데이터 마이닝을 위한 계층적 대표값 군집화 기법

안병주, 김은주, 이일병
연세대학교 컴퓨터과학과
{joo, outframe, yblee}@csai.yonsei.ac.kr

A Hierarchical Representatives Clustering Technique for Data Mining

Byung-Joo An, Eun-Ju Kim, Yill-Byung Lee
Dept. of Computer Science, Yonsei University

요 약

군집화는 데이터 집합을 유사한 데이터 개체들의 군집들로 분할하여 데이터 속에 존재하는 의미 있는 정보를 얻는 과정이다. 대부분의 군집화 기법들은 비교적 적은 양의 데이터를 대상으로 한 것이고 다차원 대용량의 데이터 처리에 관한 문제는 다루지 않고 있어서 데이터 마이닝을 위한 군집화 기법으로는 부적절하다. 따라서 본 논문을 통해 대용량의 데이터에 적용할 수 있는 새로운 군집화 알고리즘인 계층적 대표값 군집화(HRC) 기법을 제안한다. HRC는 자기조직화지도와 계층적 군집화 기법을 접목한 하이브리드 방법으로 두 단계에 거쳐 군집화를 수행한다. 첫 번째 단계에서 자기조직화지도를 통해 데이터를 요약하고, 두 번째 단계에서 요약된 대표값 정보만을 가지고 계층적인 군집화를 수행한다. 또한, 두 번째 단계의 계층적 군집화 적용시 양질의 군집을 발견하기 위해 군집간의 유사도를 측정하는 새로운 척도를 고안하였다. 그리고 실험을 통해 HRC와 기존 군집화 알고리즘이 발견한 군집의 질을 비교하여 성능을 평가했다.

1. 서론

데이터 군집화는 데이터 분류와 이미지 처리 같은 많은 실용적 문제영역에 적용할 수 있는 탐색적 데이터 분석의 중요한 기법 중 하나이다.

군집화는 입력 데이터집합을 유사한 관찰치들의 군집들로 구분하여 데이터집합 속에 존재하는 의미 있는 정보를 얻는 과정이다[1][2]. 즉, 군집내의 유사성은 최대화하고 군집들간의 유사성은 최소화 시키도록 데이터 집합을 분할하는 것이다[3]. 이러한 군집발견 과정은 우리에게 군집의 데이터 분포가 갖고 있는 특징을 설명하며 다른 분석 기법을 위한 토대를 마련 해주는 역할을 수행 할 수 있다[4]. 또한 군집화 기법을 이용하면, 기업의 고객을 구매패턴에 근거해서 분류하거나, 웹 문서의 범주별 분류에 이용하거나, 유사한 기능을 하는 유전자와 단백질을 분류하는 데 이용하는 등, 다양한 응용 분야에 적용이 가능하다. 최근에는 대용량의 데이터베이스로부터 유용한 정보를 발견하고 데이터간에 존재하는 연관성을 탐색하고 분석하는 데이터 마이닝[5]에 대한 많은 연구들이 진행되고있다. 데이터 마이닝의 출현으로 인해 원시데이터에 대한 접근회수를 줄이고 알고리즘이 다루어야 할 데이터구조의 크기를 줄이는 군집화 기법에 관한 연구들이 활발하다.

본 논문은 다차원 대용량 데이터에 적용하여 효율적으로 양질의 군집을 발견할 수 있도록 하는 새로운 군집화 알고리즘인 계층적 대표값 군집화(HRC) 기법을 제안한다. 대부분의 기존 군집화 알고리즘은 소량의 데이터를 대상으로 비슷한 크기를 갖는 구형의 군집들로(hyperspherical clusters) 데이터를 나누는 경향이 있다[6][7]. 제안하는 방법은 이러한 한계를 극복하고 대용량의 데이터에 존재하는 다양한 모양의 군집을 발견할 수 있는 새로운 하이브리드 군집화 알고리즘이다. HRC는 인공지능적 군집화 방법과 통계적 군집화 방법을 접목한 방법으로 두 단계를 거쳐 군집화를 수행한다. 첫 번째 단계는 초기의 소군집을 발견하기 위한 단계로 대용량의 데이터에 효율적으로 적용할 수 있는 [7] 인공지능적 군집화 방법인 자기조직화지도(Self-Organizing Map) 를 이용하고, 두 번째 단계는 통계적 군집화 방법으로 다양한 형태의 군집을 발견할 수 있는 계층적 군집화를 이용하여 첫 번째 단계에서 발견된 초기소군집들을 반복적으로 병합하여 원하는 군집을 얻는다.

2. 관련연구

데이터 군집화는 기계학습, 통계학, 데이터베이스 분야에서 여러가지 기법들에 대한 연구가 활발히 진행되어 왔다.

군집화 알고리즘은 두 가지 부류로 분류 할 수 있는데, 분할적 군집화(Partitional Clustering) 와 계층적 군집화(Hierarchical clustering) 로 나눌 수 있다[3][8]. 분할적 군집화[3][8]는 주어진 목적함수를 최적화 하도록 데이터 집합을 k 개의 군집으로 나누는 것으로 K-MEANS, K-MEDOID, PAM 등이 이에 속한다. 임의의 초기 분할로부터 시작하여, 데이터 개체에 대한 소속 군집의 재할당 과정과 목적함수의 평가를 반복적으로 수행하여 목적함수를 최적화 한다. 계층적 군집화[3][9]는 가장 유사한 두 개체들을 선택하여 병합해 가는 병합적 계층군집방법과 가장 먼 개체들을 선택하여 나누어 나가는 분할적 계층군집방법이 있다. 두 군집의 유사도를 측정하는 기준에 따라 최단연결법, 최장연결법, 중심연결법, 평균연결법등으로 나뉜다[3].

최근에는 대용량의 데이터 처리에 중심을 두고 요약된 군집표현을 이용하거나, 표본추출 기법을 이용하거나, 혹은 특별한 자료구조를 이용하는 새로운 군집화 알고리즘들이 발표되었다. BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)[1][2] 는 원시 데이터를 직접 다루지 않고 군집에 속한 데이터

개체의 수, 중심값, 군집의 반지름으로 구성된 군집요약정보인 군집특징(cluster feature)을 이용한다. 새로운 데이터가 추가되면 새로운 군집특징은 이전의 군집특징으로부터 계산할 수 있는 점진적(incremental) 방법이다. Chameleon[10] 은 계층적 군집화 기법에 기반하여 두 단계로 구성되어 있으며 다양한 특징을 띠는 군집간 유사도를 측정할 수 있는 새로운 동적 모델을 제시했다. 이 방법은 다양한 모양, 밀도, 크기를 갖는 자연스런 군집을 발견할 수 있어서 공간데이타마이닝(Spatial Data Mining) 에 응용할 수 있다.

3. 계층적 대표값 군집화(HRC) 기법

본 논문에서는 대용량의 데이터에 적용할 수 있는 하이브리드 군집화 기법인 HRC 를 제안한다. HRC는 군집화를 수행하는 과정에서 소군집의 특징을 나타내는 요약 정보인 대표값을 이용하기 때문에 계산량을 효율적으로 감소시켜 확장성(Scalability) 이 뛰어나며 잡음에도 강하다.

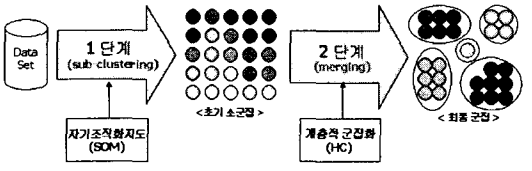

그림 1. 계층적 대표값 군집화(HRC)의 개요

이 기법은 인공지능적 군집화 방법과 통계적 군집화 방법을 접목한 방법으로 두 단계를 거쳐 군집화를 수행한다. 첫 번째 단계는 초기의 소군집을 발견하기 위한 단계로 자기조직화지도를 이용하고, 두 번째 단계는 계층적 군집화 방법에 기반 하여 첫 번째 단계에서 발견된 초기군집들 중 가장 유사한 두 개의 군집을 새로운 유사도 측정 방법으로 찾아서 하나의 군집으로 병합한다. 이와 같은 소군집들의 반복적인 병합과정을 통해 원하는 군집을 발견해 낸다.

HRC 는 자기조직화지도와 계층적 군집화기법을 혼합하여 두 방법의 장점을 이용한다. 자기조직화지도로 소군집화를 수행하여 대용량 데이터를 대표값으로 표현되는 소군집들로 요약하고, 이 대표값 정보를 이용해 다양한 특징을 갖는 군집을 발견할 수 있는 계층적 군집화를 수행한다. 자기조직화지도는 데이터의 수에 선형 비례하는 계산 복잡도를 갖기 때문에 대용량의 데이터에 적용 가능하다. 특히 온라인 학습방법을 이용할 경우 많은 양의 메모리 사용 없이 연결강도 벡터들과 현재 학습 벡터만을 위한 공간만 있으면 된다. 계층적 군집화 기법은 군집간의 유사도 측정 방법에 따라 다양한 특징을 갖는 군집을 발견 할 수 있다. 최단 연결법은 이상치와 잡음에 민감해서 체인효과를 통해 기다란 모양의 군집 발견에 효과적이며, 최장연결법은 비슷한 크기로 잘 뭉쳐있어서 경계가 뚜렷한 군집들을 포함한 데이터에 효과적이다. 평균 연결법과 중심 연결법은 최단연결법과 중심연결법을 절충한 방법이다 [3][7].

3.1. 1 단계

초기의 소군집을 발견하는 단계로 자기 조직화 지도를 이용한다. 자기조직화지도의 각 노드에 해당하는 데이터들이 소군집을 형성하게 되고 각각의 소군집은 두 개의 대표값으로 표현된다. 두 개의 대표값은 소군집에 속하는 데이터개체들을 아이겐밸류가 가장 큰 아이겐벡터 위에 투영(projection) 시켰을 때 나타나는 두 중심으로 그림 3 과 같다. 초기 소군집의 대표값은 2 단계의 병합과정에서 두 군집 간의 유사도 측정에 이용된다. 따라서 군집의 요약 정보인 대표값만을 가지고 병합과정을 수행하기 때문에 계산량을 줄이고 잡음제거 효과도 얻는다.

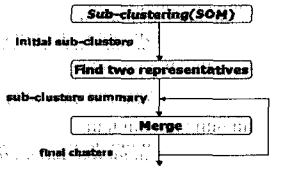

그림 2. HRC 알고리즘

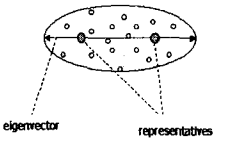

그림 3. 초기소군집의 대표값

3.2. 2 단계

계층적 군집화 기법에 기반 하여 첫 번째 단계에서 발견된 초기군집들 중 가장 유사한 두 개의 군집을 찾아서 하나의 군집으로 병합하는 과정을 반복적으로 수행하여 최종의 군집을 발견해내는 단계이다. 이 때 가장 유사한 두개의 군집을 찾기 위한 새로운 유사도 측정방법을 고안했다.

계층적 군집화 기법의 가장 중요한 핵심은 군집간의 유사도를 어떻게 측정할 것인가 하는 문제이다[10]. HRC에서는 계층적 군집화 기법에서 사용되는 유사도 측정방법 이외에 새로운 측정방법으로 군집간의 연관성과 동적 특징을 고려하는 유사도 측정법을 고안했다. 이 방법은 군집의 내부적 특징인 응집도와 군집간의 외부적 거리를 나타내는 근접도를 이용하여 두 군집간의 유사도를 나타내는 데 그 수식은 다음과 같다.

$$\text{유사도}_{ij} = \text{응집도}_{ij} \times \text{근접도}_{ij}^{2}$$

$$\text{응집도}_{ij} = \frac{\text{연결강도}_{ij}}{\frac{\text{연결강도}_i + \text{연결강도}_j}{2}}, \quad \text{근접도}_{ij} = \frac{\sum_{a}^{n_i}\sum_{b}^{n_j} \|r_a\| \times \|r_b\| \times \|r_a - r_b\|^2}{n_i \times n_j}$$

$$\text{연결강도}_i = \sum_{a}^{n_i}\sum_{b}^{n_i} \|r_a\| \times \|r_b\| \times \|r_a - r_b\|^2, \quad \text{연결강도} = \frac{\sum_{a}^{n_i}\sum_{b}^{n_j} \|r_a\| \times \|r_b\| \times \|r_a - r_b\|^2}{2}$$

$\|r_a\|$: 대표값 r_a가 대표하는 데이터 개체수, n_i: 소군집 i에 속하는 대표값의 개수

응집도$_{ij}$는 군집i와 군집j의 연결강도의 평균으로 정규화된 연결강도$_{ij}$로 정의되고, 근접도$_{ij}$는 연결강도$_{ij}$를 n_i와 n_j로 정규화한 것으로 군집i와 군집j의 평균연결강도이다. 각 군집의 연결강도를 나타내는 연결강도$_i$와 연결강도$_j$의 평균에 비해 군집을 병합했을 때의 특징을 나타내는 연결강도$_{ij}$가 높다면 응집도$_{ij}$가 높기 때문에 두 군집의 유사도는 높게 된다. 또한 근접도$_{ij}$는 군집i와 군집j사이의 평균적 거리로 해석할 수 있다.

3.3. 복잡도 분석

군집화 기법	시간복잡도	공간복잡도
자기조직화지도	$O(nkl)$	$O(k+n)$
K-MEANS	$O(nkl)$	$O(k+n)$
계층적 군집화	$O(n^2)$	$O(n^2)$
HRC	$O(nkl + m^2)$	$O(k+n+m^2)$

표 1. 군집화기법들의 복잡도

데이터의 수를 n, 군집의 수를 k, 알고리즘이 수렴할 때 까지의 반복회수를 l, 그리고 소군집의 수를 m이라 할 때 자기조직화지도, K-MEANS, 계층적 군집화, HRC의 복잡도는 표1과 같다.

자기조직화지도와 K-MEANS 의 시간복잡도와 공간복잡도는 데이터집합의 크기에 선형비례 하고 계층적 군집화는 데이터집합 크기의 제곱에 비례 한다. HRC의 시간복잡도와 공간복잡도는 데이터집합의 크기에 선형비례 하는데, 자기조직화지도와 계층적 군집화 방법을 혼합했기 때문에 계산량이 자기조직화지도 보다 많지만 m 이 n 에 비해 적은 수이기 때문에 계층적 군집화에 비해 적다.

4. 실험

실험을 위해 3 개의 2 차원 공간데이터와 7개의 UCI Machine Learning Repository[11] 데이터집합 (australian, diabetes, heart, iris, soybean, wine, zoo)을 사용하였다.

그림 4. 2차원 공간 데이터

군집화 성능은 다음 식의 Q로 측정하였는데, D_i는 군집 i에 속하는 모든 데이터개체들 간의 평균거리이고 Q 는 D_i의 평균이다.

$$Q = \sum_{i=1}^{k} \frac{D_i}{k}, \quad D_i = \frac{\sum_{a=1}^{n_i} \sum_{b=1}^{n_i} \sqrt{(x_a - x_b)^2}}{n_i(n_i - 1)}$$

성능 평가를 위해 기존 군집화 알고리즘들 중에서 K-MEANS와 계층적 군집화 기법(최단연결법, 최장연결법, 평균연결법)을 비교대상으로 하였다.

4.1 실험결과

	HRC	K-MEANS	계층적 군집화		
			최단	최장	평균
Data set 1	**4.369**	5.609	8.063	6.012	5.673
Data set 2	**5.274**	6.217	9.700	8.042	5.854
Data set 3	**1.960**	2.708	2.495	2.620	2.577

표 2. 공간데이터집합에 대한 군집화 기법들의 성능(Q) 비교

그림 4 의 2 차원 공간데이터에 대한 군집화 결과는 K-MEANS와 평균연결법은 큰 군집을 잘게 쪼개려는 경향을 보이고, 최단연결법은 이상치에 민감해서 올바른 군집을 발견하지 못 했으나, HRC 가 가장 양질의 군집을 발견했으며 군집의 질은 표 2와 같다.

	HRC	K-MEANS	계층적 군집화		
			최단	최장	평균
Australian	1.535	**1.322**	1.538	1.334	1.538
Diabetes	**0.521**	0.540	0.604	0.625	0.604
Heart	**1.284**	1.475	1.673	1.624	1.378
Iris	**0.786**	0.918	0.844	0.890	0.917
Soybean	**2.933**	3.399	3.095	3.095	3.095
Wine	**0.663**	0.715	0.988	0.719	0.988
Zoo	**1.090**	1.185	1.354	1.361	1.306

표 3. UCI 데이터집합에 대한 군집화 기법들의 성능(Q) 비교

실세계 데이터인 7 개의 UCI 데이터집합에 대한 결과는 표 3 에서와 같이 HRC 의 군집화 성능이 가장 좋다.

5. 결론 및 논의

본 논문을 통해 대용량의 데이터에 적용할 수 있도록 효율적으로 계산량을 줄일 수 있는 하이브리드 군집화 기법인 HRC 를 제시하였다. HRC 는 자기조직화지도와 계층적 군집화 기법을 접목한 하이브리드 방법으로 두 단계를 거쳐 군집화를 수행한다. 또한 군집화를 수행하는 과정에서 소군집의 특징을 나타내는 요약 정보인 대표값을 이용하기 때문에 계산속도가 향상되어 대용량의 데이터를 대상으로 하는 데이터마이닝에 응용할 수 있다. 그리고 실험 결과를 통해 HRC 가 비교적 양질의 군집을 발견하는 기대효과를 얻었다.

첫 번째 단계의 소군집화 결과에 따라 최종 군집의 결과가 좌우되는 현상을 보였는데, 향후 이에 대한 보완을 위한 연구를 진행할 것이다.

참고문헌

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

[2] Tian Zhang, Raghu Ramakrishnan, and Miron, Birch: A New Data Clustering Algorithm and Its Applications." Data Minning and Knowledge Discovery, 1, 141-182, 1997.

[3] Richard O. Duda and Peter E. Hard, Pattern Classification and Scene Analysis, A Wiley-Interscience Publication, New York, 1973.

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

[5] Fayyad, Piatetsky-Shapiro, Smyth, "Advances in knowledge discovery and data mining ", 1996.

[6] 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.

[7] A. K. Jain, M. N. Murty and P. J. Flynn, " Data clustering: a review ", the ACM Comput. Surv. 31, 3, Pages 264 - 323, Sep. 1999.

[8] Kaufman, Leonard and Rousseuw, Peter J., Finding Groups in Data - An Introduction to Cluster Analysis, Wiley Series in Probability and Mathematical Statistics, 1990.

[9] Murtagh, F., "A Survey of Rescent Advances in Hierarchical Clustering Algorithms ", The Computer Journal, 1983.

[10] George Karypis, Eui-Hong Han, Vipin Kumar, "Chameleon: Hierarchical Clustering Using Dynamic Modeling", IEEE Computer, Vol 32, No 9, August 1999.

[11] UCI Maching Learning Repository, http://www.ics ,uci.edu/~mlearn.