대용량 데이터 처리를 위한 하이브리드형 클러스터링 기법

김 만 선† · 이 상 용††

요 약

데이터 마이닝은 지식발견 과정에서 중요한 역할을 수행하며, 여러 데이터 마이닝의 알고리즘들은 특정의 목적을 위하여 선택될 수 있다. 대부분의 전통적인 계층적 클러스터링 방법은 적은 양의 데이터 집합을 처리하는데 적합하여, 제한된 리소스와 부족한 효율성으로 인하여 대용량의 데이터 집합을 다루기가 곤란하다. 본 연구에서는 대용량의 데이터에 적용되어 알려지지 않은 패턴을 발견할 수 있는 하이브리드형 신경망 클러스터링 기법인 PPC(Pre-Post Clustrering) 기법을 제안한다. PPC 기법은 인공지능적 방법인 자기조직화지도(SOM)와 통계적 방법인 계층적 클러스터링을 결합하여 두 과정을 거쳐 데이터를 군집화한다. 전처리 클러스터링 과정에서는 자기조직화지도를 통해 데이터를 요약한다. 그리고나서 후처리 클러스터링 과정에서는 군집의 내부적 특징을 나타내는 응집거리와 군집간의 외부적 거리를 나타내는 인접거리에 따라 유사도를 측정한다. 최종적으로 PPC 기법은 측정된 유사도를 이용하여 대용량 데이터 집합을 군집화한다. PPC 기법은 UCI Repository 데이터를 이용하여 실험해 본 결과, 다른 클러스터링 기법들 보다 우수한 응집도를 보였다.

A Hybrid Clustering Technique for Processing Large Data

Man-Sun Kim† · Sang-Yong Lee††

ABSTRACT

Data mining plays an important role in a knowledge discovery process and various algorithms of data mining can be selected for the specific purpose. Most of traditional hierachical clustering methods are suitable for processing small data sets, so they have difficulties in handling large data sets because of limited resources and insufficient efficiency. In this study we propose a hybrid neural networks clustering technique, called PPC for Pre-Post Clustering that can be applied to large data sets and find unknown patterns. PPC combinds an artificial intelligence method, SOM and a statistical method, hierarchical clustering technique, and clusters data through two processes. In pre-clustering process, PPC digests large data sets using SOM. Then in post-clustering, PPC measures similarity values according to cohesive distances which show inner features, and adjacent distances which show external distances between clusters. At last PPC clusters large data sets using the simularity values. Experiment with UCI repository data showed that PPC had better cohensive values than the other clustering techniques.

키워드 : 신경망 클러스터링(NN clustering), 자기조직화지도(SOM), 계층적 클러스터링(Hierarchical clustering), PPC 기법(PPC), 데이터 마이닝(Data Mining)

1. 서 론

지식탐사 프로세스의 핵심적인 역할을 담당하는 데이터 마이닝 단계에서는 여러 가지 목적에 따라 알고리즘을 선택하여 사용한다. 데이터 마이닝에서 클러스터링 방법은 기존의 통계적 기법, 기계학습, 패턴인식에 쓰이던 방법이 사용되어 왔다. 그런데 최근에 와서 부가적으로 데이터베이스 지향적인 제약 사항들을 첨가시켜, 다양하고 다차원적인 데이터를 효율적으로 분류해 나가기 위한 방안들이 연구되고 있다[1].

클러스터링은 입력 데이터 집합을 유사한 관찰값들의 군집들로 구분하여 데이터집합 속에 존재하는 의미 있는 정보를 얻는 과정이다[2]. 즉, 군집내의 유사성은 최대화하고, 군집들 간의 유사성은 최소화하도록 데이터 집합을 분할하는 것이다[3]. 이러한 군집 발견 과정은 우리에게 군집 데이터 분포가 갖고 있는 특징을 설명하며, 다른 분석 기법을 위한 토대를 마련 해주는 역할을 수행 할 수 있다[4].

클러스터링 기법은 기업의 고객을 구매 패턴에 근거해서 분류하거나, GIS에서 공간 상의 군집을 분류하거나, 웹 문서의 로그를 통해 비슷한 접근 패턴을 그룹화 하거나, exploratory clustering of gene expression profiles[5] 등 다양한 응용 분야에 적용이 가능하다. 데이터 마이닝의 출현으로 인해 대용량 데이터를 대상으로 원시 데이터에 대한 접근 횟수를 줄이고 알고리즘이 다루어야 할 데이터 구조의 크기를 줄이는 클러스터링 기법에 관한 연구들이 활발하다.

본 연구에서는 대용량 데이터에 적용하여 효율적으로 양질의 군집을 발견할 수 있도록 하는 클러스터링 기법인 PPC

※ 본 연구는 2002년도 두뇌한국21 지원에 의하여 수행되었음.
† 준 회 원 : 공주대학교 대학원 컴퓨터공학과
†† 종신회원 : 공주대학교 정보통신공학부 교수
논문접수 : 2002년 5월 24일, 심사완료 : 2002년 11월 21일

(Pre-Post Clustering) 기법을 제안한다. 대부분의 계층적 클러스터링 알고리즘은 소량의 데이터를 대상으로 비슷한 크기를 갖는 구(sphere)형의 군집들로 데이터를 나누는 경향이 있다. PPC 기법은 이러한 한계를 극복하고 대용량의 데이터에 존재하는 다양한 모양의 군집을 효율적으로 발견할 수 있는 하이브리드형 클러스터링 기법이다.

PPC 기법은 인공지능적 방법과 통계적 클러스터링 방법을 접목하여 두 단계를 거쳐 클러스터링을 수행한다. 첫 번째 단계는 초기의 소군집을 발견하기 위한 단계로, 대용량의 데이터에 효율적으로 적용할 수 있는 인공지능적 방법인 자기조직화지도(SOM)를 이용한다. 그리고 두 번째 단계는 통계적 클러스터링 방법인 다양한 형태의 군집을 발견할 수 있는 계층적 클러스터링을 이용하여 첫 번째 단계에서 발견된 초기 소군집들을 반복적으로 병합하여 최종의 군집을 얻는다.

2. 관련 연구

클러스터링 알고리즘은 일반적으로 통계적 클러스터링 방법과 인공지능적 클러스터링 방법의 두 가지로 분류할 수 있다.

2.1 통계적 클러스터링 방법

통계적 클러스터링 방법에는 계층적 클러스터링과 분할적 클러스터링이 사용된다.

2.1.1 분할적 클러스터링

분할적 클러스터링(partitioning clustering)은 계층적 클러스터링과 달리 중첩된 분할 계층구조가 아닌 평평한 하나의 분할 구조로 형성된 군집을 생성한다.

여러 가지 분할적 클러스터링 알고리즘 중에서 k-means는 유클리디안 거리(Euclidean distance)를 이용하여 가깝게 위치한 점들을 찾아 군집으로 묶어주는 기법으로 차원의 제약이 전혀 없고 간단하다는 장점 때문에 널리 사용되는 방법이다.

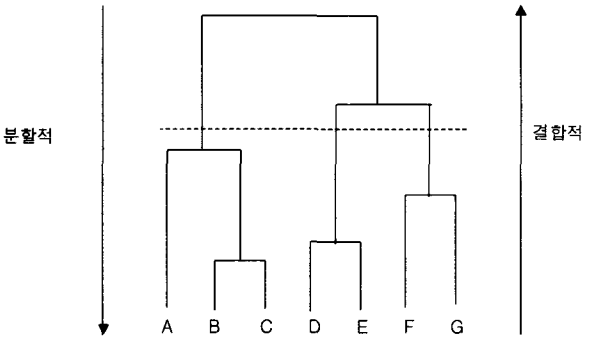
k-means는 임의의 초기 분할로부터 시작하여 군집의 중심값과 데이터 개체들과의 유사도에 근거하여 목적함수가 수렴 조건을 만족할 때까지 데이터의 소속 군집을 재할당한다. 이렇게 발견된 군집은 중심값(centroid)으로 표현되는데, 이는 해당군집에 속하는 데이터들의 평균 혹은 중앙값이다.

분할적 클러스터링은 군집이 벡터 평면상에서 구(sphere)의 형태를 가지고 있어야 효율이 좋다고 알려져 있다. 그러나 군집의 결과가 항상 구의 형태를 나타낸다고 볼 수 없기 때문에 본 논문에서는 계층적 클러스터링을 사용하였다.

2.1.2 계층적 클러스터링

계층적 클러스터링(hierarchical clustering)은 가장 유사한 두 개체들을 선택하여 병합해 가는 방법과 가장 먼 개체들을 선택하여 나누어 나가는 방법이 있다. 두 군집의 유사도를 측정하는 기준에 따라 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법 등으로 나뉜다.

계층적 클러스터링은 상향식 방법(bottom-up)인 병합적 계층군집 방법과 하향식 방법(top-down) 방법인 분할적 계층적 군집방법이 있다. 각각의 모든 데이터 개체가 하나의 분할을 이루는 최하위 계층에서부터 모두 하나의 군집으로 합쳐진 최상위 계층까지 (그림 1)과 같은 중첩된 분할의 계층순서를 생성한다.



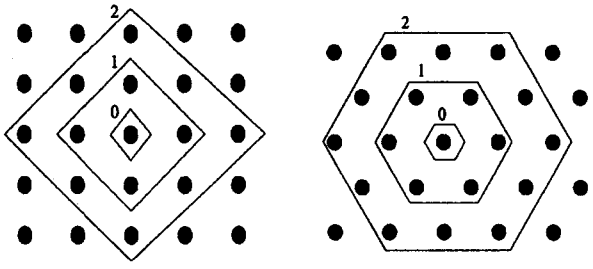
(그림 1) Dendrogram

2.2 인공지능적 방법

인간 두뇌의 현상에 가장 가깝다고 할 수 있는 인공신경망에는 자기조직화지도(SOM)와 ART가 있다. ART는 실세계를 시뮬레이션할 수 있을 정도로 정교하지 못하나, 자기조직화지도는 모델을 구축하여 클러스터링을 수행하는 방법으로 생물학적 신경망을 모형화한 인공신경망의 일종이다.

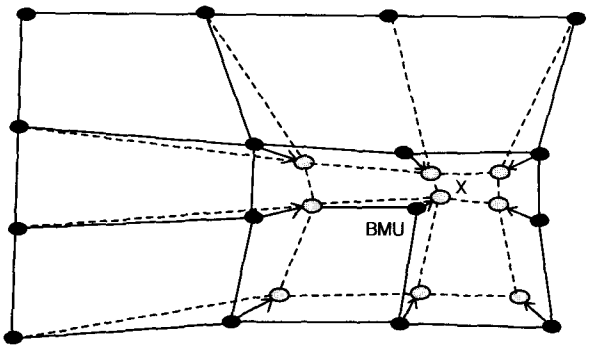
자기조직화지도는 입력층과 격자모양의 뉴런들로 구성된 경쟁층으로 이루어져 있다[6]. 경쟁층의 각 뉴런은 d차원의 연결강도 벡터($m = [m_1, m_2, \cdots m_d]$)로 표현되는데 모든 뉴런들은 위상적 연관성을 나타내는 이웃관계에 의해 이웃한 뉴런들과 연결되어있다. 이러한 뉴런들의 관계는 위상지도로 나타나게 된다.

뉴런들간의 이웃관계 구조는 그 형태에 따라 (그림 2)와 같이 사각형과 육각형 격자구조가 있다[7].



(그림 2) 이웃관계를 정의하기 위한 격자의 유형

자기조직화지도의 학습과정은 신경망의 연결강도 벡터들을 입력패턴과 유사해지도록 조정하기 위해 경쟁식 학습(competitive learning)방법을 사용한다. 이러한 학습방법은 On-line k-means와 유사하지만 가장 두드러진 차이점은 입력벡터와 가장 유사한 연결강도 벡터를 갖는 뉴런인 승자뉴런뿐만 아니라, 위상적으로 주변의 이웃관계에 있는 뉴런의 연결강도 벡터까지도 조정된다는 점이다. 즉, 승자뉴런(BMU : Best Matching Unit) 주변의 연결강도도 (그림 3)과 같이 x를 향하여 학습데이터와 유사해지는 것이다. 따라서 최종의 지도상에 유사한 연결강도 벡터들을 갖는 뉴런들이 이웃하도록 정렬되는 것이다.



(그림 3) 승자뉴런과 이웃 뉴런들의 연결강도 갱신

자기조직화지도의 학습과정은 반복적으로 수행되는데 매 단계에서 입력데이터 집합으로부터 임의의 입력벡터 x를 선택하여 x와 모든 연결강도벡터 사이의 거리를 계산한다[8].

이 때 거리는 일반적으로 유클리디안 거리인 식 (2.3)이 사용된다. 승자뉴런을 찾은 후 자기조직화지도의 승자뉴런의 연결강도 벡터와 이웃 뉴런의 연결강도 벡터는 식을 이용하여 식 (2.3)과 같이 입력벡터와 유사해지도록 갱신된다.

$$m_i(t+1) = m_i(t) + \alpha h_{ci}(t)[x(t) - m_i(t)] \qquad (2.3)$$

t : 시간 $x(t)$: t시점의 입력벡터
h_{ci} : 이웃함수 α : 학습율

이웃함수 $h_{ci}(t)$는 학습이 진행됨에 따라 이웃의 반경이 감소하는 함수로 일반적으로 식 (2.4)와 같은 가우시안 함수를 사용한다.

$$h_{ci}(t) = \exp[-(\frac{d_{ci}^2}{2s_t^2})] \qquad (2.4)$$

d_{ci} : 뉴런 c와 i의 거리 s_t : t시점의 이웃반경

2.3 대용량 데이터 처리에 중심을 둔 기법

본 연구에서는 통계적 클러스터링 방법과 인공지능적 방법을 접목하여 대용량 데이터를 처리하고자 한다.

최근 대용량의 데이터 처리에 중심을 두고 표본추출 기법을 이용하거나, 요약된 군집표현을 이용하거나, 특별한 자료구조를 이용하는 등 새로운 클러스터링 알고리즘이 발표되고 있다.

표본추출 기법을 이용한 CURE[9]는 각 군집으로부터 잘 분포된 몇 개의 데이터 개체를 선정하고 이 점들이 군집의 중심값을 향해 일정 비율 모이게 하여 이상치의 영향을 감소시킬 수 있으나, 계층적 클러스터링 중 최단 연결법을 이용하여 대표값 1개를 사용해 군집간의 유사도를 판별한다.

요약된 군집표현을 이용한 BIRCH[10]는 원시 데이터를 직접 다루지 않고 군집에 속한 데이터 개체의 수, 개체들의 선형합, 개체들의 제곱 합으로 구성된 군집의 요약정보인 군집 특성(cluster feature)을 이용한다. 전처리 과정에서 CF 트리를 적용시킨 후, 계층적 클러스터링을 수행하는 단계를 거친다. 그러나 계층적 클러스터링 기법을 수행하므로 큰 군집은 작게 쪼개고 작은 군집은 합쳐지는 현상이 일어나는 단점이 있다.

특별한 자료 구조를 이용한 Chameleon은 계층적 클러스터링 기법에 기반하여 두 단계로 구성된다. 다양한 특징을 띠는 군집간 유사도를 측정할 수 있는 새로운 동적 모델을 제시하였고, 다양한 모양, 밀도, 크기를 갖는 자연스런 군집을 발견할 수 있다는 특징을 갖고 있다. 그러나 원시 데이터를 직접 다루기 때문에 수행속도 면에서 현저한 성능 저하가 약점이다.

3. PPC 기법

본 논문에서 제안하는 하이브리드형 클러스터링 기법인 PPC (Pre-Post Clustering) 기법의 특징, 전후 클러스터링 과정을 설명하고 시간복잡도를 분석한다.

3.1 PPC 기법의 특징

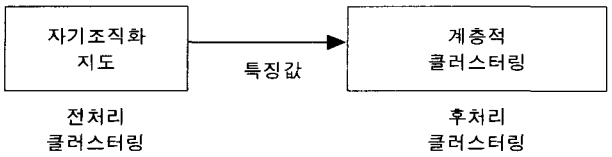
PPC 기법은 인공지능적 기법인 자기조직화지도와 통계적 클러스터링 방법인 계층적 클러스터링을 결합하여 두 방법의 장점을 이용한다. 자기조직화지도로 클러스터링을 수행하여 대용량 데이터를 특징값으로 표현되는 소군집들로 요약하고, 이 특징값 정보를 이용해 다양한 특징을 갖는 군집을 발견할 수 있는 계층적 클러스터링을 수행한다.

자기조직화지도는 데이터 집합을 비슷한 크기의 원형 군집들로 분할하는 단점이 있지만, 데이터 수에 선형 비례하는 시간복잡도 O(nkl)를 갖기 때문에 대용량의 데이터에 적용 가능하다. 특히 온라인 학습방법을 이용할 경우 많은 양의 메모리 사용없이 연결강도 벡터들과 현재 학습 벡터만을 위한 공간만 있으면 된다.

계층적 클러스터링 기법은 데이터 수(n)의 제곱에 비례하는 시간복잡도 $O(n^2)$를 갖기 때문에 계산량이 많지만, 군집 간의 유사도 측정 방법에 따라 다양한 특징을 갖는 군집을 발견

할 수 있다.

PPC 기법은 (그림 4)와 같이 두 단계를 거쳐 클러스터링을 수행한다.



(그림 4) PPC 기법의 두 단계

첫 번째 단계는 초기의 소군집을 발견하기 위한 단계로 자기조직화지도를 이용하고, 소군집에서 특징값 f_a, f_b를 생성한다. 두 번째 단계는 계층적 클러스터링 방법에 기반하여 첫 번째 단계에서 발견된 소군집들 중 가장 유사한 두 개의 군집을 새로운 유사도 측정 방법으로 찾아서 하나의 군집으로 병합한다. 이와 같은 소군집들의 반복적인 병합과정을 통해 원하는 군집을 발견해 낸다.

3.2 전처리 클러스터링 과정

초기의 소군집을 발견하는 단계로 자기조직화지도를 이용한다. 입력 데이터에 의해 자주 자극되는 모든 입력 데이터는 2차원 격자모양의 평면상에 승자군집과 이웃관계에 의해 이웃한 군집들과 완전 연결되어 있다. 이런 과정이 원시 데이터와 신뢰성 있는 관계를 갖게 된다. 또한 일반적으로 네트워크의 크기가 큰 경우에 잘 동작하기 때문에 대용량의 데이터를 적용하기에 보다 효율적이다.

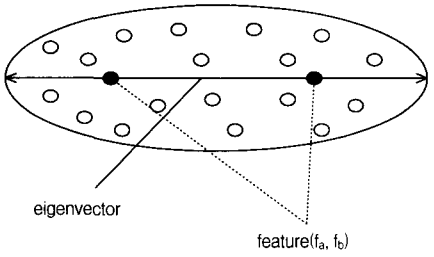
이웃관계를 정의하기 위해 사각형 격자모양을 사용하였다. 육각형 격자모양을 사용하면 가장 작은 이웃반경 범위에서 6개의 군집이 포함되나, 사각형 격자모양을 사용하면 4개의 군집이 형성된다. 승자군집과 이웃한 군집이 적을수록 계산량을 감소시킬수 있기때문에 보다 빠르게 수행된다. 승자군집은 이웃한 군집들과 연결강도를 갱신하고 반복적인 재할당이 이루어진다.

이웃 군집에 속하면서 연결강도를 갱신하면 식 (3.1)과 같이 조정되고, 이웃군집에 속하지 않으면 식 (3.2)와 같아진다.

$$\mathrm{mi}(t+1)=\begin{cases} mi(t)+\alpha\, hci(t)[x(t)-mi(t)] \\ \qquad \text{if } i \in N(t)\ (N(t): \text{이웃함수}) & (3.1) \\ mi(t) \quad otherwise & (3.2) \end{cases}$$

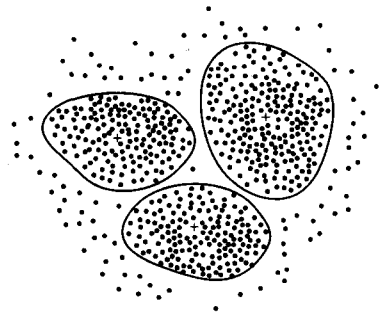
자기조직화지도는 단순히 연결강도 벡터가 입력패턴 벡터의 차이를 구한 다음, 그것의 일정한 비율을 원래의 연결강도 벡터에 더한다. 승자뉴런만이 연결강도 벡터를 갱신하는 것이 아니라, 이웃군집에 속한 군집들도 유사한 조정을 한다. 이웃군집에 포함되지 않은 군집은 원래의 연결강도 벡터를 갖게된다.

자기조직화지도는 각 노드에 해당하는 데이터들이 소군집을 형성하게 되고, 각각의 소군집은 두 개의 특징값(f_a, f_b)으로 표현된다. 두 개의 특징값은 소군집에 속하는 데이터 개체들을 고유값(eigenvalue)이 가장 큰 고유벡터(eigenvector) 위에 투영(projection)시켰을 때 나타나는 두 중심으로 (그림 5)와 같다.



(그림 5) 초기 소군집의 특징값

자기조직화지도를 수행한 후 얻은 초기 소군집의 특징값(f_a, f_b)은 2번째 단계의 후처리 클러스터링 과정에서 새로운 계층적 클러스터링 기법의 거리를 측정하는데 사용된다. 즉, 두 군집간의 새로운 유사도 측정에 이용된다. 따라서 군집의 요약 정보인 특징값만을 가지고 후처리 클러스터링을 수행하기 때문에 계산량을 줄이고 잡음제거 효과도 얻을 수 있다. (그림 6)은 3개의 소군집이 형성되었을 때 잡음처리된 결과이다. 자기조직화지도의 이웃 군집 안에 포함되지 않은 군집은 잡음으로 간주되는 것으로 해석할 수 있다.



(그림 6) 3개의 소군집이 형성되었을 때 잡음처리

3.3 후처리 클러스터링 과정

계층적 클러스터링 기법에 기반하여 초기 군집들 중 가장 유사한 두 개의 군집을 찾아서 하나의 군집으로 병합하는 과정을 반복적으로 수행하여 최종의 군집을 발견해내는 단계이다. 이 때 가장 유사한 두 개의 군집을 찾기 위한 새로운 유사도 측정방법을 고안했다.

계층적 클러스터링 방법의 가장 중요한 핵심은 군집 간의 유사도를 어떻게 측정할 것인가 하는 것으로, 양질의 군집을 발견하기 위한 중요한 문제이다. 여기서 말하는 '양질의 군집'이란 high intra-class similarity, low inter-class simi-

larity를 뜻한다. high intra-class similarity를 내부적 특징으로 규정하고 응집거리를 계산하고, low inter-class similarity를 외부적 특징으로 규정하고 인접거리로 계산한다.

PPC 기법은 계층적 클러스터링 기법에서 사용되는 기존의 유사도 측정방법인 최단 연결법, 최장 연결법, 평균 연결법, 중심연결법 등의 측정방법 이외의 측정방법으로 군집 간의 연관성과 특징을 고려하는 유사도 측정법을 고안했다.

인접거리의 분모를 2로 설정하는 것을 중위수(median) 연결법이라 하는데, 이 연결법은 군집의 전체적인 크기를 고려하지 못하고 단순히 중심만을 연결하는 단점이 있다. 따라서 본 연구에서는 군집의 전체적인 크기를 고려하기 위하여 분모를 $n_i n_j$로 측정하였다.

응집거리를 표현하는데 사용되는 연결거리$_{ij}$와 연결거리$_i$는 식 (3.3)과 식 (3.4)로 정의된다. 연결거리$_{ij}$의 식 (3.3)은 군집 i와 군집 j에 속하는 특징값들의 가중거리의 합이고, 연결거리$_i$의 식 (3.4)는 군집 i에 속하는 특징값들의 가중거리의 합이다. 각 특징값의 가중치 Wf_a와 Wf_b는 특징값이 대표하는 소군집의 개체수와 같다.

$$\text{연결거리}_{ij} = \sum_{a}^{n_i}\sum_{b}^{n_j} Wf_a \times Wf_b \times \| f_a - f_b \|^2 \quad (3.3)$$

$$\text{연결거리}_i = \frac{\sum_{a}^{n_i}\sum_{b}^{n_i} Wf_a \times Wf_b \times \| f_a - f_b \|^2}{2} \quad (3.4)$$

$$\text{연결거리}j = \frac{\sum_{a}^{n_i}\sum_{b}^{n_i} Wf_a \times Wf_b \times \| f_a - f_b \|^2}{2} \quad (3.5)$$

f_a, f_b : 특징값 벡터

n_i : 소군집 i에 속하는 특징값의 개수

Wf_a : 특징값 f_a가 대표하는 소군집의 개체수

Wf_b : 특징값 f_b가 대표하는 소군집의 개체수

군집의 내부적 특징인 응집거리 식 (3.6)과 군집간의 외부적 거리를 나타내는 인접거리 식 (3.7)을 이용하여 두 군집 간의 유사도를 식 (3.8)과 같이 나타낸다.

$$\text{응집거리}_{ij} = \frac{\text{연결거리}_{ij}}{\frac{\text{연결거리}_i + \text{연결거리}_j}{2}} \quad (3.6)$$

$$\text{인접거리}_{ij} = \frac{\sum_{a}^{n_i}\sum_{b}^{n_j} Wf_a \times Wf_b \times \| f_a - f_b \|^2}{n_i n_j} \quad (3.7)$$

$$\text{유사도}_{ij} = \frac{1}{\text{응집거리}_{ij} \times \text{인접거리}_{ij}^2} \quad (3.8)$$

응집거리$_{ij}$는 연결거리$_i$와 연결거리$_j$의 평균으로 정규화된 연결거리$_{ij}$로 정의되고 인접거리$_{ij}$는 연결거리$_{ij}$를 소군집의 대표값 개수인 n_i와 n_j로 정규화한 것으로 군집 i와 군집 j의 평균 연결거리이다. 각 군집의 연결거리를 나타내는 연결거리$_i$ 연결거리$_j$의 평균에 비해 군집을 후처리 클러스터링했을 때의 특징을 나타내는 연결거리$_{ij}$가 작다면 응집거리$_{ij}$가 작기 때문에 두 군집의 유사도는 높게 된다. 또한 인접거리$_{ij}$ (3.8)는 군집 i와 군집 j사이의 평균적 거리로 해석할 수 있다.

3.4 시간 복잡도 분석

데이터 마이닝 알고리즘의 실행시간은 예상 가능해야 하고, 어느 정도 받아들일 수 있는 시간이어야 한다. 공간복잡도와 시간복잡도가 사용되고 있으나, 프로그램의 입출력의 횟수나 크기와 관계없는 공간이 요구되는 공간복잡도로는 계산량을 감소시킬 수 없다. 따라서 시간복잡도를 통해서 분석한다. 사용되는 명령의 수와 그 명령이 실행되는 시간을 곱해서 합한 것을 O표 기법으로 시간복잡도(time complexity)로 산출한다[11].

〈표 1〉 시간복잡도 산출 방법

베이직 프로그램	시간복잡도	반복 횟수
for i = 1 to l s = s+1 next I	O (l)	l 회
for i = 1 to l for j = 1 to l s = s + i + j next j next i for i = 1 to l for j = 1 to l s = s * i * j next j next I	O (l^2)	l^2 회

일반적으로 지수(exponential) 혹은 다항식(polynomial) 시간복잡도를 갖는 알고리즘은 실용적이지 못하다.

데이터의 수를 n, 군집의 수를 k, 알고리즘이 수렴할 때까지의 반복횟수를 l, 샘플 데이터의 수를 s 그리고 소군집의 수를 p라고 할 때, 자기조직화지도, k-means, BIRCH, CURE, 계층적 클러스터링, PPC의 시간복잡도는 <표 2>와 같다.

〈표 2〉 클러스터링 기법들의 시간복잡도

클러스터링 기법	시간복잡도
자기조직화지도	O (nk l)
k-means	O (nk l)
BIRCH	O (nk)
CURE	O (logs × s^2)
계층적 클러스터링	O (n^2)
PPC	O (nk l + p^2)

<표 2>에 의하면 CURE가 가장 좋은 시간복잡도를 갖으며, 자기조직화지도, BIRCH, k-means의 시간복잡도는 데이

터집합의 크기에 선형비례하고, 계층적 클러스터링은 데이터 집합 크기의 제곱에 비례한다. 만약 데이터의 수 n을 100, 군집의 수 k를 2, 알고리즘이 수렴할 때까지의 반복횟수 l을 3, 그리고 소군집의 수 p를 2라 하고 실험하였을 경우 자기조직화지도의 시간복잡도는 600, 계층적 클러스터링 기법의 시간복잡도는 10000, PPC의 시간복잡도는 604가 된다.

PPC 기법의 시간 복잡도는 데이터의 크기에 선형비례 하는데, 자기조직화지도와 계층적 클러스터링 방법을 혼합했기 때문에 계산량이 자기조직화지도와 k-means 보다 많지만, p가 n에 비해 극히 작은 수이기 때문에 계층적 클러스터링에 비해 적다. 따라서 PPC 기법은 자기조직화지도와 k-means에 비해 계산량은 많지만, 데이터개체 수인 n에 선형비례하는 복잡도를 갖으면서 계층적 클러스터링보다 적은 계산량으로 군집을 발견 할 수 있다는 기대효과를 얻을 수 있다.

4. 실험 내용

4.1 실험 데이터

PPC 기법의 타당성을 검증하기 위해 몇 가지 벤치마크 데이터를 이용한다. 실세계 데이터에 대한 실험을 위해 <표 3>과 같이 UCI Machine Learning Repository[16]의 데이터 집합 중, 데이터 개체수가 700개를 넘는 Car evaluation, Covertype data, Diabetes, Letter Recognition, Nursery DB, Solar Flare를 사용하였다.

〈표 3〉 실험 데이터의 특징

실험 데이터	특징수	개체수	군집수
Car evaluation	4	1,728	6
Covertype data	54	581,012	8
Diabetes	9	768	2
Letter Recognition	17	20,000	1
Nursery DB	5	12,960	8
Solar Flare	13	1,389	3

4.2 클러스터링 성능 평가

4.2.1 학습율 α

자기조직화지도는 기계학습을 적용한 학습이기 때문에 설정값에 따라 학습의 결과에 영향을 미칠 수 있다. 본 실험에서는 임의로 설정해 주었다. α는 0과 1 사이의 값을 갖는 것이 일반적인 학습율(learn rate)인데, 너무 큰 값의 학습률은 빠르게 학습이 진행될 수도 있지만, 자칫하면 학습이 안 되는 상황도 발생할 수 있다. 반대로 너무 작은 학습률은 오차가 적어지는 형태로 학습이 이루어져서, 최종적으로 오차 최소점에 수렴은 하지만 각 학습 단계에서의 연결강도 변화량이 미세하여 전체 학습시간이 길어지는 단점을 들 수 있다.

4.2.2 응집도 Q

클러스터링 결과의 성능은 식 (4.1)과 같이 D_i의 평균인 Q로 측정하였다. D_i는 군집 i에 속하는 모든 데이터 개체들 (x_a, x_b) 사이의 평균거리로 군집이 얼마나 잘 응집되어 있는가를 나타내다. Q의 값이 적을수록 잘 응집된 군집이다.

$$Q = \sum_{i=1}^{k} \frac{D_i}{k} \tag{4.1}$$

$$D_i = \frac{\sum_{a=1}^{n_i} \sum_{b=1}^{n_i} \sqrt{(x_a - x_b)^2}}{n_i(n_i - 1)} \tag{4.2}$$

성능 평가를 위해 기존 클러스터링 알고리즘들 중에서 k-means와 4가지의 계층적 클러스터링 기법(최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법)을 비교대상으로 하였다. 이와 더불어 자기조직화지도를 수행한 후 특징값 f_a, f_b를 고려하지 않은 결과를 이용하여 4가지의 계층적 클러스터링 기법을 수행하는 심플 하이브리드 방식으로도 실험을 해보았다. 즉, PPC, k-means, 특징값 f_a, f_b를 고려한 4가지 기법, 특징값 f_a, f_b를 고려하지 않은 4가지 기법 이렇게 총 10가지 기법으로 실험을 하였다

이를 통해 기존의 군집간 유사도를 측정하는 방식인 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법에 비해 응집거리와 근접거리에 기반한 새로운 유사도 척도가 얼마나 효과적인지 비교한다.

4.3 실험 결과 및 분석

실세계의 6가지 UCI Machine Learning Repository의 데이터[12]를 이용하여, PPC 기법에 대하여 학습율 α를 0.3, 0.6, 0.9로 설정한 후 실험해 보았다<표 4>.

〈표 4〉 PPC 기법에 대한 실험 결과

데이터 \ 학습율	α = 0.3	α = 0.6	α = 0.9
Car evaluation	0.578	0.558	0.576
Covertype data	1.414	1.444	1.414
Diabetes	1.872	1.852	1.872
Letter Recognition	3.255	3.275	3.255
Nursery DB	2.209	2.249	2.205
Solar Flare	1.467	1.407	1.463

신경망의 학습시 연결 가중치의 변화율을 결정하는 0과 1 사이의 상수 즉, 학습율은 1에 가까울 수록 오버트레이닝 현상이 일어나고, 0에 가까울 수록 학습 시간이 길어지는 단점이 있다.

본 실험에서는 세 가지 학습율에 따른 별 다른 차이점을 발견하지 못했기 때문에, 치우치지 않는 값(α = 0.6)에 대하여 여러 가지 클러스터링 기법들과 비교하였다<표 6>.

연결강도 전처리 클러스터링 과정을 완전히 수행한 후, 새로운 유사도 측정법으로 응집도 Q를 구하기 때문에 응집도를 구하는 과정에서 α 값이 배제되기 때문이라는 판단했다. 따라서 신경망의 학습율은 양질의 군집을 얻는데 큰 영향을 미치지 못한다는 결과가 나왔다.

새로운 유사도 측정방법을 평가하기 위해 응집도 Q값을 실험하기 위하여 PPC, k-means, 일반적 계층적 클러스터링(4가지), 심플 하이브리드(4가지) 등 총 10가지 방법으로 나누어 실험하였다.

심플 하이브리드란 특징값 f_a, f_b를 고려하지 않고 자기조직화지도과 통계적인 계층적 클러스터링을 결합하는 방법으로 평가하는 방법이다.

아래의 <표 6>을 보면 PPC 기법으로 구한 Q값이 k-means와 최단연결법, 최장 연결법, 평균 연결법, 중심 연결법으로 구한 Q값 보다 적은 수치가 나왔으며, 심플하이브리드로 평가한 4가지 값과 비교해도 적은 수치가 나왔다.

〈표 6〉 각 클러스터링 방법과 비교

기법 / 데이터	PPC	k-means	계층적 클러스터링 기법			
			최단 연결	최장 연결	평균 연결	중심 연결
Car evaluation	**0.576**	0.659	0.668	0.683	0.649	0.660
Covertype	**0.205**	0.343	0.521	0.510	0.436	0.51
Diabetes	**1.872**	1.022	2.045	2.254	1.946	1.924
Letter Recognition	**1.576**	1.659	1.668	1.683	1.649	1.66
Nursery DB	**2.205**	2.343	2.521	2.510	2.436	2.510
Solar Flare	**2.255**	2.572	2.653	2.642	2.456	2.264
기법 / 데이터	심플하이브리드 (자기조직화지도 + 4가지 계층적 클러스터링 기법)			
	최단 연결	최장 연결	평균 연결	중심 연결
Car evaluation	0.592	0.692	0.622	0.596
Covertype	0.357	0.461	0.531	0.465
Diabetes	1.041	1.061	1.061	1.035
Letter Recognition	1.592	1.692	1.622	1.596
Nursery DB	2.357	2.461	2.531	2.465
Solar Flare	2.47	2.419	2.435	2.286

<표 6>에 따르면 응집도 Q의 값이 적을수록 잘 응축된 군집을 얻는 것으로 평가하였고, 대부분의 실험 데이터에 대해서 비슷한 성능을 보이는 letter recognition을 제외한 PPC 기법이 다른 방법들과 비교해 우수한 것으로 나타났다.

PPC 기법은 클러스터링을 수행하는 과정에서 소군집의 특징을 나타내는 요약 정보인 특징값을 이용하기 때문에 계산량을 효율적으로 감소시켜 확장성이 뛰어나며 잡음에도 강하다.

표본추출 기법을 이용한 CURE를 PPC와 비교해 보면 잘 분포된 몇 개의 데이터 개체를 선정하는 pre clustering을 수행하는 단계는 유사하지만 최단 연결법을 이용해 특징값 1개를 사용해서 유사도를 측정한 후 최종군집을 판별하는 방법의 단점을 보완하여 PPC는 특징값을 2개로 선정하여 보다 연관성있는 특징을 갖는 군집을 얻을 수 있다. 원시 데이터를 직접 다루지 않은 BIRCH에 비해서 신뢰할 수 있는 군집을 얻는 것이 특징이다.

크게 세 가지로 분류하면(통계적인 클러스터링방법, k-means) < 심플 하이브리드 < PPC 이런 순으로 좋은 응집도를 보였다.

최단 연결법과 최장 연결법은 군집간의 유사도를 측정하기 위해 특정 데이터 개체 하나에만 의존하기 때문에 잡음에 민감하고, 평균 연결법과 중심 연결법은 하나 이상의 데이터 개체를 기반으로 중심값을 갖기 때문에 잡음에 대해덜 민감하였다. PPC 기법은 자기조직화지도로 전처리 하는 과정에서 데이터를 요약, 압축했고 두 개의 데이터 개체를 기반으로 인접거리, 연결거리를 기반으로 최종군집을 얻었기 때문에 가장 좋은 응집도를 보였다.

5. 결 론

본 논문을 통해 대용량의 데이터에 적용할 수 있도록 효율적으로 계산량을 줄일 수 있는 하이브리드 클러스터링 기법인 PPC 기법를 제안하였다. PPC 기법은 인공지능적 방법인 자기조직화지도와 통계적 기법인 계층적 클러스터링 방법을 접목한 방법으로 두 단계를 거쳐 클러스터링을 수행한다. 전처리 클러스터링 과정에서 자기조직화지도를 통해 데이터를 요약하고, 후처리 클러스터링 과정에서 클러스터링 적용시 군집의 내부적 특징을 나타내는 응집거리와 군집간의 외부적 거리를 나타내는 인접거리에 기반한 새로운 유사도 측정방법을 고안하여 적용하였다. 또한 클러스터링을 수행하는 과정에서 소군집의 특징을 나타내는 요약 정보인 특징값을 이용하기 때문에 계산속도가 향상되어 대용량의 데이터를 대상으로 하는 데이터 마이닝에 응용할 수 있다.

PPC 기법은 성능을 평가하기 위해 기존 방법들과의 비교실험을 통해 그 유효성을 검증해 보았다. 또한 새로운 유사도 척도를 검증하기 위해 자기조직화지도와 계층적 클러스터링 기법을 특징값 f_a, f_b를 고려하지 않고 단순히 결합한 심플 하이브리드 방법과의 비교에서도 PPC의 클러스터링에서 실험한 응집도값이 상대적으로 적은 수치를 보임으로서 양질을 군집을 발견할 수 있었다.

따라서 대용량의 데이터를 대상으로 효율적으로 비교적 양질의 군집을 발견할 수 있는 클러스터링 방법이라는 사실을 확인할 수 있었다.

향후 클러스터링 수행결과를 사용자에게 어떤 형태로 서비스할 것인가(visualization)는 클러스터링 과정 못지 않게 매우 중요하다. 따라서 클러스터의 내용을 보다 쉽게 이해할 수 있는 레이블을 제공하는 방법에 대한 연구가 필요할 것이다.

참 고 문 헌

[1] 장미희, 이혜영, "고차원 데이터에서 2차원 프로젝션을 이용한 클러스터링", 정보과학회 추계학술대회, 2001.

[2] Tian Zhang, Raghu Ramakrishnan, and Miron, "Birch : an efficient data clustering method for very large database," the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June, 1996.

[3] Richard. Pyle, Duda and Peter E. Hart, "Pattern Classfication and Scene Analysis," A Wiley-Interscience Publication, NewYork, 1973.

[4] Berry, Linoff, "Data Mining Techniques for Market, Sales, and Customer Support," Jone Wiley & Sons, 1997.

[5] http://www.cis.hut.fi/sami/mipapers/bioinformatics.shtml.

[6] Kohonen, Teuvo, "The self-organizing map," Neurocomputing, Vol.21, pp.1-6, 1998.

[7] 김대수, 신경망 이론과 응용(Ⅰ), 하이테크정보, 1992.

[8] K. C. Gowda, E. Diday, "Symbolic Clustering Using a Similarity Measure," IEEE Trans on System, Man, and Cybernetics, Vol.22, No.2, p.341, 1992.

[9] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim, "CURE : An Efficient Clustering Algorithm for Large Databases," the ACM SIGMOD Conference on Management of Data, Seattle, Washington, June, 1998.

[10] Tian Zhang, Raghu Ramakrishnan, and Miron, "Birch : A New Data Clustering Algotithm and Applications," Data Mining and Knowledge Discovery, 1, pp.141-182, 1997.

[11] http://user.chollian.net/~leesc12/lecture/start/tcomplex.htm.

[12] http://www.ics.uci.edu/~mlearn/MLRepository.html.

김 만 선

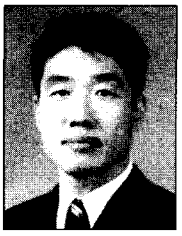
e-mail : mansun@kongju.ac.kr

2000년 홍익대학교 전자전기컴퓨터공학부 (공학사)

2002년 공주대학교 대학원 전자계산학과 (이학석사)

2002년~현재 공주대학교 대학원 컴퓨터공학과 박사과정

관심분야 : 데이터마이닝, 신경망, 에이전트 시스템



이 상 용

e-mail : sylee@kongju.ac.kr

1984년 중앙대학교 전자계산학과(공학사)

1988년 일본동경공업대학 총합이공학연구과 (공학석사)

1988년~1989년 일본 NEC 중앙연구소 연구원

1993년 중앙대학교 일반대학원 전자계산학과(공학박사)

1993년~현재 공주대학교 정보통신공학부 교수

1996년~1997년 University of Central Florida 방문교수

관심분야 : 인공지능, 기계학습, 에이전트, 바이오인포매틱스