經營科學
第19卷 第1號
2002年 5月

데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬

황 인 수*

Two Phase Hierarchical Clustering Algorithm for Group Formation in Data Mining

Insoo Hwang*

Abstract

Data clustering is often one of the first steps in data mining analysis. It identifies groups of related objects that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation.

This paper purpose to present the development of two phase hierarchical clustering algorithm for group formation. Applications of the algorithm for product-customer group formation in customer relationship management are also discussed. As a result of computer simulations, suggested algorithm outperforms single link method and k-means clustering.

Keyword : Data Mining, Data Clustering, Crm, Algorithm

1. 서 론

데이터 마이닝(data mining)은 대량의 데이터로부터 패턴 인식, 통계적 기법, 인공지능 기법 등을 이용하여 데이터간의 상호 관련성, 패턴, 추세 등 의사결정에 유용한 정보를 추출해 내는 과정으로서 지식탐사의 핵심적인 역할을 담당한다(Bruce, 1996). 최근에는 효율적인 고객관리 및 마케팅전략

논문접수일 : 2001년 9월 17일 논문게재확정일 : 2002년 4월 17일
* 전주대학교 정보기술학부

을 수립하기 위해 고객을 세분화하거나 성향을 분석하여 그룹별 혹은 개인별로 차별화된 마케팅을 지원하는 고객관계관리(Customer Relationship Management, CRM) 등에서 데이터 마이닝이 광범위하게 사용되고 있다.

데이터 마이닝은 문제의 영역이나 그 목적에 따라서 다양한 방법들이 존재하는데, Cooley(1997)는 군집분석(clustering analysis), 분류규칙 발견(classification rule discovery), 연관규칙 발견(association rule discovery), 연속패턴 발견(sequence pattern discovery), 시각화(visualization) 등의 다섯 가지로 분류하였다.

본 연구에서 다루는 클러스터링은 하나의 데이터 집합을 서로 유사성을 갖는 몇 개의 클러스터로 분할해 나가는 과정으로서, 동일한 클러스터에 속하는 개체들간에는 상당한 유사성이 존재하지만, 클러스터간에는 이질성을 갖도록 클러스터를 구성한다(Michel, 1997). 클러스터링 기법들은 크게 계층적 알고리듬(hierarchical algorithm)과 최적화 알고리듬으로 구분할 수 있는데, 클러스터링이 NP-hard 문제이기 때문에 대부분의 현실문제에서는 계층적 알고리듬 및 이를 변형한 휴리스틱을 적용한다(Hartigan, 1974 ; Michael *et al*, 1979). 그러나, 계층적 알고리듬은 최적해를 보증하지 못하기 때문에 branch & bound 기법 등 최적해를 찾기 위한 알고리듬들이 제안되기는 하지만 계산의 복잡도로 인해 현실의 문제에서는 많이 사용되지 못하고 있다(Koontz, 1975).

계층적 클러스터링은 각각 한 개씩의 개체를 갖는 클러스터로부터 시작하여 적절한 개수의 클러스터가 만들어질 때까지 각 계층에서 가장 유사한 두 개의 클러스터를 병합하는 과정을 반복적으로 시행한다(Peter, 1988). 이와는 반대로, 모든 개체를 포함하는 한 개의 클러스터를 분할함으로써 적절한 개수의 클러스터를 만들어 가는 방법을 택하기도 한다. 따라서, 계층적 클러스터링에서는 각 개체가 특정 클러스터에 속하게 되면, 클러스터링이 진행되는 동안 보다 적합한 다른 클러스터가 존재하는 경우에도 다른 클러스터로 이동할 수 없기 때문에 바람직하지 못한 성과를 도출하는 단점이 있다.

이에 따라, 본 연구에서는 단순하면서도 전통적으로 가장 많이 사용되고 있는 계층적 클러스터링 기법을 변형한 2단계 클러스터링 알고리듬을 개발하였으며, 이에 대한 컴퓨터 시뮬레이션 결과를 제시한다. 이 알고리듬의 첫 번째 단계에서는 Single Linkage 알고리듬을 이용하여 1차 클러스터링을 수행하며, 두 번째 단계에서는 각 클러스터에 속한 개체들을 한 개씩 차례대로 제거하여 새로운 클러스터를 생성한 후 Average Linkage 알고리듬에 따라 병합하는 과정을 반복적으로 수행한다. 본 논문에서는 이 알고리듬의 효율성을 검증하기 위해 최근에 많은 관심의 대상이 되고 있는 CRM에서 서로 배타적인 제품-고객 그룹을 작성하는 문제에 대한 컴퓨터 시뮬레이션 결과를 함께 기술한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구에서 제안하는 2단계 계층적 클러스터링에 대해 기술하며, 제 3장에서는 이 알고리듬을 CRM의 제품-고객그룹의 구성문제에 적용하는 컴퓨터 시뮬레이션 결과에 대해 기술한다. 그리고, 제 4장에서는 본 연구의 결과에 대해 정리하면서 향후의 연구방향을 제시한다.

2. 2단계 계층적 클러스터링

본 논문에서는 계층적 클러스터링 기법을 기반으로 2단계로 분리하여 클러스터링을 수행하는 새로운 알고리듬을 제안한다. 알고리듬의 1단계에서는 각각 한 개씩의 개체로 구성된 클러스터를 구성한 후 Single Linkage(Sibson, 1973) 알고리듬에 따라서 인접한 두 개의 클러스터를 계층적으로 병합하는 1차 클러스터링을 수행한다. 이 때, 최종적으로 구성되는 클러스터의 개수는 클러스터링 목표함수에 따라서 자동적으로 결정되거나, 혹은 k-means 알고리듬에서와 같이 사전에 설정될 수 있다. 다음으로 2단계에서는 각 클러스터에 속한 개체들을 한 개씩 차례대로 제거하여 새로운 클러

스터를 만든 후 Average Linkage(Ellen, 1986) 알고리듬에 따라서 다시 병합하는 과정을 반복한다. <그림 1>은 목표로 하는 클러스터의 개수가 사전에 정해져 있는 문제에서 클러스터링을 수행하는 2단계 클러스터링 알고리듬을 자바 프로그래밍 언어의 문법에 따라 표현한 것이다.

```
public void clustering(int targetGroups) {
  int group, product, count=0 ;

  // 1st Phase
  initialize() ;  // 각 제품을 그룹으로 만듦
  for (group = numberOfProducts ; group> tergetGroups ;
      group--) {
      computeMinDistance() ;  // 그룹간 거리계산
      mergeMinGroups() ;  // 두 그룹을 병합
  }

  // 2nd Phase
  while (notFinished())
    for (group=0 ; group<targetGroups ; group++)
      while (moreProductIn(group)) {
        groupSplit(group)   // 그룹 분할
        computeAvgDistance() ;  // 그룹간 거리계산
        mergeMinGroups() ;  // 두 그룹을 병합
      }
}
```

〈그림 1〉 2단계 클러스터링 알고리듬

첫 번째 단계의 initialize() 메소드에서는 먼저 개별 제품을 각각의 클러스터로 구성하여 제품 개수만큼의 클러스터를 생성한다. 다음으로 computeMinDistance() 메소드에서는 Single Linkage 알고리듬에 따라서 각각의 클러스터에 속한 원소들간에 거리를 계산하여 가장 작은 값을 두 클러스터간의 거리로 설정한다. 여기서, 거리는 유클리디안 거리(Euclidean distance)를 사용한다. 끝으로, mergeMinGroup() 메소드에서는 거리가 가장 가까운 두 개의 클러스터를 병합하여 클러스터의 개수를 한 개 감소시킨다. 위의 과정을 반복적으로 수행하여 목표로 하는 개수의 클러스터가 생성되면 1단계를 종료한다.

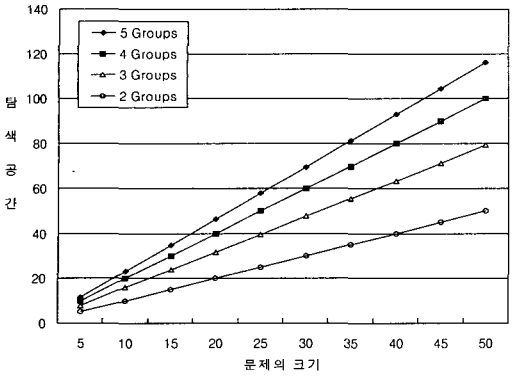
두 번째 단계의 groupSplit() 메소드에서는 각 클러스터에 배정되어 있는 제품을 한 개씩 차례대로 제거한 후 가장 적합한 클러스터에 재배정 한다. 이 때에는 Average Linkage 알고리듬에 따라서 각 클러스터의 중심점으로부터의 거리를 계산하여 가장 가까운 클러스터에 배정한다. 여기서 notFinished() 메소드는 2단계 클러스터링의 종료조건을 점검하는 것으로서, 클러스터에 속하는 제품들의 구성이 하나라도 변경된 경우에는 true를 리턴하여 클러스터링을 계속하며, 그렇지 않은 경우에는 false를 리턴하여 클러스터링을 종료한다.

3. 고객그룹 형성 문제에의 적용

3.1 고객관계관리 문제의 개요

본 연구에서 다루고자 하는 문제는 고객을 중심으로 마케팅 및 서비스 전략을 수행하는 고객관계관리(customer relationship management)의 한 유형으로서 (Kalakota & Robinson, 1999), 서로 구별되는 N_p 가지의 제품과 N_c명의 고객으로 이루어진 제품-고객 관계테이블을 이용하여 사전에 설정된 N_g개의 그룹을 구성하는 것이다. 이는 기업 조직내에 정해진 개수의 고객관리팀이 존재할 때 각 팀에 유사 제품 및 고객을 할당하는 의사결정에 유용하게 사용될 수 있을 것이다. 제품-고객 그룹의 형성문제는 조합(combinations) 이론에 의하면 상태공간의 크기가 $N_g^{N_p}$로 계산된다(Sheldon, 1988). <그림 2>는 제품의 종류가 5에서 50까지 증가할 때 구성하고자 하는 그룹의 수에 따른 탐색공간의 크기 변화를 $Log_2\ N$ 의 값으로 나타낸 그림이다.

따라서, 이 문제의 복잡도는 $O(N_g^{N_p})$로 나타나며, 이는 기업에서 판매하는 제품의 종류가 증가하거나 구성하고자 하는 그룹의 수가 증가하면 탐색공간(search space)의 크기가 기하급수적으로 증가하기 때문에 현실적으로 존재하는 큰 규모의 문제에서는 최적해를 발견하는 것이 거의 불가능함을 의미한다.



〈그림 2〉 제품의 종류 및 그룹에 따른 탐색공간의 크기

3.2 시뮬레이션

3.2.1 시뮬레이션 문제의 구성

본 연구에서 제시한 2단계 클러스터링 알고리듬의 효율성을 평가하는 시뮬레이션을 수행하기 위해서 문제생성 프로그램을 이용하여 다음의 조건에서 각각 100개씩의 문제세트를 생성하였다. 문제생성 프로그램에서 문제세트의 밀도는 6~12%, 그룹의 개수는 2~5개, 제품의 개수는 20~50개, 그리고 고객의 수는 300명을 가정하였다. 여기서, 밀도는 제품-고객 관계테이블에서 전체 셀 중에서 0이 아닌 값을 갖는 셀이 차지하는 비율을 의미하는 것으로, 각 고객이 많은 종류의 제품을 구매할수록 밀도는 높아진다. 또한, 예비 시뮬레이션을 실시한 결과 제품의 개수와 고객의 수는 시뮬레이션 소요시간에는 영향을 미치지만 최적해를 찾는 비율에는 큰 영향을 미치지 않는 것으로 나타남에 따라, 제품의 개수와 고객의 수는 임의로 설정한 것이다.

본 연구에서 제안하는 클러스터링 알고리듬의 성과를 평가하기 위해 구현한 모든 프로그램은 자바(Java) 프로그래밍 언어로 작성되었으며, 펜티엄 Ⅲ-1GHz Dual CPU의 윈도우즈 2000서버 운영체제에서 시뮬레이션을 실시하였다.

3.2.2 시뮬레이션 결과

<표 1>은 위와 같은 방법에 따라 생성된 문제세트에 대해 계층적 클러스터링의 Single Linkage 알고리듬으로 클러스터를 구성하여 최적해를 발견한 비율을 보여주고 있다. 표에서 보는 바와 같이 고객들이 여러 가지 종류의 제품을 구입할수록, 그리고 생성하고자 하는 그룹의 개수가 증가할수록 최적해를 찾는 비율은 높아지는 것으로 나타났다.

〈표 1〉 1단계에서 최적해를 발견한 비율(%)

밀도 \ 그룹	2	3	4	5	평 균
0.06	0.00	0.00	0.00	1.00	0.25
0.08	0.14	1.43	7.14	37.29	11.5
0.10	2.86	19.71	62.00	93.14	44.43
0.12	20.71	70.71	98.00	99.57	72.25
평균	5.92	22.96	41.79	57.75	32.12

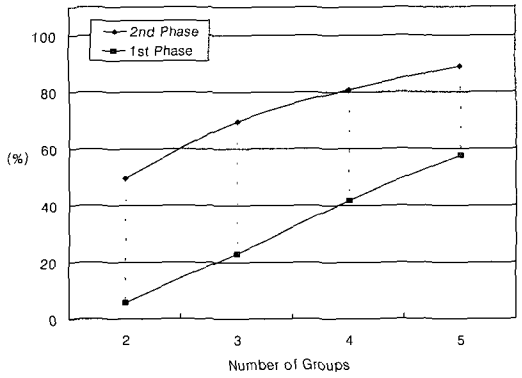
참고로, 본 연구에서는 시뮬레이션의 성과측정이 용이하도록 프로그램에서 자동적으로 g개의 그룹을 갖는 제품-고객관계 테이블을 임의로 생성하여 그룹핑 성과를 계산한 후, 이 테이블의 행과 열을 임의로 교환하여 만든 새로운 제품-고객관계 테이블에 클러스터링 알고리듬을 적용하여 시뮬레이션 하였다. 따라서, 각 문제에 대한 최적해를 사전에 알 수 있었다.

<표 2>는 1단계 클러스터링 결과에 대해 계속하여 2단계 클러스터링을 수행하였을 때 최적해를 발견한 비율을 보여주고 있는데, 1단계에 비해서 성과가 현저하게 향상되었음을 알 수 있다. 또한, <그림 3>은 생성하고자 하는 그룹의 개수에

〈표 2〉 2단계에서 최적해를 발견한 비율(%)

밀도 \ 그룹	2	3	4	5	평 균
0.06	1.75	12.75	29.75	57.50	25.43
0.08	25.50	67.50	92.50	98.50	71.00
0.10	74.50	98.50	100.00	100.00	93.25
0.12	96.25	99.75	100.00	100.00	99.00
평균	49.50	69.63	80.56	89.00	72.17

따라서 각 단계에서 최적해를 찾은 비율을 그림으로 보여준다.



〈그림 3〉 각 단계에서 최적해를 발견한 비율

3.2.3 성과척도의 개발

앞에서도 기술한 바와 같이, 현실적으로 존재하는 큰 규모의 문제에서는 최적해를 찾는다는 것이 사실상 불가능하기 때문에 제품-고객 그룹생성의 성과를 측정하기 위한 성과척도를 개발하여 클러스터링의 성과를 비교하는 것이 바람직하다. 이에 따라, 본 연구에서는 제조셀의 생성에 활용되는 그룹 테크놀로지(Kusiak, 1987 ; Chan, 1982)에서 클러스터 내에 포함되지 못한 원소들의 비율과 제품의 이용율(utilization)을 함께 고려하여 개발한 GE 척도(Chen, 1995)에 기반하여 다음과 같이 성과척도를 개발하였다.

먼저 GE 성과척도는 제품-고객군에 포함되지 못한 제품-고객 원소들의 개수뿐만 아니라 제품의 이용율까지 고려하여 만든 성과척도로서, 제품-고객군 내에 포함되는 원소들의 비율과 제품-고객군에서 제품이 선호되는 비율을 종합하여 평가한다.

$$\eta = \alpha\eta_1 + (1-\alpha)\eta_2$$

위에서 η_1은 그룹내에 포함된 원소의 비율, η_2는 그룹밖에 있는 원소에 관한 비율을 나타내며, α는 가중치를 나타내는 데 일반적으로 0.5로 설정한다(Chen, 1995). 따라서, η 척도의 값이 클수록 그룹핑의 성과는 높다고 볼 수 있다.

$$\eta_1 = \frac{E_i}{\sum_{i=1}^{N_g} p_i c_i}$$

$$\eta_2 = 1 - \left[\frac{E_o}{N_p N_c - \sum_{i=1}^{N_g} p_i c_i} \right]$$

여기서, E_i와 E_o는 각각 각 그룹에 포함된 원소의 개수와 벗어난 원소의 개수를 의미하며, N_p, N_c, N_g 는 각각 제품의 개수, 고객의 수, 그리고 그룹의 개수를 의미한다. 또한, p_i와 c_i는 각각 특정 그룹 i에 포함된 제품의 개수와 고객의 수이다.

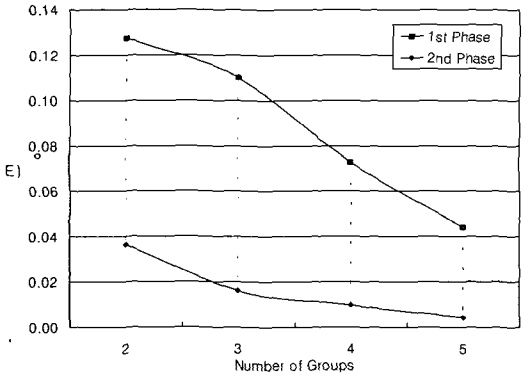
위에서 기술한 그룹테크놀로지의 성과척도는 성과척도의 값이 클수록 그룹핑이 잘되었음을 나타내지만, 최적해에서의 성과척도의 값을 알 수가 없기 때문에 최적해에 얼마나 가까운지를 평가하기가 어렵다. 따라서, 본 연구에서는 최적해에서 성과척도의 값이 0이 되도록 GE 성과척도를 변형하였으며, 각 그룹별로 성과를 계산한 후에 이를 그룹의 크기에 따라 가중치를 주어서 전체적인 성과척도를 계산하도록 하였다.

$$E = \sum_{i=1}^{N_g} \left[\left(\alpha \frac{e_i}{p_i c_i} + (1-\alpha)\frac{e_i}{t_i} \right) \cdot \frac{p_i}{N_p} \right]$$

여기서, α는 두 가지 성과척도에 대한 가중치로서 본 연구에서는 0.5로 설정하였으며, t_i는 그룹 i에 속한 고객들이 구매했거나 혹은 관심을 갖고 있는 제품들의 개수, e_i는 이 중에서 그룹을 벗어난 제품의 개수이다. 또한, p_i와 c_i는 각각 그룹 i에 속한 제품과 고객의 수이며, N_p와 N_g는 각각 제품의 전체 개수와 그룹의 개수를 의미한다.

<그림 4>는 위에서 기술한 성과척도를 이용하여 측정한 각 단계에서의 성과의 평균값을 그림으

로 보여주고 있다. 그림에서 보는 바와 같이 Single Linkage를 사용하는 1단계에 Average Linkage 기법을 이용하여 2단계 클러스터링을 수행할 경우 위에서 제시한 성과척도 값이 훨씬 더 낮아졌다. 따라서, 2단계 클러스터링을 수행할 경우 클러스터링의 성과가 현저히 향상됨을 알 수 있다.

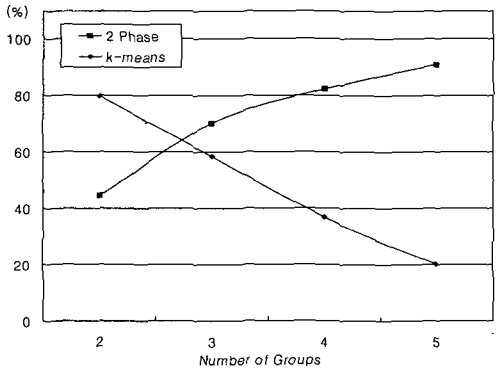


〈그림 4〉 각 단계의 그룹핑 성과 비교

3.3 k-means 알고리듬과의 비교

본 연구에서 제시한 2단계 계층적 클러스터링 알고리듬은 각 단계에서 계층적 클러스터링을 이용하기 때문에 사전에 클러스터의 개수가 정해지지 않은 경우에도 성과척도의 증감에 따라 클러스터의 개수를 자동적으로 결정하는 장점을 갖고 있다. 그러나, 위에서 예로 든 고객그룹의 형성과 같이 사전적으로 클러스터의 개수가 정해져 있는 경우에는 일반적으로 k-means 알고리듬(Forgery, 1965 ; MacQueen, 1967)을 많이 사용하고 있다. k-means 알고리듬에서는 먼저 임의로 k개의 개체를 선택하여 각각을 클러스터로 설정한 후 나머지 개체들을 가장 근접한 클러스터에 할당한다. 다음으로, 각 클러스터의 중심점을 다시 계산한 후에 모든 개체를 가장 근접한 클러스터에 다시 할당하는 과정을 반복한다. 이에 반하여, 본 연구에서는 1단계에서 계층적 클러스터링 기법의 Single Linkage 기법에 따라 초기해를 구한 후에, 2단계에서 각 개체가 새로운 클러스터에 할당될 때마다 중심점을 다시 계산하도록 함으로써 최적해에 빠르게 접근해 가도록 하였다.

<그림 5>는 생성하고자 하는 그룹의 개수에 따라서 각 알고리듬이 최적해를 찾는 비율을 비교하여 나타낸 것으로 그룹의 개수가 증가할수록 k-means 알고리듬의 성과는 낮아지지만 본 연구에서 제시하는 2단계 알고리듬의 성과는 향상되는 것으로 나타났다. 또한, 그림에서도 볼 수 있는 바와 같이 세 개 이상의 그룹을 생성할 경우 2단계 알고리듬이 훨씬 더 나은 성과를 제시하는 것을 볼 수 있다.



〈그림 5〉 그룹의 개수에 따른 알고리듬간 성과 비교

여기서, 그룹의 개수가 증가함에 따라 두 가지 알고리듬의 성과가 반대방향으로 움직이고 있는데, 이것은 각 알고리듬이 갖고 있는 기본적인 특성에 기인하는 것으로 판단된다. 즉, k-means 알고리듬에서는 그룹의 개수만큼의 중심점을 임의로 설정한 후에 중심점의 이동에 따라 그룹핑이 변경되므로, 그룹의 개수가 많아질수록 각 그룹의 중심점이 최적으로 이동하는 것이 어려워지기 때문에 클러스터링의 성과는 낮아지는 것으로 분석된다. 그러나, 본 연구에서 적용한 계층적 클러스터링에서는 1단계에서 보다 많은 개수의 그룹으로부터 작은 개수의 그룹으로 병합한 후 2단계에서 조정

하는 과정을 반복적으로 수행하기 때문에 그룹의 개수가 많아질수록 클러스터링의 성과는 향상되는 것으로 분석된다. 이것이 두 알고리듬의 성과는 반대 방향으로 움직이게 하는 원인으로 판단된다.

4. 결론 및 향후 연구계획

본 논문에서는 대표적인 데이터 마이닝 기법의 하나인 클러스터링에 있어서 전통적으로 많이 사용되고 있는 계층적 클러스터링 기법을 변형한 2단계 클러스터링 알고리듬을 제시하였다. 이 알고리듬의 첫 번째 단계에서는 Single Linkage 알고리듬을 이용하여 클러스터링을 수행하며, 두 번째 단계에서는 각 클러스터에 속한 개체들을 한 개씩 차례대로 분리하여 새로운 클러스터를 만든 후 Average Linkage 알고리듬을 이용하여 병합하는 과정을 반복적으로 수행한다.

본 논문에서는 이 알고리듬의 효율성에 대한 성과측정을 위해 최근에 많은 관심의 대상이 되고 있는 고객관계관리에서 서로 배타적인 제품-고객 그룹을 작성하는 문제에 대한 컴퓨터 시뮬레이션 결과를 함께 기술하였다. 시뮬레이션 결과, 1단계 클러스터링에 비해 성과가 현저히 증가하는 것으로 나타났으며, 클러스터의 개수가 사전에 결정되어 있는 문제에 많이 사용되고 있는 k-means 알고리듬에 비해서도 훨씬 더 나은 성과를 나타냈다. 본 연구에서는 2단계 계층적 알고리듬의 성과를 k-means 알고리듬의 성과와 비교하였는데, 최근에 개발된 보다 많은 알고리듬과의 성과를 비교하지 못한 것을 한계점으로 들 수 있겠다. 따라서, 후속 연구에서는 알고리듬을 발전시킬 뿐만 아니라 기존에 개발된 다른 많은 알고리듬과의 성과 비교를 추가하고자 한다.

또한, 향후 연구에서는 제품-고객 관계테이블의 밀도, 문제의 규모, 클러스터의 개수 등에 따른 성과차이를 보다 심층적으로 분석하여 현실의 문제에 위의 알고리듬을 적용하는 연구를 수행할 계획이며, 이와 함께 1단계의 계층적 클러스터링 뿐만 아니라 2단계에서도 성과척도에 따라서 클러스터의 개수를 증감시켜서 최적의 클러스터를 생성하는 방안에 대한 연구를 계속하고자 한다.

참 고 문 헌

[1] Bruce M., "Defining Data Mining, The Hows and Whys of Data Mining, and How It differs From Other Analytical Techniques," *Online Addition of DBMS Data Warehouse Supplement*, August, 1996.

[2] Chan, H. and D. Milner, "Direct clustering algorithm for group formation in cellular manufacturing," *Journal of Manufacturing Systems*, Vol.1(1982), pp.65-75.

[3] Chen, S. and C. Cheng, "A neural network-based cell formation algorithm in cellular manufacturing," *International Journal of Production Research*, Vol.33, No.2(1995), pp. 293-318.

[4] Cooley R., B. Mobasher, and J. Srivastava, "Web Mining : Information Pattern Discovery on the World Wide Web," *Proc. of the 9th IEEE International Conference*, 1997, pp.558-567.

[5] Ellen M., "Implementing Agglomerative Hierarchic Clustering Algorithm for Use in Document Retrieval," *Information Processing & Management*, Vol.22, No.6(1986), pp. 465-476.

[6] Forgery, E., "Cluster Analysis of Multivariate Data : Efficieny vs. Interpretability of Classifications," *Biometrics*, Vol.21(1965), pp.768.

[7] Hartigan J., *Clustering Algorithms*, John Wiley & Sons, New York, 1974.

[8] Kalakota, R. and M. Robinson, *e-business : Roadmap for Success*, Addison Wesley, 1999.

[9] Koontz, W., P. Narendra and K. Fukunaga, "A branch and bound clustering algorithm,"

*IEEE Transactions on Computer*s, Vol. C-24(1975), pp.908-915.

[10] Kusiak, A. and W. Chow, "Efficient Solving of The Group Technology Problem," *Journal of Manufacturing Systems*, Vol.6, No.2 (1987), pp.117-124.

[11] MacQueen, K. "Some Methods for Classification and Analysis of Multivariate Observation," *Proc. of the 5th Berkeley Symposium on Mathematical Statics and Probability*, 1967, pp.281-297.

[12] Michael R. and S. Johnson, *Computers and Intractability : A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, 1979.

[13] Michel J., A. Berry, and Gorden Linoff, *Data Mining Techniques : For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc., 1997.

[14] Peter, W., "Recent Trends in Hierarchical Document Clustering : A Critical Review," *Information Processing & Management*, Vol. 24, No.5(1988), pp.577-597.

[15] Sheldon, R., *A First Course in Probability*, 3rd Eds., Maxwell Macmillan, 1988.

[16] Sibson, R., "Slink : An Optimal Efficient Algorithm for a Complete Link Method," *The Computer Journal*, Vol.16(1973), pp.30-34.