Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 759~768

K-means Clustering using Grid-based Representatives

Hee Chang Park[1] · Sun Myung Lee[2]

Abstract

K-means clustering has been widely used in many applications, such that pattern analysis, data analysis, market research and so on. It can identify dense and sparse regions among data attributes or object attributes. But k-means algorithm requires many hours to get k clusters, because it is more primitive and explorative. In this paper we propose a new method of k-means clustering using the grid-based representative value(arithmetic and trimmed mean) for sample. It is more fast than any traditional clustering method and maintains its accuracy.

Keywords : arithmetic mean, data mining, k-means clustering, trimmed mean

1. 서론

데이터 마이닝(data mining)은 방대하고 다양한 형태의 데이터로부터 의사결정에 유용한 정보를 발견하려는 일련의 데이터 분석 및 모형선정 기법이다. 데이터 마이닝의 기법에는 의사결정나무, 연관성 규칙, 클러스터링, 그리고 신경망 분석 등이 있다. 이들 중에서 클러스터링은 다양한 특성을 지닌 관찰대상을 유사성을 바탕으로 동질적인 집단으로 분류하는데 쓰이는 기법으로 이를 크게 나누면 분할 군집법과 계층적 군집법이 있다. 그 중에서 분할 군집법은 데이터들을 임의의 부분집합으로 분할을 한 후 데이터들을 유사한 그룹으로 재배치하는 군집방법이다. 분할 군집법의 종류에는 본 논문에서 연구하고자 하는 k-means 알고리즘과 k-medoids 알고리즘, k-prototypes 알고리즘, k-modes 알고리즘 등이 있다.

k-means 알고리즘은 MacQueen(1967)에 의해 처음 소개되었으며, 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 무게중심 (평균)을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다. Kaufman과

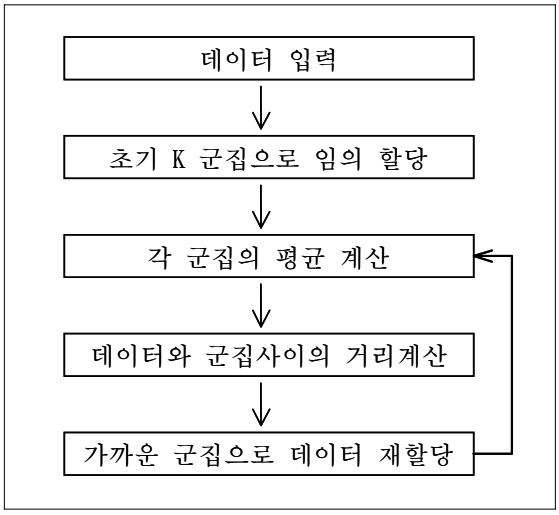
1) First Author : Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea
E-mail : hcpark@sarim.changwon.ac.kr
2 Graduate Student Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea

Rousseeuw(1990)는 군집의 대표값을 중앙값으로 하는 k-medoids 방법인 PAM(partitioning around medoids)과 CLARA(clustering large applications) 알고리즘을 제안하였다. 이 알고리즘은 데이터를 샘플링하여 PAM을 적용한 방법으로, 표본을 잘 뽑았다면 표본의 중앙값은 전체 데이터의 중앙값에 근사하며, 더 나은 근사값을 위해 CLARA는 다중 샘플을 사용한다. Ng 등(1994)은 CLARA를 더욱 향상시킨 CLARANS(clustering large applications based on randomized search)를 제안하였다. Huang(1997, 1998)은 k-means가 연속형 데이터에 대해 한정된 단점을 보완한 연속형과 범주형의 혼합된 데이터에 대한 k-prototypes 알고리즘을 제안하는 동시에 범주형 데이터에 대해서 k-modes 알고리즘을 제시하였다. Chu 등(2002)은 k-medoids의 약점을 극복하기 위해 효과적인 샘플링 기법을 추가하여 MCMRS(Multi-Centroid, Multi-Run Sampling Scheme) 알고리즘을 제시하였으며, 또한 이들은 MCMRS의 발전된 더 진보된 샘플링 기법인 IMCMRS(Incremental Multi-Centroid, Multi-Run Sampling Scheme) 알고리즘을 제안하였다.

데이터 마2이닝은 대량의 데이터를 대상으로 하므로 그 만큼 처리 속도에 대한 발전된 많은 방법이 연구되고 있다. 이러한 속도 문제를 해결하기 위해 본 논문에서는 k-means 알고리즘에 대해 그리드를 기반으로 한 대표값(산술평균, 절사평균)을 이용한 알고리즘을 제안하고자 한다. 2절에서는 k-means에 대한 일반적인 방법을 살펴본 후, 3절에서는 그리드 기반 대표값을 이용한 k-means 알고리즘을 구현하며, 4절에서는 모의실험을 통하여 본 연구에서 제시한 기법과 기존의 기법을 비교하고자 한다. 마지막으로 5절에서 본 연구의 결론을 맺고자 한다.

2. k-means 군집방법

기본적인 k-means 군집방법의 수행단계는 다음과 같다.



<그림 1> 기본적인 k-means 수행 단계

[단계 1] 데이터들을 임의의 k 개의 군집으로 분할한다.
[단계 2] 각 군집의 평균을 구한다.
[단계 3] 각 데이터 점들과 각 군집의 평균과의 거리를 구하여 데이터 점들을 가장 가까운 군집으로 재할당한다.
[단계 4] 데이터 점들의 재배치가 없을 때까지 [단계 2] 및 [단계 3]의 과정을 반복한다.

군집분석에서 군집간의 유사성 측정은 거리로써 나타낸다. 서로 다른 개체 사이의 거리 $d_{ij} = d(X_i, X_j)$를 구하는 방법에는 유클리디안(Euclidean) 거리, 유클리디안 제곱거리, 마할라노비스(Mahalanobis) 거리, 그리고 민코우스키(Minkowski) 거리 등이 있으며, 본 논문에서는 유클리디안 제곱거리를 이용하고자 한다.

3. 그리드 기반 대표값의 k-means 군집방법

데이터 마이닝은 대량의 데이터를 대상으로 계산을 수행한다. 그러므로 데이터 크기와 변수 수에 따라 처리 속도에 상당한 차이가 있다. 유용한 정보를 빠른 시간 내에 얻어야 하는 경우 속도는 큰 문제점이 된다. 본 절에서는 그리드를 기반으로 한 대표값을 이용하여 k-means를 수행하는 알고리즘을 구현하고자 한다. 여기서 그리드는 유사한 성질의 데이터 집합이다.

3.1 그리드의 설정

클러스터링의 수행과정을 최소화하기 위해 먼저 데이터 개체들을 적당한 그리드 간격으로 분할하여 각 그리드별로 대표값을 구한다. 그리드 간격이 넓으면 넓을수록 계산 과정이 줄어들며 그만큼 시간이 단축되지만 정확도는 떨어질 것이고, 그리드 간격이 좁아지면 계산 과정이 늘어나서 시간은 늘어나지만 정확도는 더욱 향상될 것이다. 따라서, 정확도 대비 속도의 적절한 균형점을 찾아야 할 것이며, 이는 곧 그리드 간격을 어떻게 설정하느냐 하는 문제로 귀착된다. 본 논문에서 제시하는 알고리즘의 그리드 간격을 GI(Grid Interval)라고 할 때 GI는 다음과 같이 설정한다.

$$GI_v = \frac{\max_v - \min_v}{n^{\frac{1}{p}}} \tag{3.1}$$

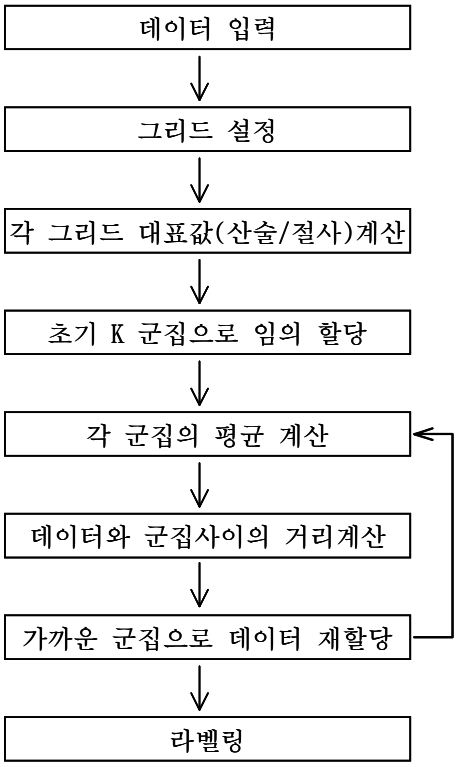
여기서 v는 v번째 변수를 나타내며, $\max_v$와 $\min_v$는 각각 v번째 변수의 최대값과 최소값을 나타낸다. n은 각 변수의 데이터가 이루는 쌍의 수가 되며, 그 쌍은 곧 거리를 위한 좌표점으로 나타낸다. p는 차원수이다. 데이터 공간이 2차원인 경우, 데이터들의 분포가 정사각형으로 고루 분포되었다고 가정했을 때 그 넓이는 n이고 한 변의 길이는 $\sqrt{n}$이 된다.

3.2 그리드 기반 대표값

기존의 전형적인 k-means 기법은 클러스터링 계산 과정에 있어서 데이터 또는 데이터 개체들의 수에 의존하는 반면에 그리드 기반 대표값의 k-means 기법은 셀의 수에 의존하므로 기존의 방법에 비해서 빠른 처리 시간을 가진다. 본 논문에서는 데이터를 동일한 간격의 그리드로 나눈 후 각 그리드별로 산술평균 및 절사 평균을 구한 후, 이를 이용하여 클러스터링을 수행한다.

3.3 알고리즘 구현

그리드 기반 대표값의 k-means 군집분석을 위한 수행 과정은 다음과 같다.



<그림 2> 그리드 기반 대표값의 k-means 군집 과정

이에 대한 알고리즘은 다음과 같이 구현된다.

```
clustering()
{
  int k, n, p;
  float Data;
  Data = InsertData();
  GridInterval = (Max − Min) / n^(1/p);
  while(Grid_X)
  {
    while(Grid_Y)
    {
      while(i <= n)
      {
        if((GridX_min < x[i] <= GridX_max) &&
              (GridY_min < y[i] <= GridY_max))
        {
          DataGrid[i] = GridNo;
        }
        i++;
      }
    }
  }

  while(G_No <= DataGrid)
  {
    G_Mean[G_No] = Σ_{v=1}^{v=G_n} x_v / G_n;
    G_No++;
  }

  k_means();

  while(Grid)
  {
    Nearst = 0;
    while(remainData)
    {
      Dist = (GridCenter − Data)^2
      if(Dist < Nearst)
      {
        Nearst = Dist;
        DataGird[i] = G_No;
      }
    }
  }
}
```

4. 모의실험

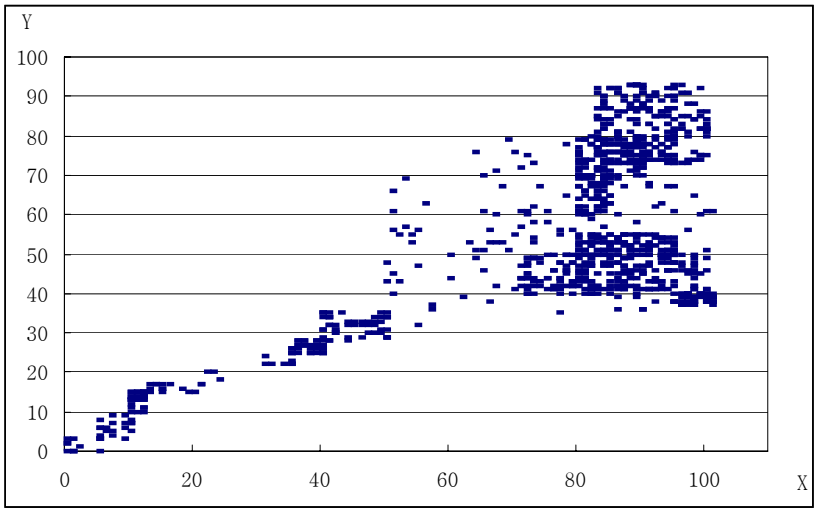
앞 절에서 구현한 그리드 기반 표본의 대표값에 의한 k-means 알고리즘을 바탕으로 수행 시간 및 정확도를 비교하기 위하여 모의실험을 실시하였다. 본 실험의 구현 환경은 다음과 같다.

CPU : Intel Pentium4-1.8GHz Northwood
RAM : 512MB
O/S : Microsoft Windows XP Professional
Language : JAVA J2SDK 1.4.0
Database : MySQL 3.23.51 (External Linux Server)

모의실험을 위해 두 개의 변수로 이루어진 1000건의 데이터를 랜덤하게 발생시켜 기본 데이터 셋으로 사용하였다. 실험은 기본 데이터 셋에서 랜덤 샘플링하여 사용하였다. 이들 데이터 셋에 대한 특징은 다음과 같다.

데이터 수 : 1000건
변수값의 범위 : X(0 ~ 101), Y(0 ~ 93)
$\overline{X}$ = 75.7, S_x = 24.9
$\overline{Y}$ = 53.9, S_y = 22.0

데이터 셋의 분포는 <그림 3>과 같다.



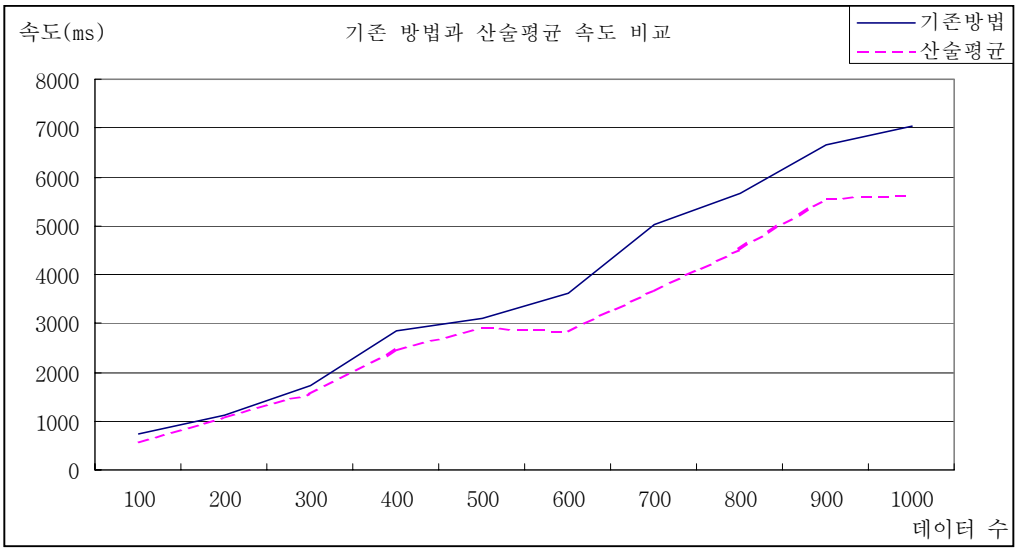
<그림 3> 데이터 셋의 분포도

그리드를 기반으로 산술평균하여 k-means를 수행하였을 때 수행 속도와 정확도는 <표 1>과 같다.

<표 1> 그리드 기반 산술평균에 의한 방법의 수행 속도와 정확도

데이터 수	속도(ms)	정확도
100	581	100%
200	1082	97.5%
300	1582	97.5%
400	2464	97.8%
500	2914	98.4%
600	2844	98.3%
700	3676	99.3%
800	4506	97.8%
900	5558	97.9%
1000	5618	97.9%

기존의 방법과 그리드 기반 산술평균에 의한 방법의 수행 속도를 비교한 결과는 <그림 4>와 같다.



<그림 4> 기존 방법과 그리드 기반 산술평균에 의한 방법의 수행 속도

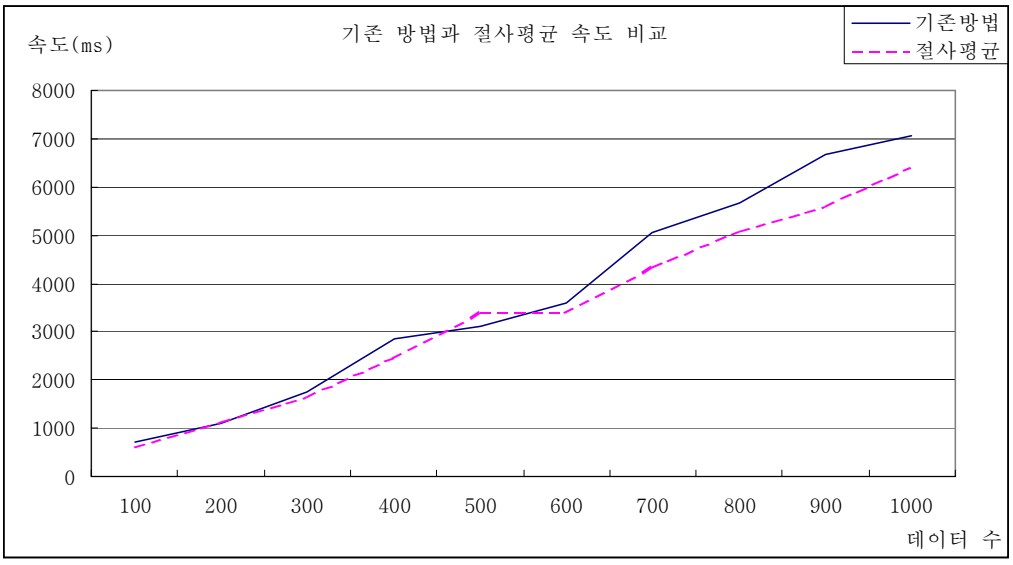
위의 결과에서 보는 바와 같이 기존의 방법에 비해 산술평균이 속도에 있어서 효과적인 것을 알 수 있다. 반면에 정확도에 있어서는 <표 1>에서 보는 것과 같이 다소 떨어지는 것을 알 수 있다. 이러한 실험 결과로 볼 때 수행 속도와 정확도의 적절한 균형점을 찾아 효율적인 방법을 찾는 것이 좋은 방향이 되는 것을 알 수가 있다.

산술평균의 그리드 기반 절사평균의 수행 속도와 정확도는 <표 2>와 같다.

<표 2> 그리드 기반 절사평균 수행 속도와 정확도

데이터 수	속도(ms)	정확도
100	620	97%
200	1122	97.5%
300	1653	97.7%
400	2474	97.8%
500	3405	98.2%
600	3395	98.2%
700	4346	99%
800	5077	97.4%
900	5598	97.9%
1000	6420	97.8%

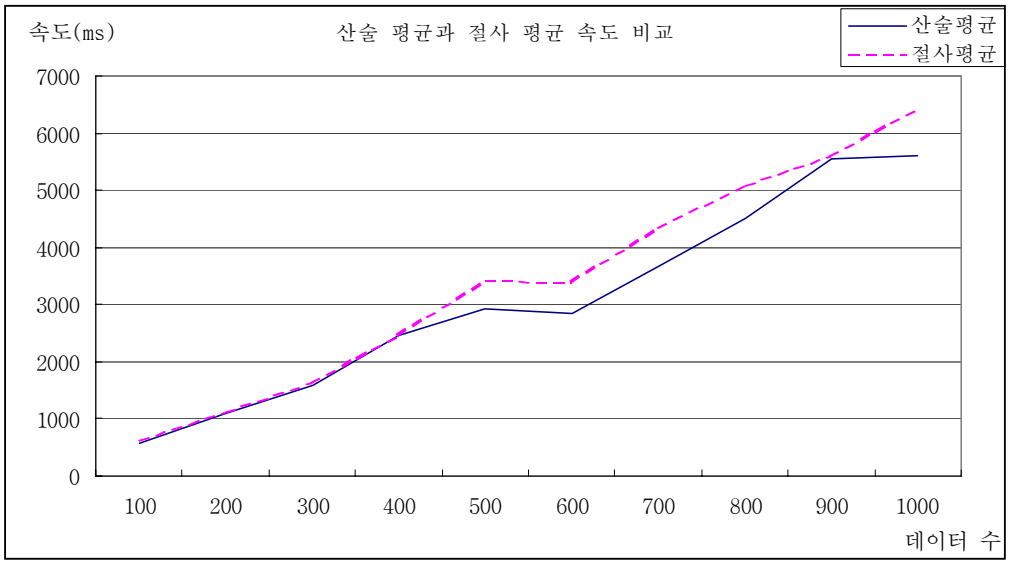
기존 방법과 그리드 기반 절사 평균의 수행 속도 비교한 결과는 <그림 5>와 같다.



<그림 5> 기존 방법과 절사평균 수행 속도

위의 결과에서 보는 바와 같이 기존의 방법에 비해 그리드 기반 절사평균의 수행 속도가 다소 효과적인 것을 알 수 있다. 산술평균과 마찬가지로 정확도 면에서는 다소 떨어지지만 속도의 향상을 가져온다.

산술평균과 절사평균에 의한 방법의 수행 속도를 비교한 결과는 <그림 6>과 같다.



<그림 6> 산술 평균과 절사 평균의 수행 속도

산술평균이 절사평균에 비해 수행 속도면에서 다소 효과적인 것을 <그림 6>을 통해서 확인 할 수 있다. 데이터 수가 증가할수록 속도의 차이가 뚜렷이 나는 것을 알 수가 있다.

위의 두 실험 결과에서 보듯이 그리드 기반 산술평균과 절사평균의 정확도는 비슷한 것을 알 수 있다. 절사평균은 산술평균의 이상점에 민감한 점을 개선하기 위한 방법이었지만 본 논문의 실험에서는 명확한 차이는 확인할 수는 없었다.

5. 결론

데이터 마이닝에서 처리 속도 문제는 해결해야 할 과제 중의 하나이다. 이러한 속도 문제를 해결하기 위해 본 논문에서는 분할 군집법에서 가장 일반적으로 사용되고 있는 k-means 알고리즘에 대해 그리드를 기반으로 대표값 알고리즘을 제안하였다. 동시에 본 연구에서 제시한 기법과 기존의 기법을 모의실험을 통하여 비교하였으며, 수행속도와 정확도에서 만족할 만한 수준의 결과가 얻어짐을 확인하였다.

참고문헌

1. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1, 281-297.
2. Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.
3. Ng, R. and Han, J. (1994). Efficient and effective clustering method for spatial data mining. *Proceedings of the 20th. Very Large Data Bases Conference*. 144-155.
4. Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific*. 21-34.
5. Huang, Z. (1998). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Tucson, Arizona, USA, May, 146-151.
6. Chu, S.C, Roddick, J.F and Pan, J.S. (2002). Efficient k-medoids algorithms using multi-centroids with multi-runs sampling scheme. *Proceedings of Workshop on Mining Data for CRM,* (Taipei, Taiwan), Springer. 14-25.
7. Chu, S.C, Roddick, J.F and Pan, J.S. (2002). An Incremental Multi-Centroid, Multi-Run Sampling Scheme for k-medoids-based Algorithms-Extended Report. *Proceedings of Second International Conference on Knowledge Discovery and Data Mining.* 197 - 202.

[2005년 8월 접수, 2005년 9월 채택]