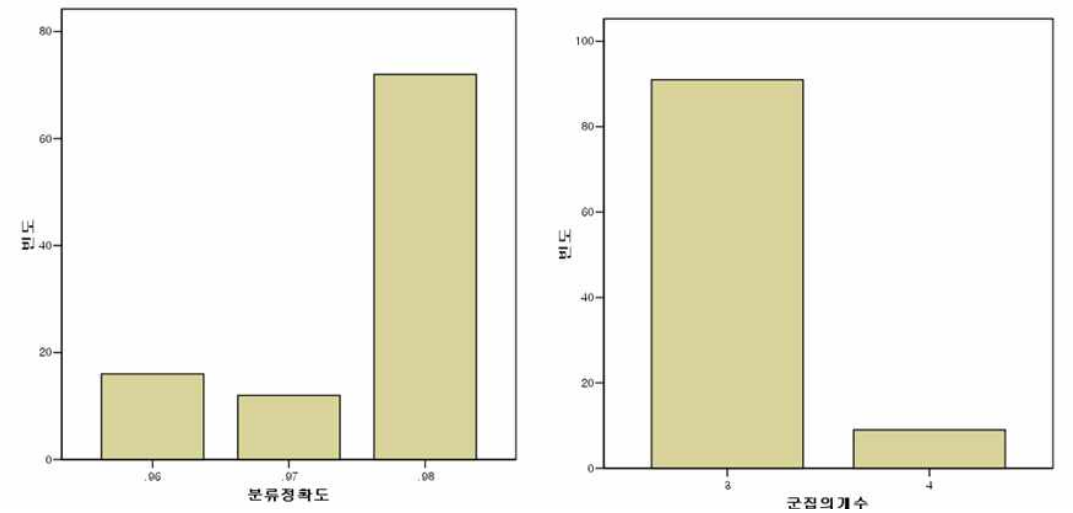


Figure 37. 실험(4.1.5)에 대한 Chain k-modes 알고리즘의 분석결과 분류 정확도와 군집의 개수 그래프

분류정확도

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	.96	16	16.0	16.0	16.0
	.97	12	12.0	12.0	28.0
	.98	72	72.0	72.0	100.0
	합계	100	100.0	100.0	

군집의개수

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	3	91	91.0	91.0	91.0
	4	9	9.0	9.0	100.0
	합계	100	100.0	100.0	

상관계수

		분류정확도	군집의개수
분류정확도	Pearson 상관계수	1	-.095
	유의확률(양쪽)		.349
	N	100	100
군집의개수	Pearson 상관계수	-.095	1
	유의확률(양쪽)	.349	
	N	100	100

Table 6. 실험(4.1.5)에 대한 Chain k-modes 알고리즘 분석결과의 분류 정확도 및 군집의 개수 결과

Table 6은 Chain k-modes 알고리즘 분석결과에 대한 분류 정확도 및 군집의 개수 결과이다. 분류정확도는 대부분 96%이상으로 나타나고 있으며, 3개의 군집으로 분류한 것은 100번 중 91회 나타나고 있으며, 4개의 군집으로 분류된 것은 100번 중 9번 나타나고 있다. 4개의 군집으로 분류한 경우, 3개의 군집은 원래의 군집의 정보를 가지고 있는 군집과 하나의 혼합된 속성으로 이루어진 군집으로 분류되고 있다. 분류정확도와 군집의 개수에 대한 상관관계에서는 일반적으로 상관성이 나타나고 있지는 않고 있다.

Figure 38은 Chain k-modes 알고리즘을 100회 반복수행한 결과이다. 1500개의 데이터에 대하여 초기 100개 이상의 초기 mode들이 생성이 되고 알고리즘의 반복을 통해 최적의 군집으로 병합 및 갱신되는 과정을 보여주고 있다.

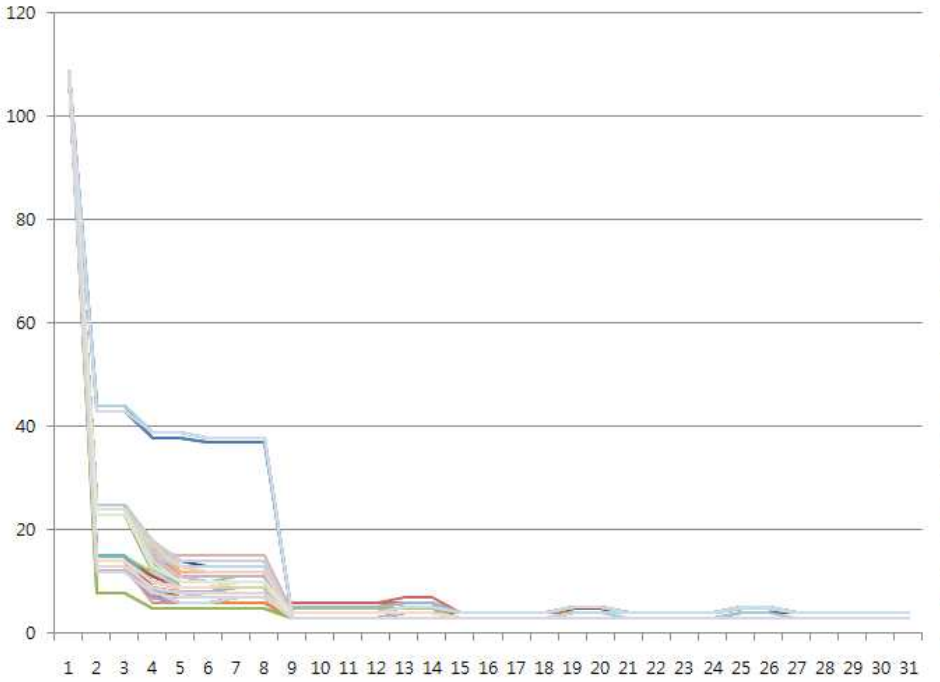


Figure 38. 실험(4.1.5)에 대한 Chain k-modes 알고리즘의 100회 반복 수행 결과 그래프

4.1.6 군집의 개수(K=4)에 따른 실험 Type I

본 실험에서는 평균이 5,5,6 및 7인 네 개의 정규난수를 발생시킨 결과 그래프이다. 평균이 5로 동일한 난수이지만, 이는 1차원에서만 동일하고 2차원이상의 고 차원에서는 다른 난수가 발생하도록 구성하여, 중복된 조건의 난수가 발생하지 않도록 하였다. 군집에 대하여 각각 100개의 난수 데이터를 발생시켰으며 총 400개의 데이터를 분석하였다. 또한 군집의 구분을 모호하게 하기 위하여 공분산행렬을 적절히 조절하였으며, Figure 39에서와 같이 중복되는 지점에서는 비교적 구분하기 양호하도록 설정하였다. 하지만 Figure 42의 군집 구분(k=4)에 대한 정보를 제공하지 않을 경우 데이터의 경계는 매우 모호하다.

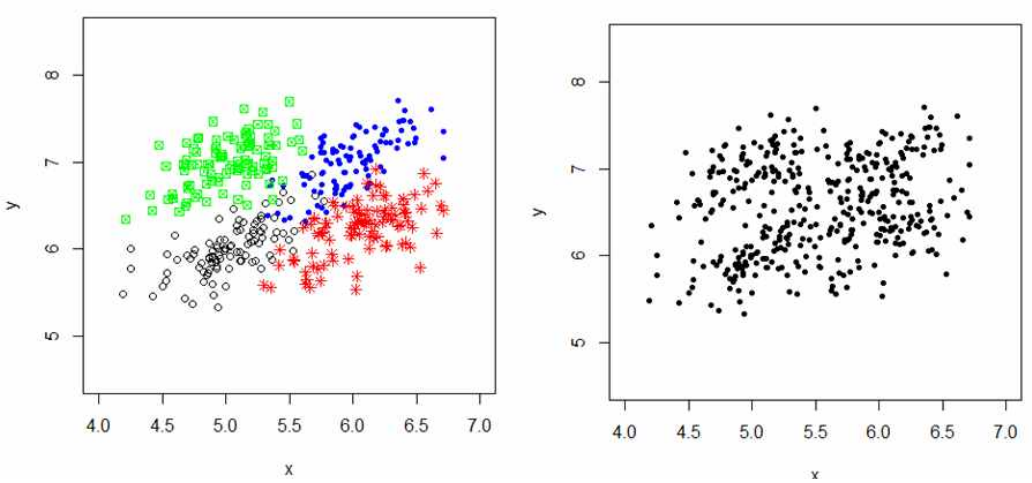


Figure 39. 비교적 덜 혼잡한 4개의 10차원 난수

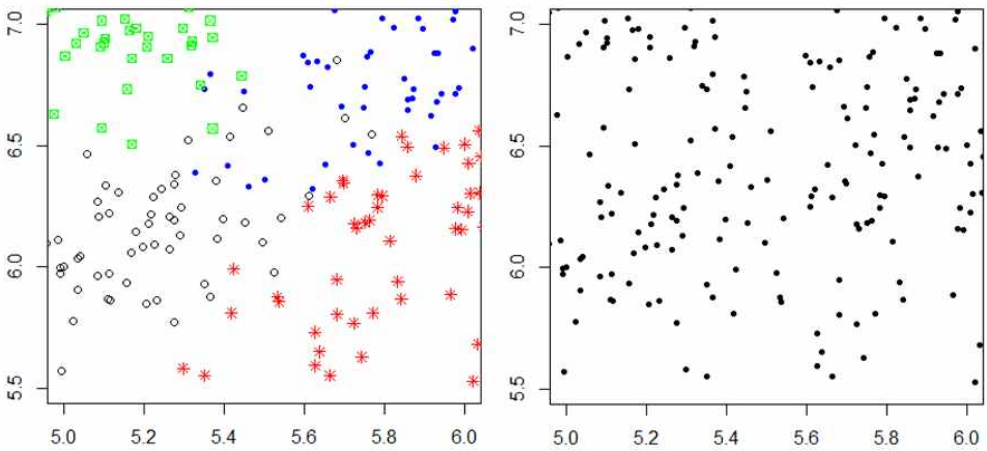


Figure 40. 4개의 10차원 난수의 정보 유무에 따른 비교

군집의개수

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	4	98	98,0	98,0	98,0
	5	2	2,0	2,0	100,0
	합계	100	100,0	100,0	

Table 7. 실험(4.1.6)에 대한 Chain k-modes 알고리즘의 100회 반복 수행 결과

Table 7은 Chain k-modes 알고리즘의 100회 반복 수행 결과이다. 대부분의 경우 4개의 군집으로 분석되고 있다.

4.1.7 군집의 개수(K=4)에 따른 실험 Type II

본 실험에서는 평균이 5, 5, 6 및 7인 정규난수를 발생시킨 결과 그래프이다. 평균이 5로 동일한 난수이지만, 이는 1차원에서만 동일하고 2차원 이상의 고 차원에서는 다른 난수가 발생하도록 구성하여, 중복된 조건의 난수가 발생하지 않도록 하였다. 또한, 공분산 행렬을 적절히 조절하여 군집이 경계가 모호하도록 설정하였으며, Chain k-modes 알고리즘 분석 시 군집의 정보를 제공하지 않고 분석하였다. 각 군집에 대하여 100개의 난수 데이터를 발생시켰으며 총 400개의 데이터를 분석하였다. 또한, 군집의 구분을 모호하게 하기 위하여 공분산행렬을 적절히 조절하였으며, Figure 41에서와 같이 중복되는 지점에서 데이터가 혼잡하도록 설정하였다. Figure 42는 Figure 41에서의 혼잡한 곳을 확대한 그림으로서, 군집에 대한 정보를 제공하지 않을 경우 데이터의 경계는 매우 모호하다는 것을 알 수 있다.

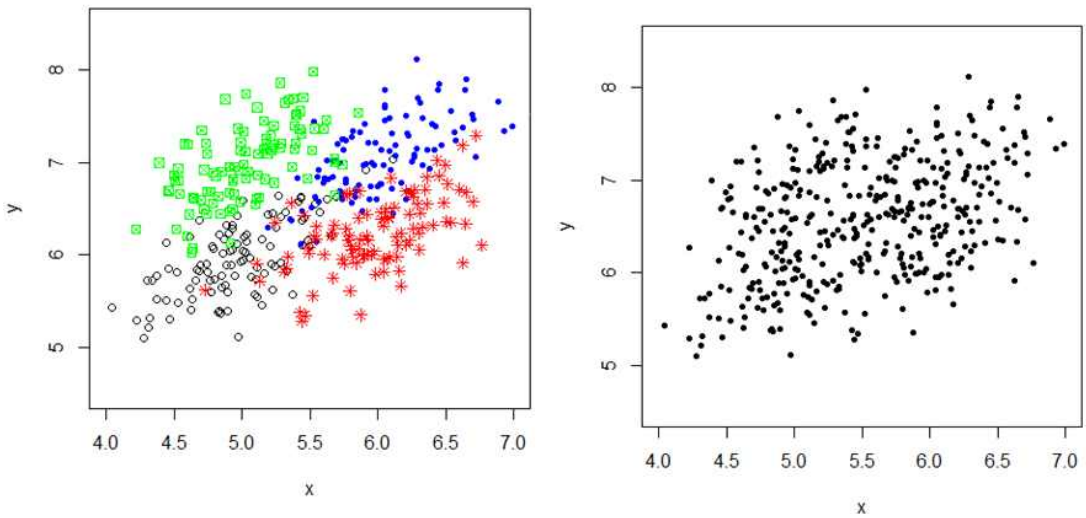


Figure 41. 혼잡한 4개의 10차원 난수

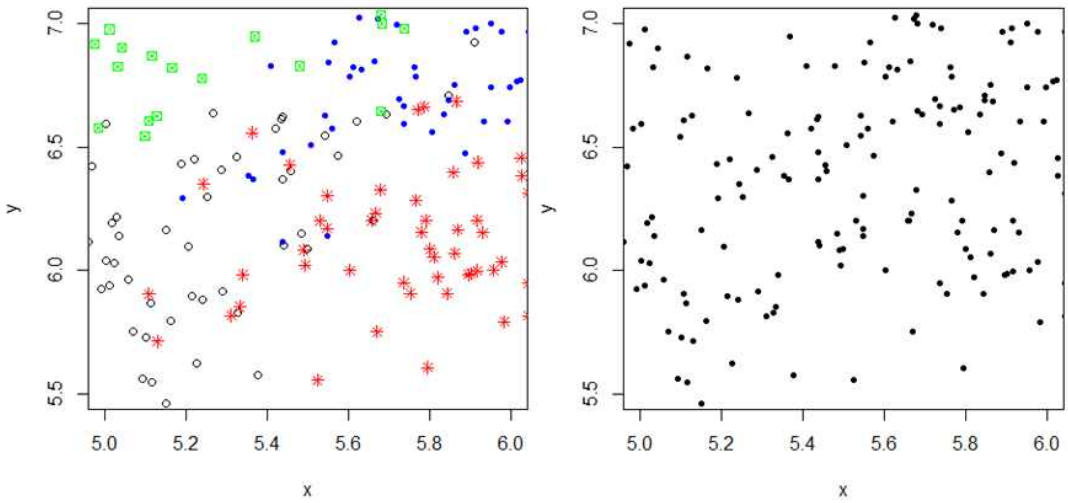


Figure 42. 혼잡한 4개의 10차원 난수의 정보 유무에 따른 비교

분류정확도

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	.92	2	1.7	1.7	1.7
	.93	5	4.3	4.3	6.1
	.94	28	24.3	24.3	30.4
	.95	32	27.8	27.8	58.3
	.96	29	25.2	25.2	83.5
	.97	4	3.5	3.5	87.0
	.98	15	13.0	13.0	100.0
	합계	115	100.0	100.0	

군집의개수

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	4	65	56.5	56.5	56.5
	5	35	30.4	30.4	87.0
	6	13	11.3	11.3	98.3
	7	2	1.7	1.7	100.0
	합계	115	100.0	100.0	

상관계수

		분류정확도	군집의개수
분류정확도	Pearson 상관계수	1	-.323**
	유의확률 (양쪽)		.000
	N	115	115
군집의개수	Pearson 상관계수	-.323**	1
	유의확률 (양쪽)	.000	
	N	115	115

**. 상관계수는 0.01 수준(양쪽)에서 유의합니다.

Table 8. 실험(4.1.7)에 대한 Chain k-modes 알고리즘의 100회 반복 수행 결과

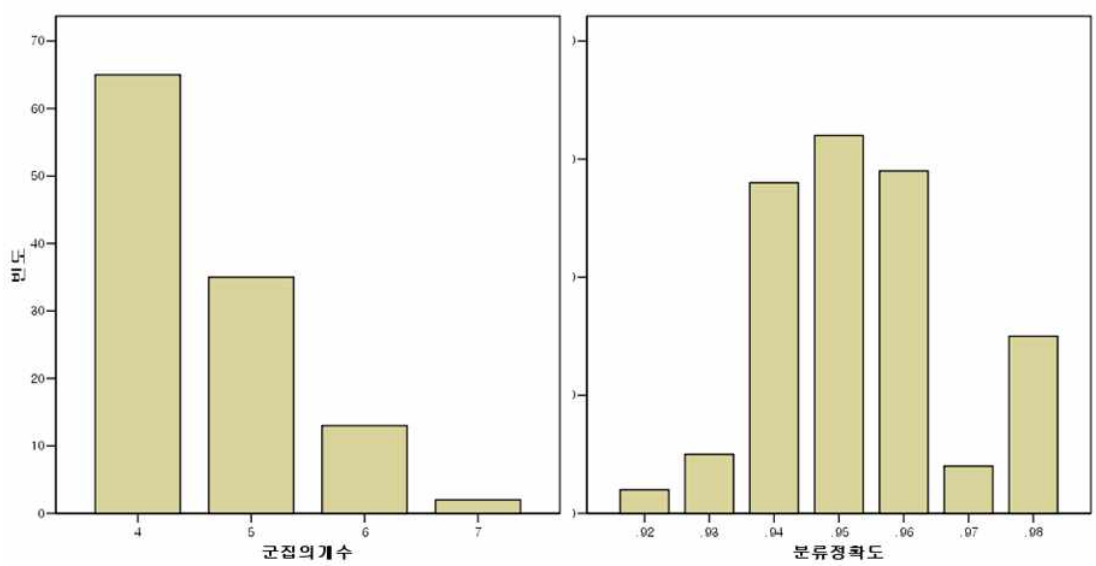


Figure 43. 실험(4.1.7)에 대한 Chain k-modes 알고리즘의 결과 그래프

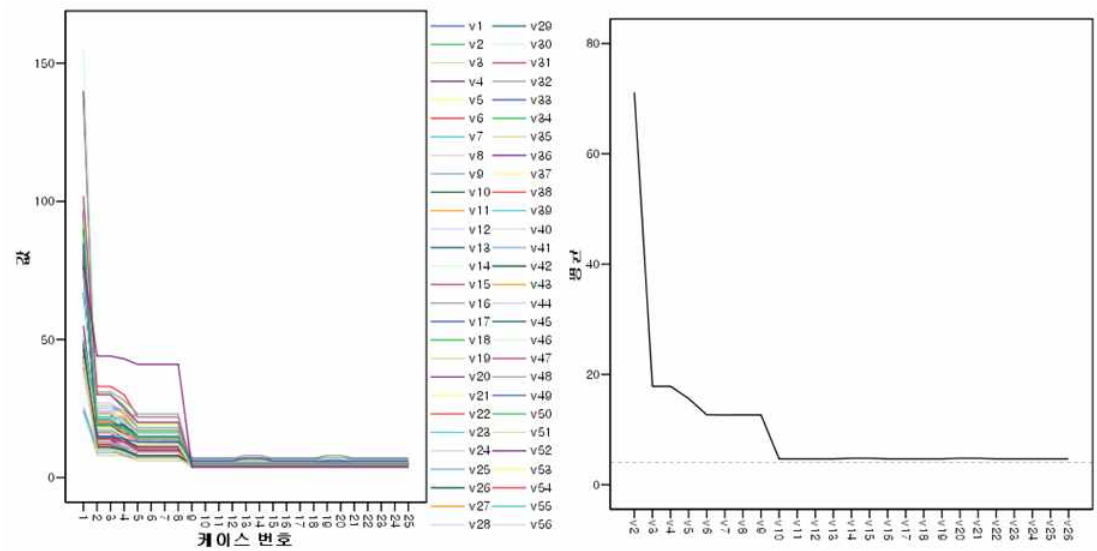


Figure 44. 실험(4.1.7)에 대한 Chain k-modes 알고리즘의 100회 반복 수행 결과와 평균그래프

경계가 모호한 4개의 군집에 대한 분석에서도 Chain k-modes 알고리즘은 우수한 성능을 보여주고 있다. 4개의 군집으로 하였으며, 5개 이상의 군집으로 분류하는 경우는 데이터의 속성이 혼합된 데이터들을 새로운 군집으로 구분하고 있다. Figure 44의 평균그래프에서는 Chain k-modes 알고리즘의 일정한 군집의 수로 수렴하고 있음을 잘 보여주고 있다.

4.2 Mushroom dataset

Mushroom data set은 UCI Machine Learning Repository에서 제공하는 범주형데이터로서, 22개의 과 8124개의 레코드로 구성되어있다.

속성	속성값	수치형표현
1 cap-shape	bell=b,conical=c,convex=x,flat=f, knobbed=k, sunken=s	1~6
2 cap-surface	fibrous=f,grooves=g,scaly=y,smooth=s	1~4
3 cap-color	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y	1~10
4 bruises	bruises=t,no=f	1~2
5 odor	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s	1~9
6 gill-attachment	attached=a,descending=d,free=f,notched=n	1~4
7 gill-spacing	close=c,crowded=w,distant=d	1~3
8 gill-size	broad=b,narrow=n	1~2
9 gill-color	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y	1~12
10 stalk-shape	enlarging=e,tapering=t	1~2
11 stalk-root	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?	1~7
12 stalk-surface-above-ring	fibrous=f,scaly=y,silky=k,smooth=s	1~4
13 stalk-surface-below-ring	fibrous=f,scaly=y,silky=k,smooth=s	1~4
14 stalk-color-above-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y	1~9
15 stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y	1~9
16 veil-type	partial=p,universal=u	1~2
17 veil-color	brown=n,orange=o,white=w,yellow=y	1~4
18 ring-number	none=n,one=o,two=t	1~3
19 ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z	1~8
20 spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y	1~9
21 population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y	1~6
22 habitat	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d	17
23 Class	edible , poisonous	1,2

Table 9. Mushroom Dataset의 속성 성분

Mushroom data set은 버섯의 상처, 외관 및 색상 등에 대한 정보를 제공하고 있으며 모든 속성은 범주형 속성이다. 최종적인 군집의 개수는 독이 없는 것과 독이 있는 것으로 구분하고 있다.

・샘플링

Mushroom data set은 총 8124개의 레코드로 구성되어있으며, 매 실험에서 8124개의 레코드 중 임의로 100개의 레코드를 선정하여 실험하였으며, 초기치 mode의 값은 100개의 레코드 중 U(0,1)에서 난수를 선정하였다.

・한계기준

한계기준 θ는 초기 mode를 기반으로 새로운 mode를 생성할 것인지, 기존 mode의 소속으로 병합될 것인지를 결정하는 한계기준으로서 초기 mode의 개수를 결정하는데 중요한 역할을 한다. 즉, 비교하는 레코드와 mode간의 거리 값 δ)이 θ보다 크면 비교하는 레코드는 기존 mode들에 흡수될 확률이 증가한다. 본 논문에서는 그 한계기준을 임의로 선정된 100개의 레코드 중에서 2개를 임의로 선정하여 유사도를 계산하고 이를 30회 반복하여 그 유사도 값들을 기반으로 θ를 구하였다.

반복횟수	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	18	11	9	11	18	12	10	11	8	13	12	9	12	9	11
반복횟수	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ_i	8	15	12	10	18	10	18	13	7	11	10	14	8	18	11

Table 10. 한계기준 의 계산을 위한 30회 랜덤추출 유사도 계산결과

Table 10에서는 샘플 100개중에서 30쌍의 레코드를 추출하여 그 유사도를 계산한 결과이다. 본 논문에서는 유사도의 10%의 상한을 기준으로 θ의 값을 정하고자 하고 있으며, 만약 데이터가 정규성을 따르게 되면 θ는 에서의 유사도 값을 취할 수 있다. 하지만 일반적으로 데이터가 정규성을 따르는지 확인되어야

할 것이다. 하지만 Figure 45과 Table 11에서는 샘플 유사도의 정규성은 만족하고 있지 않다. 즉, 일반적으로 샘플데이터간의 유사도 값들은 정규성을 보장할 수는 없다. 그래서 본 논문에서는 의 값을 유사도 값의 θ분위수로 하였다. 분위수는 표본에서 관측된 유사도 값의 θ% 위치에 있는 유사도 값으로서, 본 실험에서는 $\theta = 18$로 설정되었다.

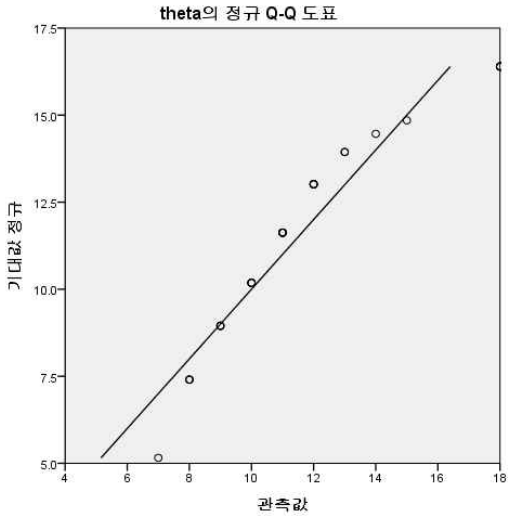


Figure 45. 샘플 유사도 값들에 대한 Q-Q Plot

Kolmogorov-Smirnov			Shapiro-Wilk		
통계량	자유도	유의확률	통계량	자유도	유의확률
.188	30	.008	.885	30	.004

Table 11. 샘플 유사도 값들의 정규성 검정

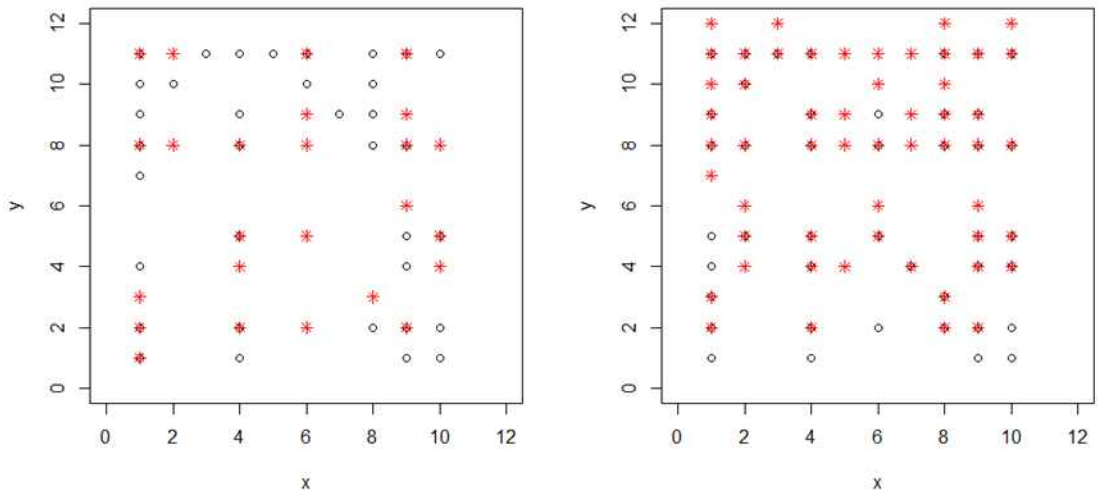


Figure 46. Mushroom data set의 두 군집(*:독있음, ◦:식용)의 비교 plot

Figure 46은 Mushroom data set의 두 군집(*:독있음, ◦:식용)의 비교 plot을 보여주고 있다. Figure 48(左)과 Figure 48(友)은 Mushroom data set의 3번 속성과 5번 속성의 값들을 비교한 것이다. Figure 48(左)에서는 두 군집의 차이가 비교적 뚜렷하게 나타나고 있다. 하지만 Figure 48(友)에서는 두 군집에서 중복되는 곳이 많이 나타나고 있음을 알 수 있다. 결국, 3번 속성은 두 집단을 구분하는 요인으로 중요하다고 할 수 있다.

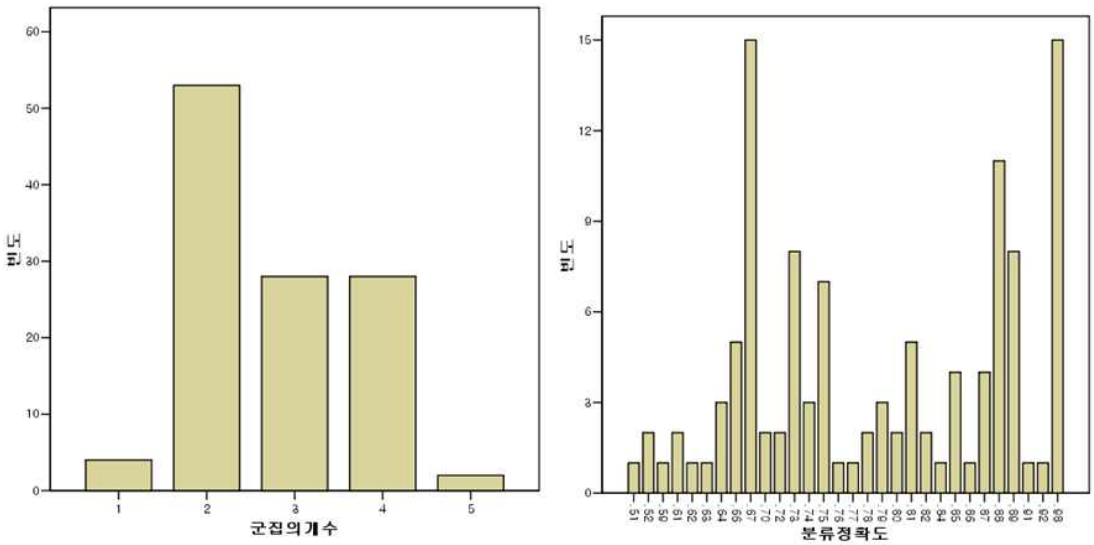


Figure 47. Mushroom data set에 대한 Chain k-modes 알고리즘 분석 결과 그래프

Figure 47은 Mushroom data set에 대한 Chain k-modes 알고리즘의 분석 결과 그래프이다. 100회 반복 실험 중 50회 이상에서 2개의 군집으로 군집되고 있으며, 3개 혹은 4개의 군집으로 나타나는 경우도 나타나고 있다. 군집의 분류 정확도는 2개의 군집일 경우 85%이상으로 잘 나타나고 있지만, 3개 혹은 4개의 군집으로 분류되고 있을 때는 분류 정확도가 낮아지고 있다. 분류의 정확도는 데이터에 대한 정보(군집의 개수)가 없을 경우 분류의 정확도보다는 분류된 군집의 특성을 분석하는 것이 더욱 중요하다.

분류 정확도	10회		
	$=2$		$k_0=1$
	기존 알고리즘	개선 k-mode 알고리즘	제안
100%			
97%			
95%	1	3	2
90%	2	1	2
85%		3	1
80%		2	2
75%	1		1
70%	1	1	1
65%	1		1
60%	2		
50%	2		
40%			
평균	0.70	0.86	0.83
표준편차	0.16	0.08	0.1

Table 12. 알고리즘의 반복 최종수행 결과

Table 12에서는 알고리즘의 반복 수행 결과를 보여주고 있다. 기존의 방법으로는 평균적으로 약 70%정도의 정확도를 보여주고 있으며, 개선된 방법은 86% 정도의 정확도를 보여주고 있다. 제안 알고리즘에서는 정확도가 83%정도로 나타나고 있으며, 이는 개선된 방법보다는 정확도면에서는 다소 낮게 나타나고 있다.

반면, 기존 알고리즘과 개선 k-mode 알고리즘은 초기치를 미리 주어져야 분석이 가능하지만, 제안 알고리즘에서는 초기치를 제공하지 않아도 높은 분류 정확도를 보이고 있다.

4.3 Soybean(Small data set)

Soybean (Small) data set은 UCI Machine Learning Repository에서 제공하는 범주형데이터로서, 35개의 속성과 47개의 레코드로 구성되어있다.

	속성	속성값	수치형표현
1	date	April, may, June, July, August, september, october, ?.	0,1,2,3,4,5,6,7
2	plant-stand	normal, lt-normal, ?.	0,1,2
3	precip	lt-norm, norm, gt-norm, ?.	0,1,2
4	temp	lt-norm, norm, gt-norm, ?.	0,1,2
5	hail	yes, no, ?.	0,1,2
6	crop-hist	diff-lst-year, same-lst-yr, same-lst-two-yrs,same-lst-sev-yrs, ?.	0,1,2,3,4
7	area-damaged	scattered, low-areas, upper-areas, whole-field, ?.	0,1,2,3,4
8	sevrity	minor, pot-severe, severe, ?.	0,1,2,3
9	seed	none, fungicide, other, ?.	0,1,2,3
10	germination	90-100%, 80-89%, lt-80%, ?.	0,1,2,3
11	plant-growth	norm, abnorm, ?.	0,1,2
12	leaves	norm, abnorm.	0,1
13	leafspots-halo	absent, yellow-halos, no-yellow-halos, ?.	0,1,2
14	leafspots-marg	w-s-marg, no-w-s-marg, dna, ?.	0,1,2,3
15	leafspot-size	lt-1/8, gt-1/8, dna, ?.	0,1,2,3
16	leaf-shread	absent, present, ?.	0,1,2
17	leaf-malf	absent, present, ?.	0,1,2
18	leaf-mild	absent, upper-surf, lower-surf, ?.	0,1,2,3
19	stem	norm, abnorm, ?.	0,1,2
20	lodging	yes, no, ?.	0,1,2
21	stem-cankers	absent, below-soil, above-soil, above-sec-nde, ?.	0,1,2,3,4
22	canker-lesion	dna, brown, dk-brown-blk, tan, ?.	0,1,2,3,4
23	fruiting-bodies	absent, present, ?.	0,1,2
24	external decay	absent, firm-and-dry, watery, ?.	0,1,2,3
25	mycelium	absent, present, ?.	0,1,2
26	int-discolor	none, brown, black, ?.	0,1,2,3
27	sclerotia	absent, present, ?.	0,1,2
28	fruit-pds	norm, diseased, few-present, dna,?.	0,1,2,3
29	fruit spots	absent, colored, brown-w/blk-specks, distort, dna, ?	0,1,2,3,4,5
30	seed	norm, abnorm, ?.	0,1,2
31	mold-groth	absent, present, ?.	0,1,2
32	seed-discolor	absent, present, ?.	0,1,2
33	seed-size	norm, lt-norm, ?.	0,1,2
34	shriveling	absent, present, ?.	0,1,2
35	roots	norm, rotted, galls-cysts, ?.	0,1,2,3

Table 13. Soybean data set의 속성 성분

Table 13는 Soybean data의 속성들을 보여주고 있다. 총 35개의 속성에 대하여 정리하였으며 속성들의 범주형 속성값을 수치형으로 변환하여 실험하였다.

Soybean data는 기존 연구를 통해 군집의 개수는 4로 알려져 있다. 하지만 본 실험에서는 초기치의 개수를 정하지 않고 군집분석을 수행하였으며, 초기치 k 으로부터 군집을 생성, 갱신하도록 하였다. 초기치를 새롭게 갱신하는 기준인 θ는 95백분위수로 정하였다.

분류 정확도	10회			100회			1000회		
	$=4$		$k_0=1$	$k_0=4$		$k_0=1$	$k_0=4$		$k_0=1$
	기존	개선	제안	기존	개선	제안	기존	개선	제안
100%	1	5	3	16	57	20	223	659	171
97%			1			21			209
95%	2	4	3	11	31	26	98	253	288
90%			2	7	3	9	58	41	87
85%			1		1	20	8	12	172
80%						2			29
75%						2			44
70%	3	1		27	7		266	20	
65%					1		296	14	
60%	4			28			45	1	
50%				9			6		
40%									
평균	0.74	0.95	0.94	0.76	0.95	0.94	0.79	0.97	0.94
표준편차	0.027	0.008	0.004	0.026	0.007	0.004	0.023	0.005	0.003

Table 14. 알고리즘의 반복 수행 결과

Table 14는 제안 알고리즘에 대하여 초기치 k_0를 임의로 4개 선정하여 수행한 기존 k-modes 알고리즘과 유클리드거리를 기반으로 잘 선택된 초기치를 4개 선정하여 수행하는 개선 k-modes 알고리즘을 비교한 결과이다. 분류 정확도가 100%인 경우는 잘 선택된 초기치를 k_0를 4개 선정하여 군집하는 경우가 가장 좋은 결과를 보이고 있다. 97% 이상의 높은 정확도를 보인 경우에는 제안 알고리즘에서는 10회 반복 수행에서는 기존 알고리즘은 1회, 개선 알고리즘에서는 5회, 제안 알고리즘에서는 4회로 나타났다. 100회 반복 수행에서는 기존은 16회, 개선 알고리즘은 57회, 제안 알고리즘에서는 41회 나타났다. 1000회 반복에서는 기존 알고리즘은 223회, 개선 알고리즘은 659회, 제안 알고리즘은

380회 나타났다. 전반적으로 개선알고리즘이 분류정확도가 높게 나타났다. 하지만 개선 알고리즘은 Soybean Small 데이터의 군집이 4라는 정보를 기반으로 4개의 초기치를 선택하는 방법이다. 본 논문에서 제안 알고리즘은 초기치를 1개에서 최적의 4개로 접근하였다. 또한, 10회 반복 수행의 경우 정확도 평균은 74%이며 개선은 95%, 제안 알고리즘은 94%이며, 100회 반복 수행의 경우 기존 알고리즘은 76%, 개선 알고리즘은 95%, 제안 알고리즘은 94%이다. 1000회 반복 수행의 경우, 평균은 기존 알고리즘은 79%, 개선 알고리즘은 97%, 제안 알고리즘은 94%이다. 제안 알고리즘이 97%이상의 높은 분류 정확도에서는 개선 방식보다 다소 낮게 나타나고 있지만, 개선 알고리즘은 70%이하의 분류 정확도가 나타나는 반면, 제안 알고리즘에서는 75%이하의 분류 정확도는 나타나지 않고 있다. 결과적으로, 개선 알고리즘과 제안 알고리즘의 분류 정확도에서는 비슷하게 나타나고 있으며, 초기치 정보를 요구하는 기존 알고리즘과 개선 k-modes 알고리즘보다는 제안 Chain k-modes 알고리즘의 활용 가치가 높다고 판단 할 수 있다.

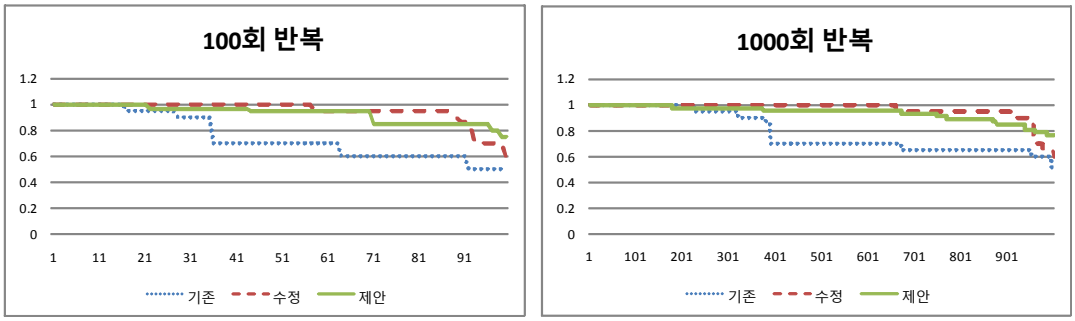


Figure 48. 100회, 1000회 반복 수행 결과 그래프

Figure 48은 Soybean data set에 대하여, 기존 k-modes 알고리즘과, 수정(개선 k-modes)알고리즘, 제안(Chain k-modes)알고리즘의 100회, 1000회 반복수행 결과이다. 제안하는 Chain k-modes 알고리즘은 군집에 대한 최적의 정보를 제공받는 수정(개선 k-modes)알고리즘과 비슷한 정확도를 보여주고 있으며, 이는 제안 알고리즘이 미지의 데이터에 대하여 보다 효율적이라고 할 수 있다.

4.4 알고리즘 결과 비교

4.4.1 Initial points refining algorithm

분류 정확도	기존 알고리즘	*itial points refining algorithm*	제안 Chain k-modes 알고리즘
95%~100%	7	14	16
85%~95%	2		4
0.75%~85%	3	5	
65%~75%	5		
55%~65%	3	1	
~55%			

Table 15. Initial points refining algorithm과의 분류 정확도 결과 비교

Table 15에서 *Initial points refining algorithm*은 좋은 결과를 보여주고 있다. 98%의 정확도를 보여주는 것이 14회 이상으로서 기존 $k-modes$ 알고리즘에 비해 *Initial points refining algorithm*은 좋은 결과를 보여주고 있다. 하지만 제안 알고리즘은 16회로 더욱 우수한 성능을 보여주고 있다.

4.4.2 k-representative algorithm

분류 정확도	$-representative\ algorithm$	제안 Chain k-modes 알고리즘
98%~100%	519	380
94% ~ 98%	87	497
89% ~ 94%	80	259
85% ~ 89%	89	
81% ~85%	86	
77% ~ 81%	94	29
72% ~ 77%	28	44
68% ~ 72%	8	
64% ~ 68%	7	
53% ~ 64%	2	

Table 16. k-representative algorithm과의 분류 정확도 결과 비교

Table 16에서는 $k-representative\ algorithm$의 1000회 반복 수행 결과를 보여주고 있으며, $k-representative\ algorithm$ 역시 초기치 정보를 미리 제공하고 군집분석하는 기법이다. 군집분석결과 98%이상의 정확도에서는 $k-representative\ algorithm$이 좀 더 높게 나타나고 있으나, 94%이상의 정확도에서는 제안 알고리즘이 높게 나타나고 있으며, $k-representative\ algorithm$은 70% 이하의 분류정확도가 나타나지만, 제안 알고리즘은 72% 이하의 낮은 분류 정확도는 나타나고 있지 않다.

5 결론

최근 범주형데이터에 대한 연구는 빅데이터에 대한 관심과 더불어 매우 중요한 이슈가 되고 있다. 실시간으로 나타나는 새로운 이슈를 빠르고 정확하게 분석해 낼 수 있어야 하며, 이러한 분석은 컴퓨터를 기반으로 하는 기계학습기법이 주를 이루고 있다. 이는 분석에 필요한 모든 과정과 정보를 스스로 수집하고 수행하는 것을 요구하고 있음을 의미한다. 이는 신경망(neural networks)분야와도 관련지을 수 있다. 데이터 마이닝은 데이터를 설명하거나 예측하는 다양한 분석기법에서 활용되며, 시간이 지날수록 좀 더 강력하고 우수한 분석도구를 필요로 하고 있다. 데이터 마이닝은 하나의 기법이 아닌 어떤 결과를 위한 프로세스 과정을 의미한다. 본 논문에서는 데이터마이닝기법 중 범주형데이터에 대한 군집분석 알고리즘의 전반적인 과정을 설명하고, 이에 필요한 세부적인 개념들을 연구하였다.

임의의 데이터를 분석하는데 있어 군집분석의 주요한 관심사는 '데이터가 몇 개의 군집으로 이루어져 있는가?' 이다. 하지만 우리가 접하고 있는 대다수의 데이터들은 이러한 정보가 부족하거나 주어지지 않는다. 군집의 개수에 대한 정보를 알고 있다면 우리는 매우 정확하고 효율적인 군집분석을 수행할 수 있으나, 군집의 개수를 예측하는 문제는 매우 민감한 문제이다.

본 논문에서는 범주형데이터에서의 군집분석을 위한 새로운 방식의 초기치 선정방법을 제안하였으며, 이를 기반으로 군집의 개수를 예측할 수 있는 방법을 제안하였다. 임의로 선정된 한 개의 초기치를 기반으로 다수의 mode를 생성시키고, 제안 알고리즘에 따라 병합, 갱신과정을 거쳐 최적의 군집으로 군집화 하였으며, 이를 통해 군집의 개수를 예측할 수 있었다.

초기 mode들의 속성 값들은 범주형데이터이며, 속성 값의 비율은 다항분포를 취한다. 이 다항분포의 비율은 각 mode에 대한 고유한 값으로서 속성의 개수만큼 다항분포는 존재한다. 우리는 데이터간의 비교 시 이러한 분포의 대표 값들을

비교함으로서 다른 mode들과의 유사도를 구할 수 있었다. 이를 초기 mode들 간의 유사성으로 정의하고 mode들을 병합 및 갱신을 통하여 최적의 군집을 찾을 수 있는 Chain k-modes 알고리즘을 제안하였다.

제안 Chain k-modes 알고리즘은 기존 알고리즘 및 다른 알고리즘과의 비교에서도 분류정확도면에서 동등하거나 그 이상의 성능을 보여주었다. 중요한 점은 제안 하는 알고리즘은 군집의 개수에 대한 정보가 주어지지 않았음에도 불구하고 우수한 성능을 보여준다는 점이다. 우리가 접하는 실제 데이터들은 군집의 개수가 정해지지 않으며, 상황에 따라 수시로 변할 수 있다. 기존 알고리즘 및 다른 알고리즘에서는 군집의 개수가 주어지지 않을 경우 알고리즘이 수행되지 않거나 임의로 군집의 개수가 주어질 경우 군집분석의 결과가 좋지 않거나 매우 낮은 정확도를 보이는 경우가 발생한다. 이는 대용량의 데이터 혹은 미지의 데이터를 분석하는데 있어 매우 치명적인 문제가 될 수 있다. 하지만, 제안 Chain k-modes 알고리즘에서는 군집의 수를 미리 정하지 않으면서 최적의 결과를 예측할 수 있다.

본 연구에서는 군집분석의 기본적인 조건이 만족하도록 알고리즘을 설계하였다. 데이터가 추가되면 유연하게 대처할 수 있도록, 주어진 데이터를 모두 분석하는 것이 아닌 반복 샘플링을 통해서 군집분석을 수행하였으며, 이는 새로운 성질의 데이터가 나타날 때도 적용하기가 유용하다. 또한, 제안 Chain k-modes 알고리즘은 임의의 군집 형태에 대해서도 군집이 가능하다. 수치형 속성 데이터에 대한 군집분석은 대부분 구형태가 아닌 경우 군집의 결과가 좋지 않다. 하지만 제안 알고리즘은 수치가 아닌 속성의 연관성을 활용하므로 군집의 모양이 일정하지 않아도 분석이 가능하다.

최근 데이터의 분석범위는 대용량 데이터를 넘어선 빅데이터 환경에서의 분석을 요구하고 있다. 빅데이터는 기존에서 다루고 있는 데이터를 다루는 방법이나 도구로는 처리할 수 없는 정형 또는 비정형의 방대한 데이터를 의미한다. 아직은 빅데이터의 정의에 대하여 합의된 점은 없다. 그래서 다양한 분야에서 빅데이터를 가정하여 분석하는 방법들이 제시되고 구현되어 있다. 최근 스마트폰이 사회 전반으로 빠른 속도로 널리 쓰이고 있는 시점에서, 소셜 미디어의 성장과 맞물려 다양한 데이터가 생성되고 있으며 이 데이터들을 활용하는 방법이 개발되고 있

다. 빅데이터의 형태는 일반적으로 범주형데이터라고 할 수 있으며, 실시간을 발생하는 시계열 데이터라고도 할 수 있다. 즉, 데이터의 양은 크고, 빠르게 발생하며, 다양한 데이터 속성을 내포하고 있음을 의미한다. 지금도 데이터 마이닝 분야에서의 군집분석기법은 이러한 데이터를 다룰 수 있는 기법으로 연구, 발전되고 있다. 하지만, 지금까지는 빅데이터를 효과적으로 군집분석 할 수 있는 기법보다는 빅데이터를 효과적으로 다룰 수 있는 도구의 개발이 많이 진행되고 있다. 대표적으로 하둡(Hadoop: High-Availability Distributed Object-Oriented Platform)을 예로 들 수 있다. 하둡(Apache Hadoop)은 대용량의 데이터를 처리할 수 있는 분산 응용 프로그램을 지원하는 자바 프레임워크이다. 하둡이 분석기법의 알고리즘을 뜻하지는 않는다. 하지만 데이터 마이닝의 특성상 하둡은 대용량의 데이터를 처리할 수 있는 매우 유용한 도구가 되어주고 있으며, 이를 기반으로 다양한 군집분석 알고리즘을 적용하여 활용할 수 있다. 본 논문에서는 대용량 데이터에 유연하게 대처할 수 있도록 분석하는 데이터에 대하여 샘플을 추출하며 초기 분석과 군집의 갱신에 대한 병렬처리가 가능하도록 설계하였다. 또한, 다양한 속성의 데이터를 다루기 위하여 가장 기본적인 빈도분석을 이용하여 속성의 연관성을 계산함으로서 범주형 속성 및 수치형 속성에도 적용이 가능하도록 하였다.

본 논문에서는 범주형 속성의 데이터에 대해서 데이터의 정보가 주어지지 않았을 경우에 대한 초기치 선정방법과 그에 따른 군집분석 방법을 제안하였으며, 우수한 결과를 보여주었다. 군집의 수를 정하는 문제는 요인분석(Factor Analysis)에서 최적의 요인의 개수를 구하는 것만큼 어려운 문제이다. 최적의 군집의 수를 예측하고 이를 분석할 수 있다면, 우리는 미지의 데이터에 대한 새로운 분석이 가능하다. 하지만 지금까지 알려진 대부분의 군집분석 기법들은 이에 대한 연구 및 연구결과가 다소 미흡하였다. 본 연구에서는 데이터의 속성을 범주형 속성에 관심을 두고 연구를 진행하였다. 수치형 속성에 대한 범주형 표현은 어렵지 않으나 범주형 속성에 대한 수치형 표현은 상당히 어려운 문제이다. 즉, 범주형 속성에 대한 분석이 가능하다면 수치형 속성에 대한 분석도 쉽게 변환할 수 있다. 또한, 제안 Chain k-modes 알고리즘은 예측하는 군집의 개수를 군집내 유사성과 군집간 상의성을 이용하여 주어진 데이터를 분석하는 과정에서

예측하였다. 시뮬레이션결과 제안 알고리즘은 사용자가 미리 정해놓은 정보보다 더 많은 정보를 추정할 수 있었다. 하지만, 제안 알고리즘은 데이터의 속성에 대하여 비교하는 회수가 기존 기법들 보다 증가하는 단점을 가지고 있다. 이는 병렬처리나 하드웨어의 개발로 극복할 수는 있으나, 이에 대한 개선이 필요하다. 또한, 본 연구는 범주형데이터에 대한 분석을 목적으로 하였으며, 향후 수치형데이터 및 현실에서 나타나는 수치형데이터와 범주형데이터가 섞여있는 혼합형데이터(Mixed Data)에 대한 연구도 진행되어야 할 것이다.

6 참고문헌

Ahmad, Amir, and Shehroz S. Khan. "Generating K-Prototype Points for Unsupervised Learning." IICAI, (2003).

Ahmad, Amir, and Lipika Dey. "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set." Pattern Recognition Letters 28.1, 110-118, (2007).

Ahmad, Amir, and Lipika Dey. "A k-mean clustering algorithm for mixed numeric and categorical data." Data & Knowledge Engineering 63.2, 503-527, (2007b).

Ahmad, Amir, and Lipika Dey. "A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets." Pattern Recognition Letters 32, 1062-1069, (2011).

Ali Seman, Zainab Abu Bakar, Azizian Mohd. Sapawi. "Centre-based Clustering for Y-Short Tandem Repeats (Y-STR) as Numerical and Categorical data." IEEE Vol. 1, Issue 1, (2010).

Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm." (1997).

Andritsos, Periklis, et al.. "LIMBO: Scalable clustering of categorical data." Advances in Database Technology-EDBT 2004. Springer Berlin Heidelberg, 123-146, (2004).

Anil Chaturvedi, Kraft Foods, Paul E. Green. "K-modes Clustering.", Journal of Classification 18, 35-55, (2001).

Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm.", (1997).

Athitsos, Vassilis, and Stan Sclaroff. "Boosting nearest neighbor classifiers for multiclass recognition." Computer Vision and Pattern Recognition-Workshops, CVPR Workshops. IEEE Computer Society Conference on IEEE, (2005).

Berry and Linoff. "Data mining techniques: For marketing, sales, and customer support.", Book(ISBN:0471179809), (1997).

Bradley P. "Refining initial points for k-means clustering." - In: Proc, 15th Internat.Conf.on Machine Learning.Morgan Kaufmann, Los Altos, CA, (1998).

Bradley, P. S., K. P. Bennett, and Ayhan Demiriz. "Constrained k-means clustering." Microsoft Research, Redmond, 1-8, (2000).

Bradley P. "Refining initialization of clustering algorithm" — In: Ahsl, A.(Ed.) ,Proc.4th Internat. Conf.on Knowledge Discovery and Data Mining.AAAI Press, New York, (1998b).

Breunig, Markus M., et al.. "LOF: identifying density-based local outliers." ACM Sigmod Record. Vol. 29. No. 2. ACM, (2000).

Brito, M. R., et al.. "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection." Statistics & Probability Letters 35.1, 33-42, (1997).

Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti. "Enhancing density-based clustering: parameter reduction and outlier detection", Information System 38, 317-330. (2013).

Chaturvedi, Anil, Paul E. Green, and J. Douglas Caroll. "K-modes clustering." Journal of Classification 18.1, 35-55, (2001).

Chen, Hailin, Xiuqing Wu, and Junhua Hu. "Adaptive K-means clustering algorithm." International Symposium on Multispectral Image Processing and Pattern Recognition. International Society for Optics and Photonics, (2007).

Chen, Hung-Leng, Kun-Ta Chuang, and Ming-Syan Chen. "On data labeling

for clustering categorical data." Knowledge and Data Engineering, IEEE Transactions on 20.11, 1458-1472, (2008).

Cho, Gilsoo, Namkyoo Lim, and Yoonjung Yang. "Integrated graphical presentation of fabric sound and mechanical properties." Fibers and Polymers 11.3, 516-520, (2010).

Chu, Yi-Hong, et al. "Reducing redundancy in subspace clustering." Knowledge and Data Engineering, IEEE Transactions on 21.10, 1432-1446, (2009).

Cordeiro de Amorim, Renato, and Boris Mirkin. "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering." Pattern Recognition 45.3, 1061-1075, (2012).

Cost, Scott, and Steven Salzberg. "A weighted nearest neighbor algorithm for learning with symbolic features." Machine learning 10.1, 57-78, (1993).

Dae-Won Kim, Yigeun Chae. "More Efficient k-Modes Clustering Algorithm." Journal of Korea Data & Information Science Society, Vol. 16. No. 3, 549-556, (2005).

Day, William HE, and Herbert Edelsbrunner. "Efficient algorithms for agglomerative hierarchical clustering methods." Journal of classification 1.1, 7-24, (1984)

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society. Series B (Methodological), 1-38, (1977).

DH. Bergel and WR. Milnor. "Pulmonary vascular impedance in the dog." Circulation Research, 16, 401-415, (1965).

Duan, Kai-Bo, and S. Sathiya Keerthi. "Which is the best multiclass SVM method? An empirical study." Multiple Classifier Systems. Springer Berlin Heidelberg, 278-285, (2005).

Equitz, William H. "A new vector quantization clustering algorithm." Acoustics, Speech and Signal Processing, IEEE Transactions on 37.10, 1568-1575, (1989).

F. Hoppner, F.Klawonn, R. Kruse and T. Runkler, "Fuzzy cluster analysis." wiley, (2000).

Fuyuan Cao, Jiye Liang and Liang Bai, "A new initialization method for categorical data clustering." Expert System with Application, Vol. 36, No. 7, 102273-102284, September, (2009).

F. Samadzadegan and S. Saeedi, "Clustering Of Lidar Data Using Particle Swarm Optimizatio Algorithm In Urban Area." (2009).

Fayyad, Usama, Gregory Piatetsky-Shapiro, Padhraic Smyth, "From data mining to knowledge discovery in databases." AI magazine 17.3, 37, (1996).

Fayyad, Usama M., Cory Reina, and Paul S. Bradley. "Initialization of Iterative Refinement Clustering Algorithms." KDD, (1998).

Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim, "ROCK: A robust clustering algorithm for categorical attributes." Information systems 25.5, 345-366, (2000).

Gan G. Ma C. & Wu, J, "Data clustering: Theory, algorithms, and applications." Society for Industrial and Applied Mathematics (SIAM), (2007).

Gustafson D.E. & Kessel, W.C, "Fuzzy clustering with a Fuzzy Covariance Matrix." Proceedings IEEE on Decision and Control, 761-766, (1979).

Hamerly, Greg, and Charles Elkan, "Alternatives to the k-means algorithm that find better clusterings." Proceedings of the eleventh international conference on Information and knowledge management. ACM, (2002).

Hautamaki, Ville, Ismo Karkkainen, and Pasi Franti, "Outlier detection using

k-nearest neighbour graph." Pattern Recognition, ICPR 2004. Proceedings of the 17th International Conference on. Vol. 3. IEEE, (2004).

Hartigan, J. A., and Manchek A. Wong, "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1, 100-108, (1979).

Hartigan, J. A., Wong, M. A., "A K-Means Clustering Algorithm." Journal of the Royal Statistical Society, Series C 28 (1), 100-108, (1979).

Hautamaki, Ville, et al., "Improving k-means by outlier removal." Image Analysis. Springer Berlin Heidelberg, 978-987, (2005).

Han, Jiawei, Micheline Kamber, and Jian Pei, "Data mining: concepts and techniques." Morgan kaufmann, (2006).

Hand D.J, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining." MIT Press, (2001).

He, Zengyou, Xiaofei Xu, and Shengchun Deng, "Clustering mixed numeric and categorical data: A cluster ensemble approach." arXiv preprint cs, (2005).

Hee Chang Park, Sun Myung Lee, "K-means Clustering using Grid-based Representatives.", Journal of Korean Data& Information Science Society, Vol.16, No.4, 759-768, (2005).

Huang Z, "Clustering large data sets with mixed numeric and categorical values." Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD), 21-34, (1997)

Huang Z, "A fast clustering algorithm to cluster very large categorical data sets in data mining." Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, The University of British Columbia, Canada, pp.1-8, (1997b)

Huang Z, "Extensions to the k-means algorithm for clustering large data

sets with categorical values." Data mining Knowledge, Vol2, No.2, 283 -304, (1998).

Huang Z, Michal K, "A fuzzy k-modes algorithm for clustering categorical data." IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 7, NO. 4, AUGUST, (1999).

Huang Z, "A Note on K-modes Clustering", Journal of Classification 20, 257-261, (2003).

Jarman, Ian H., et al., "Clustering categorical data: A stability analysis framework." Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, (2011).

J.C.Bezdek, "Cluster validity with fuzzy sets." J. Cybernit, vol,3, 58-72, (1974).

J.C.Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms." New York: Plenum Press, ISBN:0306406713, (1981).

Kaufman, L. and Rousseeuw P.J., "Clustering by means of Medoids, in Statistical Data Analysis Based on the L1-Norm and Related Methods." edited by Y. Dodge, North-Holland, 405-416, (1987).

Khan, Shehroz S., and Shri Kant., "Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation." IJCAI, (2007).

Kim, Dae-Won, et al., "A k-populations algorithm for clustering categorical data." Pattern recognition 38.7, 1131-1134, (2005).

Kim, Jae-Hyun, and Hoh-Yoo Baek., "Excel macro for applying Bayes' rule." Journal of the Korean Data and Information Science Society 22.6, 1183-1197, (2011).

Kriegel, Hans-Peter, et al., "Density-based clustering." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1.3, 231-240, (2011)

Laughlin J, "Perception of texture , visually and tactually , and tactually: An

exploratory study using multidimensional scaling analysis." International Journal of Clothing Science and Technology 3.1, 28-36, (1991).

Lee, Yeong-Chyi, Tzung-Pei Hong, and Wen-Yang Lin, "Mining association rules with multiple minimum supports using maximum constraints." International Journal of Approximate Reasoning 40.1, 44-54, (2005).

Lee, Shin-Won, Dong-Un An, and Sung-Jong Chong, "Selection of Cluster Hierarchy Depth and Initial Centroids in Hierarchical Clustering using K-Means Algorithm." Journal of the Korean Society for information Management 21.4, 173-185, (2004).

Le, Si Quang, and Tu Bao Ho, "An association-based dissimilarity measure for categorical data." Pattern Recognition Letters 26.16, 2549-2557, (2005).

Mackay C, Anand S.C, Bishop D, "Effects of Laundering on the Sensory and Mechanical Properties of 1 × 1 Rib Knitwear Fabrics: Part I: Experimental Procedures and Fabric Dimensional Properties", Textile Research Journal March, 66: 151-157, (1996).

Mackay C, Anand S.C, Bishop D, "Effects of Laundering on the Sensory and Mechanical Properties of 1×1 Rib Knitwear Fabrics: Part II: Changes in Sensory and Mechanical Properties", Textile Research Journal March, 69: 252-260, (1999).

MacQueen, James, "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 281-297, (1967).

Michalski, Ryszard S., and Robert E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy." Pattern Analysis and Machine Intelligence, IEEE Transactions on 4,

396-410, (1983).

Miller, Harvey J., and Jiawei Han, eds., "Geographic data mining and knowledge discovery." CRC Press, (2009).

Mingoti, Sueli A., and Renata A. Matos, "Clustering Algorithms for Categorical Data: A Monte Carlo Study." International Journal of Statistics and Applications 2.4, 24-32, (2012).

M.K. Ng. M.J. Li. J.Z. Huang. & Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm." IEEE Transactions of Pattern Analysis and Machine Intelligence. 29(3), 503-507, (2007).

Moon, Todd K, "The expectation-maximization algorithm." Signal processing magazine, IEEE 13.6, 47-60, (1996).

Moth'd Belal. Al-Daoud, "A New Algorithm for Cluster Initialization." World Academy of Science, Engineering and Technology 4, (2007).

M. Dutta, A.Kakoti Mahanta, Arun K. Pujari, "QRock : A quick version of the ROCK algorithm for clustering of categorical data.", Pattern Recognition Letters 26, 2364-2373, (2005).

N.M.Goodall, 1996, "A New Similarity Index Based on Probability.", in Biometrics, vol. 22, 882-907, (1996).

Nam H, "k-priorities : An Efficient Clustering algorithm for Categorical Data Sets." KIST 석사학위논문, (2002).

Nazeer, KA Abdul, and M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." Proceedings of the World Congress on Engineering. Vol. 1, (2009).

Ng, Michael K., and Joyce C. Wong, "Clustering categorical data sets using tabu search techniques." Pattern Recognition 35.12, 2783-2790, (2002).

Ng,R.t. and Han,J., "Efficient and Effective Clustering Methods for Spatial Data Mining." Proceedings of the 20th VLDB Conference, Santiago, Chile, 144-155, (1994).

Ng R. & Han, J., "Efficient and effective clustering methods for spatial data mining." Proceedings of the 20th international conference on very large databases, Santiago, Chile. 144-155, (1994b).

Ohn Mar San, Van-Nam HUYNH, Yoshiteru Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data.", Int. J. Appl, Math. Compute. Sci, vol.14, No.2, 241-247, (2004).

Otey, Matthew Eric, Amol Ghoting, and Srinivasan Parthasarathy, "Fast distributed outlier detection in mixed-attribute data sets." Data Mining and Knowledge Discovery 12.2-3, 203-228, (2006).

Pal, Nikhil R., et al., "A possibilistic fuzzy c-means clustering algorithm." Fuzzy Systems, IEEE Transactions on 13.4, 517-530, (2005).

Peng Zhang, Xiaogang Wang, and Peter X.k. Song, "Clustering categorical data based on distance vectors.", Journal of the American Statistical Association , March, vol. 101, No.473, (2006).

P.A. Vijaya, M. Narasimha, D.K. Subramanian, "An efficient hierarchical clustering algorithm for large data sets." Pattern Recognition Letters 25, 505-513, (2004).

P.H.S.Torr, "Filtering Using a Tree-Based Estimator." The 9th International Conference on Computer Vision(ICCV'03), 1063-1070, (2003).

Philips S. "Acceleration of k-means and related clustering algorithms." In Mount, D and Stein, C, editors, ALENEX: International workshop on algorithm engineering and experimentation, LNCS, 2409, 166-177, (2002).

Pena, Jose Manuel, Jose Antonio Lozano, and Pedro Larranaga, "An empirical comparison of four initialization methods for the K-Means algorithm." Pattern recognition letters 20.10, 1027-1040, (1999).

Ralambondrainy, H., 1995, "A conceptual version of the K-Means algorithm.", Pattern Recognition Letters, 16, 1147-1157, (1995).

Rishi Syal , Dr V.Vijaya Kumar, "Innovative Modified K-Mode Clustering

Algorithm." ISSN, Vol.2, Issue 4, July-August, 390-398, (2012).

Roy, Dharmendra K., and Lokesh K. Sharma, "Genetic k-means clustering algorithm for mixed numeric and categorical data sets." International Journal of Artificial Intelligence & Applications 1.2, 23-28, (2010).

Smith, Lindsay I., "A tutorial on principal components analysis." Cornell University, USA 51, 52, (2002).

San, Ohn Mar, Van-Nam Huynh, and Yoshiteru Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data." International Journal of Applied Mathematics and Computer Science 14.2, 241-248, (2004).

Selim, Shokri Z., and Mohamed A. Ismail., "K-means type algorithms: a generalized convergence theorem and characterization of local optimality." Pattern Analysis and Machine Intelligence, IEEE Transactions on 1, 81-87, (1984).

Seman, Ali, Z. Abu Bakar, and A. M. Sapawi, "Centre-based clustering for Y-Short Tandem Repeats (Y-STR) as Numerical and Categorical data." Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on. IEEE, (2010).

Stegmann, Mikkel Bille, "Analysis and segmentation of face images using point annotations and linear subspace techniques." (2002).

Stenger, Bjoern, et al., "Filtering using a tree-based estimator." Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, (2003).

Sundberg, Rolf, "An iterative method for solution of the likelihood equations for incomplete data from exponential families." Communication in Statistics-Simulation and Computation 5.1, 55-64, (1976).

Sun, Ying, Qiuming Zhu, and Zhengxin Chen, "An iterative initial-points refinement algorithm for categorical data clustering." Pattern Recognition Letters 23.7, 875-884, (2002).

Taous-Meriem Laleg, Michel Sorine and Qinghua Zhang, "Input Impedance of the Arterial System Using Parametric Models" ICES, (2006).

Tapas Kanung, "The Analysis of a Simple k-Means Clustering Algorithms." in Proceedings of ACM symposium on Computational geometry. Hongkong, June 12-14, (2000).

Thayananthan, Arasanathan, et al., "Learning a Kinematic Prior for Tree-Based Filtering." BMVC, (2003).

Thu-Hien Thi Nguyen, Van-Nam Huynh, "Extending k-Representative Clustering Algorithm with an Information Theoretic-based Dissimilarity Measure for Categorical Objects." school of Knowledge Science Japan Advanced Institute of Science and Technology, Japan.

UCI Data Repository,
http://www.sgi.com/tech/mlc/db/

Wagstaff, Kiri, et al., "Constrained k-means clustering with background knowledge." ICML. Vol. 1, (2001).

Wikipedia, The Free Encyclopedia,
http://www.wikipedia.org

Wilpon, J., and L. Rabiner, "A modified K-means clustering algorithm for use in isolated work recognition." Acoustics, Speech and Signal Processing, IEEE Transactions on 33.3, 587-594, (1985).

Xu, Xiaowei, et al., "A distribution-based clustering algorithm for mining in large spatial databases." Data Engineering, Proceedings, 14th International Conference on. IEEE, (1998).

Yang, S. Y., D. H. Yuan, and G. M. Lai, "Refining initial points for k-means clustering." Computer Science and Engineering, (2007).

Z. He. X. Xu. S. Deng, "Attribute Value Weighting in k-Modes Clustering" Computer Science e-Prints: arXiv: cs/0701013v1 [cs.AI], Cornell University Library, Cornell University, Ithaca, NY, USA, v1, 1-15, (2007).

Zhang, Tian, Raghu Ramakrishnan, and Miron Livny, "BIRCH: an efficient data clustering method for very large databases." ACM SIGMOD Record. Vol. 25. No. 2. ACM, (1996).

Zhang, Dao Q., and Song C. Chen, "Kernel-based fuzzy and possibilistic c-means clustering." Proceedings of the International Conference Artificial Neural Network, (2003).

김만선, 이상용, “대용량 데이터 처리를 위한 하이브리드형 클러스터링 기법.”, 정보처리학회논문지 8, 제10-B권 제1호, (2003).

백장선, 심정욱. 2000, "K-평균 군집방법을 이용한 가중커널분류기." 응용통계연구 13.2, 447-455, (2000).

오수민, 송준모, 김철수, "범주형데이터 분석에서 속성의 영향력을 이용한 군집분석." Journal of KIISE: Computing Practices and letters, VOLUME 18, NUMBER 11, NOVEMBER, (2012).

오수민, 김철수, "기존 k-modes 알고리즘 개선을 통한 범주형 속성 데이터에 대한 효율적인 클러스터링 방법." 제주대학교 석사학위논문, (2006).

이동헌, 전우제, and 박수홍, "K 평균 군집화를 이용한 벡터 데이터 압축 방법." 한국공간정보시스템학회 논문지 7.3, 104-118, (2005).

이상용, "대용량 데이터 처리를 위한 하이브리드형 클러스터링 기법." 두뇌한국21, (2002).

이희우, 신선미, "데이터 마이닝을 이용한 서울시 교직원의 피료요인 탐색연구." 한국학교보건학회지, 제19권, 제1호. 79-88, (2006).

허경용, 우영운, 김광백, “Possibilistic Fuzzy C-means 클러스터링 알고리즘의 확장” Proceedings of KFIS autumn conference, Vol, 17, No. 2, (2007).

허명회, "이중 K-평균 군집화", 응용통계연구, 제13권, 2호, 343-352, (2000).

허명회, "k-평균 군집화와 재현성 평가 및 응용" 응용통계연구, 제17권 1호, 135-144, (2004).