Clustering Algorithm by Grid-based Sampling

Hee Chang Park[1], Jee Hyun Ryu[2]

Abstract

Cluster analysis has been widely used in many applications, such that pattern analysis or recognition, data analysis, image processing, market research on on-line or off-line and so on. Clustering can identify dense and sparse regions among data attributes or object attributes. But it requires many hours to get clusters that we want, because of clustering is more primitive, explorative and we make many data an object of cluster analysis. In this paper we propose a new method of clustering using sample based on grid. It is more fast than any traditional clustering method and maintains its accuracy. It reduces running time by using grid-based sample. And other clustering applications can be more effective by using this methods with its original methods.

1. 서 론

데이터 마이닝(data mining)은 데이터를 어떻게 활용할 것인가에 대한 중요 기술중의 하나이며, 이는 수많은 데이터(data)로부터 쉽게 드러나지 않는 의미 있는 정보들을 채굴(mine)해 내는 과정이라 할 수 있다. 마이닝에서는 의미 있는 정보를 캐내기 위한 여러 가지 기법들이 존재한다. 여기에는 각종 통계적 기법뿐만 아니라 신경망(neural networks), 의사 결정 나무(decision tree), 연관성 규칙(association rule), 유전자 알고리즘(genetic algorithm), 메모리-기반 추론(memory-based reasoning), 그리고 클러스터링(clustering) 등이 있다.

클러스터링은 다양한 특성을 지닌 관찰대상을 상사성(또는 비상사성)을 바탕으로

1) 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

동질적인 집단으로 분류하는 데 쓰이는 기법으로 아주 방대한 양의 데이터를 대상으로 하고 있으며 군집들의 개수 또는 구조, 그리고 데이터들의 분포 등에 관해 아무런 가정을 두지도 않는다. 클러스터링은 Tryon(1939)에 의해 처음 소개된 이후로 Jardine과 Sibson(1971), Anderberg(1973), Tryon과 Bailey(1973), Hartigan(1975), Jain과 Dubes 등(1988)에 의해 일반적인 클러스터링 기법의 형태를 갖추게 되었다.

기존의 클러스터링 방법은 알고리즘별로 계층적인 방법, 분할적인 방법, 밀도 기반의 방법, 그리드 기반의 방법, 모델 기반 방법 등으로 나뉘어진다. 본 논문에서는 계층적인 클러스터링의 기법에서 그리드 기반으로 하여 추출된 표본을 이용한 클러스터링 기법을 제안하고자 한다.

계층적인 클러스터링 방법은 비슷한 성질의 데이터 개체들을 그룹화 하여, 그 결과 생성된 군집을 트리 형태로 만들어 나가는 방법으로 Kaufman과 Rousseeuw(1990), Zhang 등(1996), Guha 등(1998, 1999), Karypis 등(1999)과 같은 많은 연구자들에 의해 연구가 진행되어 왔다. Kaufman과 Rousseeuw는 여러 개의 개체 또는 클러스터를 비슷한 성질의 것끼리 하나씩 묶어나가는 방법인 AGNES(AGglomerative NESting)와 하나의 클러스터로부터 시작해서 비슷한 성질의 것끼리 분할해 나가는 DIANA(DIvisive ANAlysis)를 제안하였다. Zhang 등은 BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies)를 제안, 다른 기술들을 적용하기 전에 CF 트리(Clustering Feature tree)를 이용한 계층적 클러스터링을 수행하였다. Guha 등은 이상치에 민감하지 않으면서도 방대한 데이터에 효율적인 CURE(Clustering Using REpresentative)를 제안하였고, Karypis 등은 서로 다른 두 클러스터간의 내부적 관계를 고려하는 다이나믹 모델링을 활용한 Chameleon(A Hierarchical Clustering Algorithm Using Dynamic Modeling)을 제안하였다. 그리고 Guha 등은 데이터 개체의 범주형 속성에 대해서 클러스터링을 가능하게 한 ROCK(RObust Clustering using linKs)을 제안하였다.

클러스터링은 단지 데이터들 사이의 상사(相似)성(또는 비상사성)에 근거하여 자연스럽게 군집을 찾아 나가는 아주 원시적이며 탐색적인 통계기법이다. 이러한 성질 때문에 클러스터링은 데이터 마이닝의 다른 기법들에 비해 분석 속도가 매우 뒤떨어진다는 단점이 있으며, 이는 곧 실제 여러 분야에서 클러스터링을 바로 적용하기에는 많은 문제점이 있다는 것을 의미하게 되었다. 이러한 이유로 클러스터링은 조금이라도 빠르고 정확한 알고리즘의 개발이 필수가 되었으며 지금까지도 많은 연구가 진행되고 있다.

본 논문에서는 정확성을 유지함과 동시에 수행속도를 높일 수 있는 방법으로 그리드 기반 표본을 이용한 클러스터링 기법을 제안하고자 한다. 2절에서 전형적인 클러스터링 방법에 대해서 알아보고, 3절에서는 그리드 기반의 표본을 전형적인 클러스터

링 기법에 이용하는 방법에 대해서 알아보고자 한다. 4절에서는 예제 및 모의 실험을 통해 본 연구에서 제시한 기법과 전형적인 클러스터링 기법을 비교하여 수행속도와 정확도에서 만족할 만한 수준의 결과가 얻어짐을 확인하고자 한다. 마지막으로 5절에서 본 연구의 결론을 맺고자 한다.

2. 클러스터링

클러스터링은 다양한 특성을 가진 수많은 데이터를 비슷한 성질의 데이터끼리 묶어주는 데이터 마이닝의 한 기법으로서 군집의 수 혹은 군집의 구조에 대한 가정이 없으며, 오직 데이터들 사이의 상사성(또는 비상사성)에 의하여 군집을 형성하고, 형성된 군집의 특성을 파악하여 군집들 사이의 관계를 분석하는 기법이다. 따라서 분류된 군집들은 상호 배타적이어서 한 군집에 속한 개체들은 서로 유사한 성질을 가지고 있으며 이들은 다른 군집에 속한 개체들과 상이한 성질을 가지고 있다. 기본적인 클러스터링 과정은 다음과 같다.

1) N개의 데이터 개체에 대해 p개의 변수를 관찰하여 크기 $(N\times p)$인 자료행렬을 구한다.
2) N개의 데이터 개체 사이의 크기 $(N\times N)$인 거리행렬을 구한다.
3) 거리 행렬로 군집화 방법을 통해서 군집들을 형성한다.
4) 각 군집의 성격이나 상호관계를 분석한다.

클러스터링을 수행하기 위한 자료를 정리한 다음에는 묶여지는 각 데이터간의 상사성(또는 비상사성)의 정도를 측정하는 기준척도가 필요하다. 보통 이러한 정도를 측정할 때에는 상사성(또는 비상사성)을 거리로 환산하여 거리가 가까운 대상들을 동일한 집단에 포함시키므로 어떠한 변수들을 설정할 것인가 하는 것이 먼저 해결되어야 할 것이다. 클러스터링에서는 의미 없는 변수를 제거하는 과정이 없으므로 선택된 변수들이 모두 동일한 비중을 가진다. 두 개체 사이의 거리의 종류에는 유클리드 거리, 유클리드제곱 거리, Mahalanobis 거리, 그리고 Minkowski 거리 등이 있다. 본 논문에서는 이들 중에서 식 (2.1)와 같이 정의되는 유클리드제곱 거리를 이용하고자 한다.

$$d_{ij}^2=(X_i-X_j)'(X_i-X_j) \tag{2.1}$$

기준 척도를 정한 후에는 실제로 대상들에 대해서 군집화를 해나가야 한다. 군집화

방법에는 여러 가지가 있으나 크게 계층적 군집화 방법과 비계층적 군집화 방법으로 나누어진다. 계층적 군집화 방법에는 최단 연결법(Single Linkage Method), 최장 연결법(Complete Linkage Method), 평균 연결법(Average Linkage Method), 중심 연결법(Centroid Linkage Method), 중위수 연결법(Median Linkage Method), 그리고 Ward의 방법 등이 있으며, 이는 상사성이 큰 군집끼리 묶어 나가는 방법이다.

본 논문에서는 군집간의 최단 거리 중 가장 최소 거리를 가지는 군집끼리 병합하는 방법인 최단 연결법을 이용하였다. 이 방법에서의 최단 거리는 식 (2.2)과 같이 정의된다.

$$d(U, V) = \min[d(x, y) | x \in U, y \in V] \tag{2.2}$$

여기에서 U와 V는 임의의 군집이며, x와 y는 해당 군집의 임의의 두 개체이다. 두 군집 U와 V사이의 거리 d_{UV}를 각 군집에 속하는 임의의 두 개체들 사이의 거리 중 최단거리로 정의하여 가장 상사성이 큰 군집을 묶어 나간다. 최단 연결법이 가지는 장점은 Jardine과 Sibson(1968)에 의해 연구된 바, 수리적인 면이 우수하며, 컴퓨터 처리시간이 다른 방법들에 비해 빠르고, 순서적 의미를 가지는 자료에 대해 좋은 결과를 제공하는 기법이다.

3. 그리드 기반 표본의 클러스터링

전형적인 방법의 클러스터링은 아주 방대한 양의 데이터를 대상으로 하며, 클러스터링 자체가 원시적이며 탐색적으로 접근하기 때문에 만족할 만한 최종 클러스터들을 얻기까지 매우 많은 계산과정을 거치게 되며, 그만큼 많은 시간을 요구하게 된다. 본 연구에서는 클러스터링의 고질적인 수행시간의 많은 소비에 대해 이를 최소화하기 위한 방법으로 먼저 샘플링에 대해서 알아본 후, 그리드 기반 샘플링을 이용한 클러스터링 기법을 제안하고자 한다.

본 논문에서는 클러스터링의 수행과정을 최소화하기 위해 샘플링 이전에 그리드를 사용하여 데이터 개체들을 그리드 간격으로 분할한다. 그리드 간격이 넓으면 넓을수록 계산 과정이 줄어들며 그만큼 시간이 단축되지만 정확도는 떨어질 것이고, 그리드 간격이 좁아지면 계산 과정이 늘어나서 시간은 늘어나지만 정확도는 더욱 향상 될 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.

$$GI_v = \frac{\max_v - \min_v}{n^{\frac{1}{p}}} \quad (3.1)$$

여기서 v는 각 변수를 나타내며, max와 min은 각각 해당 변수의 최대값과 최소값을 나타낸다. n은 결측치가 없다고 가정할 때, 각 변수들의 데이터들이 이루는 쌍의 수가 되며, 그 쌍은 곧 거리를 위한 좌표점으로 나타낸다. p는 차원수 즉, 변수의 개수이다. 데이터 공간이 2차원인 경우, 데이터들의 분포가 정사각형으로 고루 분포되었다고 가정했을 때 그 넓이는 n이고 한 변의 길이는 $\sqrt{n}$이 된다.

그리드 간격은 데이터들이 고루 분포되었을 경우 유사한 값을 가지는 개체들은 그리드별로 동일한 셀에 포함되도록 하였다. 먼저 각 좌표축별로 그리드 간격이 설정되면 원 데이터를 해당 그리드로 분할하며, 이는 곧 그리드의 각 셀을 하나의 군집으로 보는 첫 번째 클러스터링이 수행되는 시점이라고 볼 수 있다. 데이터가 정사각형 모양으로 고루 분포되었을 경우 그리드별 각 셀에는 한 점만을 포함할 것이며, 이 경우 클러스터의 수는 n이 될 것이다. 하지만, 셀 별로 각 한 점만을 포함할 경우는 거의 없으므로 최소한 최초 클러스터의 수는 n보다는 작게 될 것이며, 한점도 포함하지 않는 셀은 계산에서 제외된다.

기존의 전형적인 클러스터링 기법은 클러스터링 계산 과정에 있어서 데이터 또는 데이터 개체들의 수에 의존하는 반면에 그리드 기반 표본의 클러스터링은 셀의 수에 의존하므로 기존의 방법에 비해서 빠른 처리 시간을 가진다. 본 연구에서 제시하는 그리드 기반의 클러스터링 수행 단계는 다음과 같다.

[단계 1] 데이터 처리

클러스터링을 수행할 데이터를 얻어 새로운 데이터 셋을 만든다. 데이터수가 작다면 직접 입력을 해도 되겠으나, 일반적으로 클러스터링에서는 대량의 데이터를 다루므로 데이터베이스(Database) 또는 파일 시스템으로부터 데이터를 얻어온다. 일단 데이터가 얻어지면 개체 번호, 각 점들로 이루어진 개체들을 선언하며 이 개체들로 데이터 셋을 다시 구성한다.

[단계 2] 그리드 설정

새로 구성된 데이터 셋으로부터 각 차원별 그리드 간격을 얻은 후, 데이터 셋을 그리드 간격에 맞게 분할한다.

[단계 3] 랜덤 샘플링

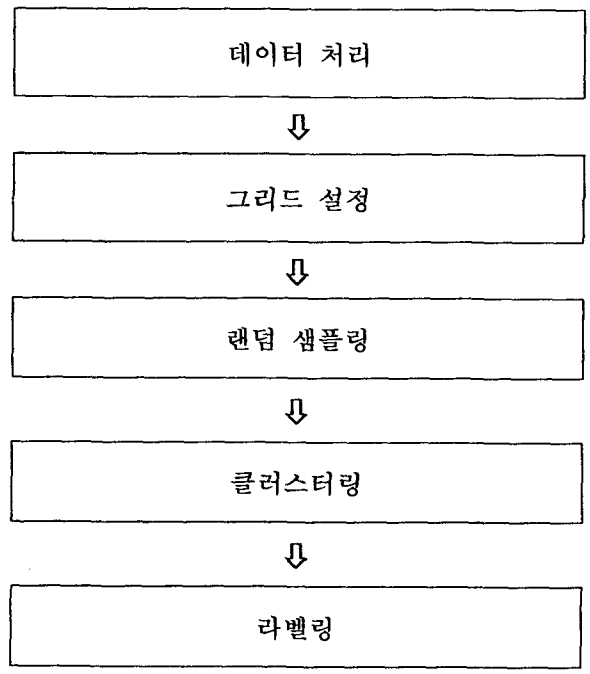
다음 단계는 그리드별 각 셀 단위로 한 점씩 랜덤 샘플링을 수행하여 클러스터링을 수행 할 대표점 셋을 만든다.

[단계 4] 클러스터링

다음 단계에서는 얻어진 대표점으로 클러스터링을 수행하며, 본 연구에서는 유클리디안 제곱거리와 최단연결법을 이용하였다.

[단계 5] 라벨링

샘플링 되지 않은 원 데이터를 해당 클러스터로 합병과 동시에 라벨링을 수행하며 최종 결과를 출력한다.



<그림 1> 그리드 기반 표본의 클러스터링 수행 단계

4. 예제 및 모의실험

본 장에서는 4절에서 구현한 알고리즘을 바탕으로 수행 시간 및 정확도를 비교하기 위하여 예제 및 모의실험을 실시하였다. 본 실험의 구현환경은 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)

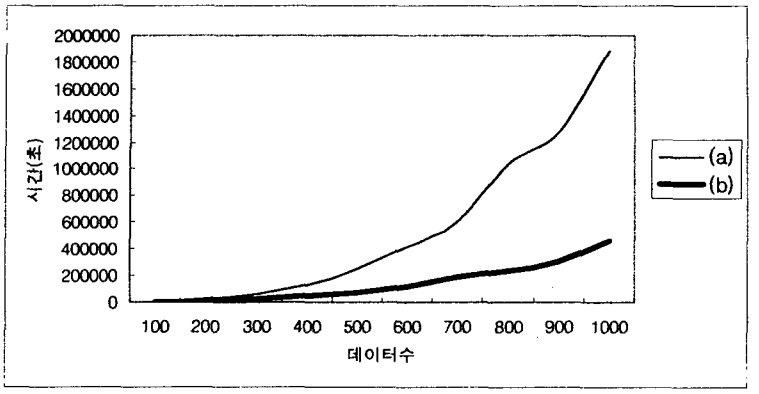
첫 번째 실험은 전형적인 클러스터링(a)과 그리드 기반 클러스터링(b)의 수행 시간을 비교하기 위하여 데이터의 특정 분포에 결과가 치우치지 않도록 데이터를 랜덤 발생시켜 고루 분포되도록 한 후 실험을 실시하였다. 데이터는 0.00부터 1000.00 까지의 값을 가지는 두 변수를 이용하였으며 결과는 다음과 같다.

<표 1> 데이터 수와 방법별 수행 시간 비교1

(단위 : 초)

데이터 수	(a)	(b)
100	2.26	0.90
200	22.90	5.49
300	62.27	19.48
400	124.74	42.15
500	249.89	68.76
600	417.01	120.62
700	603.06	192.24
800	1043.69	235.81
900	1275.71	308.58
1000	1878.88	467.00

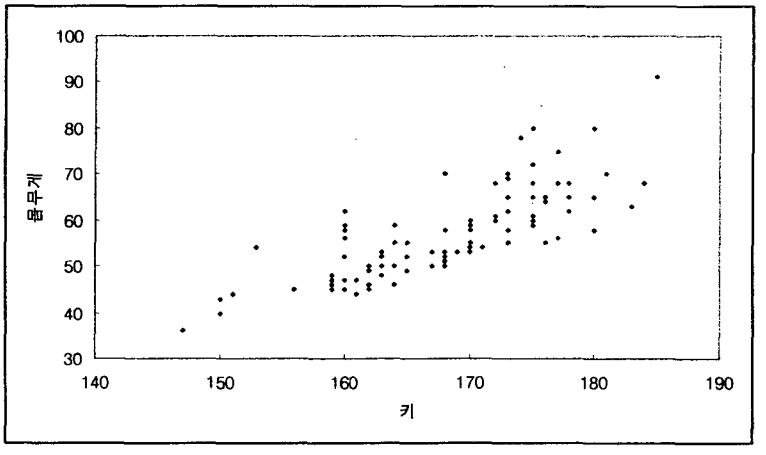
그리드 기반 클러스터링 기법에 비해서 전형적인 클러스터링 기법은 데이터 수 증가량에 비해 수행시간이 급격하게 증가되는 것을 알 수 있다. <표 1>을 그래프로 나타내면 <그림 2>과 같다.



<그림 2> (a)와 (b)의 수행 시간 비교1

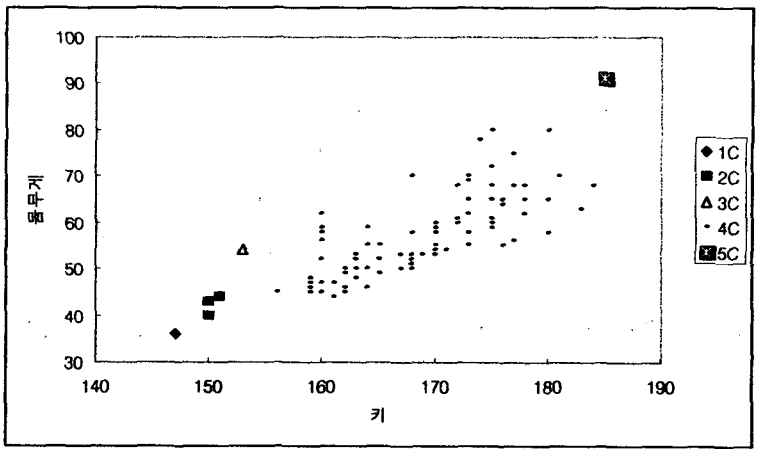
전형적인 클러스터링 기법은 <그림 2>처럼 데이터수가 1000개인 경우 수행 시간이 약 1800초로서, 실제로 방대한 양의 데이터로 클러스터링을 수행한다면 클러스터링 수행 자체가 거의 불가능하다 할 정도로 수행시간이 급격하게 증가할 것이다.

두 번째 실험은 클러스터링의 방법별 정확도를 비교하기 위해서 창원시에 소재하고 있는 모 학교의 학생 키와 몸무게의 실제 데이터 ($n=92$)를 이용하였다. 다음은 데이터의 전체 분포를 나타낸 산점도이다.



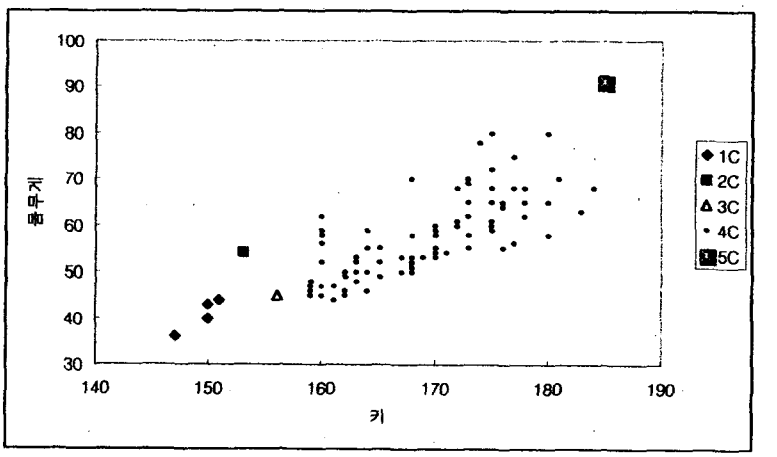
<그림 3> 전체 데이터 산포도

전형적인 클러스터링 기법에 의한 군집 결과는 다음과 같으며 5개의 군집으로 나눈 경우이다.



<그림 4> 전형적인 클러스터링 결과

다음 그래프는 그리드 기반 표본의 클러스터링 기법의 군집 결과이며 5개의 군집으로 나눈 경우이다.



<그림 5> 그리드 기반 표본의 클러스터링 결과

실험 결과, 그리드 기반 표본의 무게중심을 이용한 클러스터링 기법은 전형적인 클러스터링 기법을 기준으로 군집이 5개 일 때 100%, 군집이 10개 일 때 97.752%의 정확도를 보였으며, 무게중심을 이용하지 않은 그리드 기반 표본의 클러스터링 기법은 군집이 5개 일 때 97.8%, 군집이 10개 일 때 96.6%의 정확도를 나타내었다.

5. 결론

클러스터링은 방대한 데이터를 다루며 군집의 개수, 분포에 대해 아무런 가정을 하지 않는 원시적이며 탐색적인 방법이다. 이 기법은 수행 시간이 길어진다는 단점에도 불구하고 여러 분야에서 데이터 마이닝의 한 기법으로 많이 쓰이고 있는 기법이다.

본 연구에서는 그리드 기반으로 추출한 표본을 이용하여 클러스터링 알고리즘을 개발하였다. 실험 결과 전형적인 클러스터링 기법에 비해 그리드 기반 표본을 이용한 클러스터링 기법은 수행 시간을 단축시킬 뿐만 아니라 정확도 측면에서도 만족할 만한 수준의 결과를 나타낸다는 것을 확인하였다.

참고문헌

[1] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P.(1998), "Automatic subspace clustering of high dimensional data for data mining applications", In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), pp. 94-105, Seattle, WA.

[2] Bradley, P., Fayyad, U., and Reina, C.(1998), "Scaling clustering algorithms to large databases", In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp. 9-15, New York.

[3] Ester, M., Kriegel, H.P., and Xu, X.(1995), "Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification", In Proc. 4th Int. Symp. Large Spatial Databases (SSD'95), pp. 67-82, Portland, ME.

[4] Guha, S., Rastogi, R., and Shim, K.(1998), "CURE: An efficient clustering algorithm for large databases", In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD'98), pp. 73-84, Seattle, WA.

[5] Guha, S., Rastogi, R., and Shim, K.(1999), "Rock: A robust clustering algorithm for categorical attributes", In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), pp. 512-521, Sydney, Australia.

[6] Han J., Kamber M.(2001), "Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers.

[7] Hartigan, J.A.(1975), "Clustering Algorithms", John Wiley & Sons, New York.

[8] Huang, Z.(1998), "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, Vol. 2, pp. 283-304.

[9] Jain, A.K. and Dubes, R.C.(1988), "Algorithms for Clustering Data", Englewood Cliffs, NJ, Prentice Hall, New York.

[10] Jain, A.K., Murty, M.N., and Flynn, P.J.(1999), "Data clustering: A survey", ACM Comput. Surv., Vol. 31, pp. 264-323.

[11] Jardine, C.J. and Sibson, R.(1968), "The construction of hierarchic and nonhierarchic classifications", Comput. J., Vol. 11, pp. 177-184.

[12] Kaufman, L. and Rousseeuw, P.J., "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons, New York.

[13] Lauritzen, S.L. (1995), "The EM algorithm for graphical association models with missing data", Computational Statistics and Data Analysis, Vol. 19, pp. 191-201.

[14] MacQueen, J(1967), "Some methods for classification and analysis of multivariate observations", proc. 5th Berkeley Symp. Math. Statist, Prob., Vol.1, pp. 281-297.

[15] Ng, R and Han, J., "Efficient and effective clustering method for spatial data mining", In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pp. 144-155, Santiago, Chile.

[16] Schikuta, E.(1993), "Grid-Clustering: A Fast Hierarchical Clustering Method for Very Large Data Sets", Center for Research on Parallel Computation, Rice University.

[17] Sheikholeslami. G., Chatterjee, S., and Zhang, A.(1998), "WaveCluster: A multiresolution clustering approach for very large spatial databases", In Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98), pp. 428-439, New York.

[18] Wang, W., Yang, J., and Muntz, R.(1997), "STING: A statistical information grid approach to spatial data mining", In Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97), pp. 186-195, Athens, Greece.

[19] Zhang, T., Ramakrishnan, R., and Livny, M.(1996), "BIRCH: An efficient data clustering method for very large databases", In Proc. 1996 ACM-DIGMOD Int. Conf. Management of Data (SIGMOD'96), pp. 103-114, Montreal, Canada.