데이터 마이닝에서의 군집분석 알고리즘 비교 연구

이영섭[1] · 안미영[2]

요약

데이터베이스에 내재된 패턴이나 관계를 묘사한 것만으로도 의사결정에 필요한 정보를 제공할 수 있는데 이 데이터들의 변수들을 비슷한 특징을 가지는 소그룹으로 나누어 패턴을 찾는 것을 군집분석이라 한다. 이러한 군집 분석에는 분리군집방법과 계층적군집방법이 있는데, 재할당이 가능한 분리군집방법의 여러 알고리즘에 대해 비교해보자. 분리군집알고리즘에는 중심을 평균으로 하는 k-평균 알고리즘과, 중심을 메도이드로하는 PAM, CLARA, CLARANS 알고리즘이 있다. 이러한 알고리즘에 대한 이론과, 장단점을 설명하고, 분산과 중심들간의 평균 거리로 비교해 본다.

Keywords : 군집분석, k-평균 알고리즘, 메도이드, PAM 알고리즘, CLARA 알고리즘, CLARANS 알고리즘

1. 서론

컴퓨터와 인터넷이 급격히 발전함에 따라 기업이나 모든 조직들은 데이터를 정보의 인프라로 인식하고 데이터베이스를 구축하게 되었다. 데이터베이스는 간단한 질의 도구로 찾아내기 어려울 정도로 방대한 규모가 되었고 그 특징을 파악함으로써 이제까지 알 수 없었던 새로운 지식을 얻을 수 있게 되었다. 이렇게 새로운 지식을 찾아내는 전반적인 과정을 KDD(Knowledge Discovery in Database)과정이라고 한다. KDD 과정은 데이터로부터 유용한 지식을 발견하는 전체적인 프로세스를 의미하지만 데이터 마이닝은 원시데이터에서 패턴이나 유용한 정보, 지식을 추출하는 특별한 알고리즘을 사용하는 KDD 과정중 한 단계이다.

데이터베이스에 내재된 패턴이나 관계를 묘사한 것만으로도 의사결정에 필요한 정보를 제공할 수 있는데 이 데이터들의 변수들을 비슷한 특징을 가지는 소그룹으로 나누어 패턴을 찾는 것을 군집분석이라 한다. 군집분석은 데이터의 특성을 알지 못할 때, 유사한 항목들의 그룹을 찾아내는데 많이 사용된다. 따라서 데이터 마이닝을 수행하기 위해 데이터의 특징을 알아내기 위한 초기 작업에는 유용하지만, 이는 복잡한 데이터를 묘사하는 기법일 뿐이고, 유사성을 갖는 군집내의 관찰치들을 통한 변수들 사이의 규칙이나 패턴을 찾는 또 다른 데이터 마이닝 기법을 사용하여 유용한 결과를 얻기도 한다.

일반적으로 분리군집방법에서 많이 알려져 있는 알고리즘에는 k-평균 알고리즘이 있다. 하지만 k-평균 알고리즘은 중심 계산을 평균으로 하기 때문에 이상치에 민감한 단점이 있다. 이러한 단점을 보완할 수 있는 알고리즘으로 PAM(L. Kaufman, P.J.

1). 이영섭, 동국대학교 통계학과 조교수, 서울시 중구 필동 동국대학교
2). 안미영, 동국대학교 통계학과 석사과정, 서울시 중구 필동 동국대학교

Rousseeuw. 1978), CLARA(L. Kaufman, P.J. Rousseeuw. 1986), CLARANS(R. Ng, J. Han. 1994) 알고리즘이 있다.

기존의 논문들에서는 위에 언급한 여러 가지 알고리즘들의 복잡도(complexity)를 계산하고, 대용량 데이터들에 적용시켰을 때의 수행 속도(runtime)를 시간(second)으로 비교해 놓았다(R. Ng and J. Han. 1994). 하지만 데이터들에 여러 알고리즘을 적용시켰을 때, 통계학적인 시각으로 데이터들이 얼마나 효율적으로 분리되었는가에 대한 연구는 없었다. 본 논문에서는 분리군집방법의 여러 알고리즘의 절차와 특징에 대해 설명하고, 군집분석의 다양한 알고리즘들을 실제 데이터에 적용시켜 각각의 알고리즘의 효율성에 대해 알아보기 위해 실제 데이터를 이용하여 적용해 보고자 한다.

2. 본론

2.1 k-평균 알고리즘

k-평균 알고리즘은 Cox(1957)와 Fisher(1958)에 의해 제안되었고, Hartigar(1975)과 MacQueer(1967)에 의해 개발된 후 계속적으로 연구되고 있다. 현재 분리군집방법 중에서 가장 보편적으로 많이 쓰이는 알고리즘의 하나로, 군집의 유사성은 군집의 중심점인 평균과 객체들간의 거리로 측정한다.

k-평균 알고리즘은 중심점과 객체의 거리를 계산하여 가장 가까운 중심점에 객체를 할당하는 방법이다. k-평균 알고리즘을 사용하기 위해서는 군집의 수인 k 가 미리 결정되어 있어야 한다. k-평균 알고리즘은 〈표 1〉 의 4가지 절차를 거친다.

<표 1> k-평균 알고리즘

【단계1】 자료를 k개의 초기 군집으로 나눈다.
【단계2】 k개로 나누어진 군집의 중심을 평균을 이용하여 구한다.
【단계3】 각 객체와 중심들 사이의 거리를 거리 구하는 식을 이용하여 계산한다. 그리고 객체가 현재 속해있는 군집 중심에 가까우면 현재 군집에 포함되고, 다른 군집의 중심에 가까우면 그 군집으로 재분류한다.
【단계4】 다시 할당되는 개체가 없을 때까지 [단계2]와 [단계3]을 반복한다.

k-평균 알고리즘은 중심점을 평균으로 계산하기 때문에, 평균을 구할 수 있는 데이터에서만 사용할 수 있다. 예를 들어, 범주형 자료 같은 경우에는 k-평균 알고리즘을 적용시킬 수 없고, 잡음과 이상치에 대하여 민감한 결과를 보인다. 또한 초기에 선택한 k 값이나, 비유사성(dissimilarity)을 계산하는 방법에 따라 다른 결과를 초래하는 단점이 있다.

2.2 메도이드에 의한 알고리즘

2.2.1 PAM 알고리즘

PAM(Partitioning Around Medoids) 알고리즘은 Kaufman과 Rousseeuw에 의해 1978년에 제안된 알고리즘으로 k-평균 알고리즘에서는 군집의 중심으로 객체들 간의 평균을 계산한 가상의 점을 사용하는 것과 달리, 군집의 중심으로 실제 객체인 메도이드(medoid)를 중심으로 사용한다. 여기서는 메도이드란 군집 내에서 객체들간의 평균 비유사성이 가장 작은 객체를 말한다.

이상적인 메도이드를 찾기 위해 반복을 통하여 메도이드들을 변화시켜 나가는데, 이렇게 메도이드들을 변화시킬 때마다 객체들이 가까운 메도이드들을 중심으로 객체를 형성하기 위해 움직이게 된다. 이때 객체들이 변화된 메도이드로 인하여 재분류되었을 때 이동한 객체와 본래의 메도이드, 변화된 메도이드와 거리의 차를 비용이라 한다. PAM 알고리즘에서는 객체들이 이동하면서 발생한 비용을 모두 더한 총 비용을 이용하여 이상적인 메도이드를 찾아나간다. PAM 알고리즘의 기본적인 절차는 〈표 2〉 와 같고 이 알고리즘도 k-평균 알고리즘처럼 k의 개수를 미리 알고 있어야 한다. PAM 알고리즘은 객체들과 군집의 메도이드들 간의 평균 비유사성으로 군집의 효율성을 측정한다.

<표 2> PAM 알고리즘

【단계1】 임의로 K개의 초기치를 선택한다.
【단계2】 K개의 객체를 데이터에서 랜덤하게 추출하여, 현재 메도이드인 O_i로 설정한다.
【단계3】 현재 메도이드인 O_i에 가까이 있는 객체들로 군집을 분류한 다.
【단계4】 분류된 K개의 군집 내에서 비유사성을 계산하여 가장 좋은 메도이드 O_h를 찾는다.
【단계5】 현재 O_i와 새로운 메도이드 O_h사이의 총비용이 최소값을 가지도록 하는 메도이드를 찾는다. 만약 총 비용의 최소값이 음수면, 현재 메도이드 O_i를 새로운 메도이드 O_h로 바꾸고, [단계3]으로 되돌아 간다.
【단계6】 음수가 아닐 경우가 발생할 때까지 [단계3], [단계4], [단계5]를 반복하여 평균 비유사성이 가장 낮은 군집을 찾는다.

PAM 알고리즘은 군집의 질을 평가할 때, 모든 경우의 데이터에 대해 계산을 하므로, 데이터의 크기가 커질수록 계산량이 많아 컴퓨터의 수행 속도가 느려지는 단점이 있다. (J. Han and M. Kamber. 2000)

2.2.2 CLARA 알고리즘

CLARA(Clustering LARge Applications) 알고리즘은 Kaufman과 Rousseeuw에 의해 1986년 제안된 알고리즘으로 대용량 데이터를 효율적으로 다루기 위한 방법이다. 군집을 나눌 때, 전체 데이터 대신에 데이터의 표본을 랜덤 추출하고, 표본에 PAM 알고리즘을 적용시켜 표본에서 k개의 최적의 메도이드를 구한다. 표본 추출과 메도이드를 찾는 과정을 반복하여 최적의 메도이드를 찾는다. 표본을 랜덤 추출하고, 추출하는 과정을 반복적으로 한다면, 전체 데이터를 이용하여 구한 메도이드와 거의 유사한 메도이드가 구해진다. 하지만 군집의 정확성을 측정할 경우에는 표본에서 구하여진 메도이드들과 표본의 객체들간의 평균 비유사성을 구하는 것이 아니라 표본에서 구하여진 메도이드들과 전체 데이터의 모든 객체들의 평균 비유사성을 계산한다. CLARA 알고리즘의 기본적인 절차는 〈표 3〉 과 같고 이 알고리즘도 k-평균 알고리즘처럼 k의 개수를 미리 알고 있어야 한다.

<표 3> CLARA 알고리즘

【단계1】 전체 데이터에서 랜덤하게 표본을 추출하여 그 표본에 PAM 알고리즘을 적용시켜 k개의 메도이드를 찾는다.
【단계2】 단계 1에서 구한 k개의 메도이드를 중심으로 전체 데이터를 이용해 메도이드에 가까운 객체들로 군집을 형성한다.
【단계3】 전체 데이터로 형성된 군집을 이용해 평균 비유사성을 계산한다. 계산된 값이 현재 값보다 작다면, 계산된 값을 현재 값으로 바꾼다.
【단계4】 메도이드가 수렴할 때까지 [단계 1], [단계 2], [단계 3]을 반복한다.

데이터의 크기가 커지게 되면 PAM 알고리즘과 비교해 볼 때 CLARA 알고리즘이 더 효과적으로 군집을 형성할 수 있는 알고리즘임을 알 수 있다. CLARA 알고리즘은 표본 크기에 의존된다. PAM 알고리즘은 전체 데이터에서 메도이드를 추출하자만, CLARA는 전체 데이터에서 추출된 표본에서 최적의 메도이드를 찾는 것이다. 만약 추출된 표본에 좋은 메도이드가 없다면, CLARA는 최적의 군집을 찾지 못할 수도 있다.

2.2.3 CLARANS 알고리즘

CLARANS(Clustering Large Applications based on RANdomized Search) 알고리즘은 R. Ng와 Jiawei Han에 의해 1994년 제안된 알고리즘으로 CLARANS 알고리즘은 그래프의 개념(Graphical abstraction)을 이용한 알고리즘이다. PAM 알고리즘은 초기에 설정된 메도이드들의 집합인 초기 노드부터 모든 경우의 이웃 노드들로 바꾸어 가면서 비용을 계산하여 최소 비용을 가지는 최적의 군집을 찾는것이고, CLARA 알고리즘은 표본을 이용하여 좀 더 적은 양의 이웃 노드들로 평가하는 것이다. CLARA 알고리즘처럼 CLARANS 알고리즘은 노드의 모든 이웃들을 평가하지는 않는다. CLARANS 알고리즘은 노드의 이웃의 표본을 추출하여 평가하는 것이다. CLARANS 알고리즘의 단계는 〈표 4〉 와 같다.

< 표 4 > CLARANS 알고리즘

【단계1】 CLARANS 알고리즘에 이용할 총 반복 회수와 평가할 이웃들의 수를 설정한다.
【단계2】 반복 회수인 i 의 값을 1로 초기화한다.
【단계3】 그래프 $G_{n,k}$에서 현재 사용할 노드를 뽑아 초기 노드로 사용한다.
【단계4】 현재 노드에서 평가할 이웃들이 값의 개수만큼 이웃의 표본을 뽑아 이웃들 사이의 총 비용을 계산한다.
【단계5】 만약 표본으로 사용한 이웃들간의 비용이 더 작다면 현재 비용을 작은 비용으로 갱신하고 [단계 3]으로 돌아가 이웃들 사이의 비용을 다시 계산한다. 이 과정을 표본의 이웃들의 수만큼 반복한다.
【단계6】 반복한 후 최소 비용을 현재 최소 비용으로 저장하고, 수렴할 때까지 [단계 3]과 [단계 4]를 반복한다.
【단계7】 반복 회수가 총 반복 회수가 될 때까지 전체의 과정을 반복한다.

CLARA 알고리즘은 반복을 할때마다 정해진 표본의 개수만큼 표본을 추출하는 것이고, CLARANS 알고리즘은 단계를 거칠때 마다 노드의 이웃의 표본을 뽑는 것이다. 그러므로 CLARANS 알고리즘도 CLARA 알고리즘처럼 표본 추출에 의존된다(J. Han

and M. Kamber. 2000). 평가할 이웃들의 수의 값이 커질수록 CLARANS 알고리즘은 PAM 알고리즘과 유사해지며, 총 비용의 최소값을 찾는데 더 많은 계산과정을 필요로 한다.

3. 모의실험을 통한 군집 분석 알고리즘 비교

3.1 분석 방법

본 논문에서 사용되는 데이터는 입력변수만 있는 데이터로 목표변수가 있을 경우 제외하고 군집 분석을 하였다. k-평균 알고리즘과 메도이드에 의한 알고리즘은 연속형 변수들로만 군집분석을 할 수 있으므로, 입력 변수들은 모두 연속형인 데이터를 사용하였다. 평균 비유사성을 측정하기 위한 식은 유클리디안 거리를 사용하였다. 군집이 잘 분류되었는가에 대한 평가 기준은 평균 거리(average distance)와 군집들간의 분산들의 평균을 이용한다. 데이터는 아래의 〈표 5〉 의 데이터를 사용해 평가한다. 각각의 데이터에 k-평균, PAM, CLARA, CLARANS의 네 가지 알고리즘을 적용시키고, 군집의 수 k는 2개와 3개, 4개로 정하여 평가하겠다.

<표 5> 데이터

데이터 이름	객체 수	입력변수
vehicle	846	18
segmentation	1000	18

3.2 분석 결과

vehicle 데이터에 군집의 수를 달리하면서 네 가지 알고리즘에 적용시킨 결과는 〈표 6〉 에서 제시하고 있다. 〈표 7〉 은 segmentation 데이터의 실험 결과이다. 각각의 데이터들에 k-평균 알고리즘과 메도이드를 이용한 알고리즘들 간에 차이가 있는지를 알아보기 위하여 이상치가 있는 vehicle 그리고 segmentation 데이터에도 네 가지 알고리즘을 적용하여 군집 분석을 해보았다. 〈표 8〉 은 이상치가 있는 vehicle 데이터의 결과, 〈표 9〉 은 segmentation 데이터의 결과이다.

<표 6> vehicle 데이터의 실험 결과

	k-평균		PAM		CLARA		CLARANS	
	평균 거리	분산	평균 거리	분산	평균 거리	분산	평균 거리	분산
k=2	**76.52362**	**9591.499**	80.77152	**9559.135**	82.4734	9670.147	**79.794**	9605.12
k=3	71.32229	6534.245	**64.60056**	**6203.892**	68.56543	**6227.993**	**64.1952**	6237.519
k=4	59.91864	4613.635	**56.11855**	**4456.854**	58.29263	4539.312	**56.4237**	**4509.967**

<표 7> segmentation 데이터의 실험 결과

	k-평균		PAM		CLARA		CLARANS	
	평균 거리	분산	평균 거리	분산	평균 거리	분산	평균 거리	분산
k=2	97.77234	9839.139	91.7184	9477.385	**91.03559**	**8877.833**	**90.18463**	**8785.946**
k=3	**81.29118**	**7353.75**	81.92808	7404.528	86.37084	8076.916	84.06005	8037.3573
k=4	**75.39002**	**5566.634**	72.18646	6057.746	77.1717	6215.627	76.57865	**6164.1201**

〈표 6〉 에서도 각 알고리즘들간의 평균 거리를 비교해 보면, PAM 알고리즘과 CLARANS 알고리즘의 평균 거리와 분산이 나머지 두 알고리즘보다 비교적 작게 측정되었다. 〈표 7〉 에서는 k-평균 알고리즘과 CLARANS 알고리즘의 평균 거리와 분산이 비교적 나머지 두 알고리즘보다 작게 측정되었다.

<표 8> 이상치가 있는 vehicle 데이터의 실험 결과

	k-평균		PAM		CLARA		CLARANS	
	평균 거리	분산	평균 거리	분산	평균 거리	분산	평균 거리	분산
k=2	180.6117	29516.48	119.942	340237.3	120.8943	334753.7	119.521	3538.266
k=3	179.4243	911556.9	96.84613	240770	97.53197	240237.7	96.27335	240491.3
k=4	90.16701	14590.76	60.67178	22634.93	207.4478	393083	61.7523	25396.97

<표 9> 이상치가 있는 segmentation 데이터의 실험 결과

	k-평균		PAM		CLARA		CLARANS	
	평균 거리	분산	평균 거리	분산	평균 거리	분산	평균 거리	분산
k=2	118.7719	11511.46	92.8614	14091.46	92.3847	13431.76	92.01764	13622.97
k=3	100.0708	10841.56	82.65367	12665.15	87.52656	13218.67	86.0455	13300.92
k=4	81.80669	9376.754	73.46778	13106.38	78.36584	12840.74	76.54689	12979.93

〈표 8〉 에서 보면, k-평균 알고리즘보다 메도이드를 이용한 나머지 세 가지 알고리즘의 평균 거리가 군집 수에 관계없이 모두 감소하였고, PAM 알고리즘과 CLARANS 알고리즘의 거리들은 유사하지만, CLARA 알고리즘에서는 두 알고리즘보다 평균 거리가 약간 높은 수치가 측정되었다. 분산에서는 k-평균 알고리즘과 CLARA 알고리즘의 분산이 비교적 낮게 측정되었다. 〈표 9〉 에서 보면, 메도이드를 사용한 모든 알고리즘이 군집 수에 관계없이 k-평균 알고리즘보다 평균 거리가 작았고, 그 중 CLARANS 알고리즘의 평균 거리가 가장 작게 나타났다. 분산은 k-평균 알고리즘보다 커지는 경향을 보이고 있다.

4. 결론

알고리즘들의 특징을 알기 위해 평균 거리와 분산을 측정하여본 결과 몇 가지 결론을 제시할 수 있다. 세 가지 데이터에 적용시켰을 때, CLARANS 알고리즘은 평균 거리와 분산 모두 나머지 알고리즘보다 수치가 낮게 측정되었다. segmentation 데이터에서는 k-평균 알고리즘과 CLARANS 알고리즘이 평균 거리와 분산이 동시에 낮은

수치가 측정이 되었다. 분석 결과에서 볼 때 대부분 CLARANS 알고리즘이 나머지 알고리즘보다 평균 거리가 더 짧고, 분산도 작아지는 경향이 있다. 결과적으로, CLARANS 알고리즘은 데이터나 군집의 수에 관계없이 좋은 결과를 보여주었고, 데이터가 작을수록 PAM이 데이터가 커질수록 k-평균 알고리즘이 적합함을 알 수 있다. 이상치가 있는 데이터들에서는 메도이드를 이용한 알고리즘들의 평균 거리가 급격히 감소하였고, CLARA 알고리즘보다 CLARANS 알고리즘의 평균 거리가 대부분 작게 측정되었다. 위의 두 개의 데이터만을 가지고는 일반화 시키는데 무리가 있었다. 앞으로 더 많은 데이터를 사용하여 알고리즘들의 성격을 파악하도록 해야겠다.

이상치는 대부분의 데이터와는 차별적인 성격을 가지면서 동떨어져 있는 데이터이다. 경우에 따라서는 분석에서 제외하지만, 이상치가 어떤 과학적인 정보를 가지고 있는 경우도 허다하므로, 이러한 경우에는 제외할 수 없다. 그러므로 이상치가 존재하는 데이터에서 이상치를 제외하고 군집분석을 적용하기보다는 메도이드를 이용한 알고리즘을 적용하는 것이 더 효과적이다.

참고문헌

1. R. Ng and J. Han. (1994) Efficient and Effective clustering Methods for Spatial Data Mining. In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pages 144-155, Santiago. Chile, Sept. 1994.
2. R. Ng and J. Han (2001) CLARANS : A Method for Clustering Objects for Spatial Data Mining
3. L. Kaufman and P.J. Rousseeuw. (1990) Find Groups in Data : an Introduction to Cluster Analysis, John Wiley & Sons.
4. M. Ester. H.-P. Kriegel. and X. Xu. Knowledge discovery in large spatial databases : Focusing techniques for efficient class identification. In Proc. 4th Int. Symp. Large Spatial Database(SSD' 95), pages 67-82, Portland, ME, Aug. 1995.
5. J. Han and M. Kamber. (2000), Data Mining : Concepts and Techniques, , The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufman Publishers.