신경망 및 통계적 방법에 의한 클러스터링 성능평가 - A Study on Performance Evaluation of Clustering Algorithms using Neural and Statistical Method -

윤 석 환*
Yoon, Seok-Hwan
민 준 영**
Min, Joon-Young
신 용 백***
Shin, Yong-Back

ABSTRACT

This paper evaluates the clustering performance of a neural network and a statistical method. Algorithms which are used in this paper are the GLVQ(Generalized Learning vector Quantization) for a neural method and the k-means algorithm for a statistical clustering method. For comparison of two methods, we calculate the Rand's c statistics. As a result, the mean of c value obtained with the GLVQ is higher than that obtained with the k-means algorithm, while standard deviation of c value is lower. Experimental data sets were the Fisher's IRIS data and patterns extracted from handwritten numerals.

1. 서 론

클러스터링 알고리즘은 통계적인 방법, ISODATA알고리즘, 신경망 클러스터링으로 크게 구분할 수 있다[1]. 신경망을 이용한 클러스터링 방법중 Kohonen네트워크는 패턴과 클러스터의 중심값과의 거리를 최소화시키는 학습 알고리즘에 따라 클러스터링을 한다는 의미에서 통계적 방법중 k-means방법과 대응된다고 할 수 있다. 그리고 최근에 발표된 Pal et. al.의 GLVQ(Generalized Learning Vector Quantization) 알고리즘이 있다. 이 방법은 Kohonen의 LVQ 알고리즘에서 초기 연결강도의 설정에 따라 클러스터의 중심값이 크게 영향을 받아 오분류(misclassification)되는 패턴이 발생하는 경우가 많다는 단점을 개선시킨 방법으로 모든 출력 노드에 대해서 학습을 함으로써 학습률, 학습의 반복횟수, 초기 연결강도값에 크게 영향을 받지 않고 일정한 클러스터 중심값이 계산되기 때문에 각 클러스터별로 오분류되는 패턴이 적게 나타난다는 장점이 있다.

클러스터링의 성능을 평가하는 연구로는 각 패턴과 클러스터의 중심값에 대한 제곱오차를 계산하는 방법으로는 Ismail과 Kamel이 Best-first(ABF), First-best(AFB)알고리즘의 성능을 평가할 때 클러스터의 수를 변화시켜 가면서 k-means알고리즘과 제곱오차(sum of squared

* 한국전자통신연구소 선임연구원
** 상지대학교 병설 전문대학 교수
*** 아주대학교 산업공학과 교수

error)값을 비교하였고, Dubes는 CPU처리시간, 메모리 사용용량, 오분류 되는 패턴의 수와 패턴과 클러스터의 중심값과의 제곱오차값의 네 가지 기준을 설정하여 FORGY, WISH등 8가지의 클러스터링 프로그램을 비교하였다. 또한 Rand는 클러스터링 방법의 평가기준으로써 "Natural" 클러스터와의 비교, 원시 자료에 잡음(noise)을 추가시켜 구성된 클러스터와의 비교, 그리고 자료에 결측(missing)을 발생시켜 구성한 클러스터와 비교하는 방법을 제안하였다.

본 논문은 하드(Hard or Crisp) 클러스터링으로 국한하였을 때, 신경망을 이용한 클러스터링 알고리즘의 GLVQ와 통계적 클러스터링 알고리즘의 k-means를 비교하였으며, 비교기준으로는 Rand C 통계량을 이용하였다. 본 논문에서 이용한 데이타는 Fisher의 IRIS데이타와 대학교 재학중인 100명의 학생을 무작위로 선정하여 0-9의 숫자 중 하나씩 쓰게 한 100개의 필기체 숫자를 스캐너로 이미지를 입력받아 이 이미지 자료를 16×16의 이진자료로 변환시킨 다음 이 자료를 4×4로 특징추출한 16개변수로 구성된 패턴벡터를 클러스터링방법의 다차원 자료로 이용하였다. 비교 분석의 편의상 필기체는 각 숫자마다 10명의 학생을 할당하여 쓰게 하였다.

2. 클러스터링 방법

2.1 신경망에 의한 클러스터링 : Generalized 학습 벡터 양자화(GLVQ)

LVQ알고리즘은 초기 연결강도 $\mathbf{v}_{i,0}$에 의해서 클러스터링 결과가 많은 영향을 받기 때문에 Pal et. al.[9]은 학습을 승자노드 뿐 만 아니라 승자가 아닌 노드도 같이 학습을 시키는 GLVQ 클러스터링 알고리즘을 제안하였다. GLVQ네트워크에서의 학습규칙은 입력 패턴 $\mathbf{x}$와 출력노드간의 거리에 대한 가중값을 준 손실함수(loss function) L_x를 식 (2.1)와 같이 정의하고 이 손실함수를 최소화시키는 방법으로 학습규칙을 유도하였다.

$$L_x = \sum_{r=1}^{c} g_{ir} \| \mathbf{x} - \mathbf{v}_r \|^2$$

$$\text{여기서,} \quad g_{ir} = \begin{cases} 1 & \text{, 승자노드일 경우} \\ \dfrac{1}{\sum_{j=1}^{c} \| \mathbf{x} - \mathbf{v}_j \|^2} & \text{, 승자노드가 아닌 경우} \end{cases} \qquad (2.1)$$

로써 i번째 패턴에 대한 가중값.

여기서 승자노드일 경우에는 L_x를 승자노드와 연결된 연결강도 $\mathbf{v}_i$로 미분한 $\Delta_{\mathbf{v}_i} L_x$를 계산하였고, 승자노드가 아닐 경우에는 승자노드가 아닌 노드에 연결된 연결강도 $\mathbf{v}_j$로 미분한 $\Delta_{\mathbf{v}_j} L_x$를 계산함으로써 식 (2.2)와 식 (2.3)으로 유도하였으며, 이 학습규칙에 의하여 학습을 한다.

$$\mathbf{v}_{i,t} = \mathbf{v}_{i,t-1} + \alpha_t (\mathbf{x}_k - \mathbf{v}_{i,t-1}) \frac{D^2 - D + \| \mathbf{x}_k - \mathbf{v}_{i,t-1} \|^2}{D^2}$$

승자노드일 경우의 학습 규칙 (2.2)

$$\mathbf{v}_{r,t} = \mathbf{v}_{r,t-1} + \alpha_t (\mathbf{x}_k - \mathbf{v}_{r,t-1}) \frac{\| \mathbf{x}_k - \mathbf{v}_{i,t-1} \|^2}{D^2}$$

승자노드가 아닐 경우의 학습 규칙 (2.3)

여기서, $D = \sum_{r=1}^{c} \| \mathbf{x} - \mathbf{v}_r \|^2$, $1 \le k \le N$, $1 \le r \le c$.

$\mathbf{v}_i$는 승자노드와 연결된 연결강도 벡터

따라서 GLVQ의 학습알고리즘은 다음과 같다.

1. 승자노드의 발견

$$\| \mathbf{x}_k - \mathbf{v}_{i,t-1} \| = \min_{1 \le j \le c} \{ \| \mathbf{x}_k - \mathbf{v}_{j,t-1} \| \}$$

2. 연결강도의 수정
   승자노드일 경우에는 식 (2.2)에 의하여 연결강도 조정.
   승자노드가 아닐 경우에는 식 (2.3)에 의하여 연결강도 수정.
3. 반복에 따른 클러스터 중심값의 오차계산

$$E_t = \| \mathbf{V}_t - \mathbf{V}_{t-1} \| = \sum_{r=1}^{c} \| \mathbf{v}_{r,t} - \mathbf{v}_{r,t-1} \|$$

4. 만약 (E_t < 허용오차) 이면 끝, 아니면 1번 수행.

GLVQ의 특징으로는 연결강도의 초기값에 영향을 받지 않고 클러스터링을 할 때 오분류되는 패턴의 수가 적어지며, 자료의 구분이 확실한 경우에 학습이 완료된 후 각 클러스터의 중심값과 실제 중심값과의 오차도 매우 근소하게 나타난다는 점에서 LVQ보다는 더 좋은 알고리즘으로 평가되고 있다. 예를 들어, 초기 연결강도의 값이 모든 패턴이 포함되는 점에서 많이 벗어나서 설정되었다고 가정하여 보자. 네트워크에 $\mathbf{x}_i$가 입력되었을 때 c_i가 승자노드이면 LVQ인 경우에는 c_i에 연결된 연결강도 $\mathbf{v}_i$에 대해서만 학습을 하고, 다른 패턴 $\mathbf{x}_j$ (j≠i)가 입력되어도 역시 승자는 $\mathbf{x}_i$ 패턴의 승자였던 c_i가 된다. 왜냐하면 다른 노드들과 연결된 연결강도는 초기에 모든 패턴과 멀리 떨어진 값으로 주어졌기 때문에 이미 승자가 될 수 없기 때문이다. 이는 다음 반복이 계속되어도 $\mathbf{v}_i$에 대해서만 학습이 이루어지게 되고 반복이 완료될 때까지 조정된 연결강도는 $\mathbf{v}_i$에 국한된다. 따라서 다른 노드들은 반복이 완료될 때까지 초기에 주어진 값을 그대로 갖고 있게 되는 경우가 발생하기 때문에 모든 패턴은 c_i클러스터에 모두 포함되는 경우가 발생한다. 그러나 GLVQ는 $\mathbf{x}_i$의 승자노드에 연결된 연결강도 $\mathbf{v}_i$ 뿐만 아니라 승자가 아닌 다른 노드의 연결강도까지도 조정을 해 주기 때문에 반복이 계속되면서 $\mathbf{v}_i$가 아닌 다른 노드들도 승자가 될 수 있고, 반복이 완료된 후의 모든 연결강도는 입력된 패턴을 각 클러스터별로 포함하는 중심값이 될 수 있다.

실제로 Pal et. al.은 GLVQ를 제안한 논문에서 Fisher의 IRIS데이타를 가지고 반복횟수와 학습률을 변화시켜 가면서 LVQ와 GLVQ로 클러스터링 한 결과를 비교하였는데 150개의 IRIS데이타중 GLVQ는 오분류된 패턴이 17개로 일정한 반면에 LVQ인 경우에는 100개에서 17개의 오분류 패턴을 나타낸 결과를 제시하였다.

2.2 통계적 클러스터링 : k-means 방법

k-means방법은 각 패턴과 클러스터의 중심값과의 거리 차이를 최소화시키는 방법으로 클러스터링을 하는 것으로, 계층적 클러스터링에서 초기에 병합이 부적절했을 때 패턴을 다시 재배열할 수 없다는 단점이 있는 반면에 이 방법은 반복이 이루어지면서 각 클러스터에 대한 패턴의 재배열이 가능하도록 한 방법이다. 우선 패턴을 k개의 클러스터로 나눈 후 클러스터에 포함되어 있는 패턴들의 평균으로 클러스터의 중심값을 계산하고 이 중심값과 각 패턴과의 거리를 계산한 후 가장 거리가 가까운 클러스터에 패턴을 포함시키는 방법으로 그 조건은 식 (2.4)와 같다.

$$\mathbf{x}_i \in c_j \quad \text{iff} \quad \| \mathbf{x}_i - \mathbf{z}_j \|^2 < \| \mathbf{x}_i - \mathbf{z}_k \|^2$$

여기서, $1 \le i \le N$, $1 \le k \le c$, $j \neq k$

N : 패턴 수

c : 클러스터 수

z : 클러스터 중심값 (2.4)

그리고 이 계산은 각 클러스터의 중심값이 더 이상 변하지 않을 때까지 반복한다. 초기 k개의 클러스터의 중심값을 주는 방법에는 주어진 패턴에서 처음 k개의 패턴을 추출하여 중심값으로 하는 방법과 임의로 k개를 추출하여 중심값으로 하는 방법이 있는 데 본 논문에서는 임의로 k개를 추출하여 클러스터의 초기 중심값으로 하였다.

3. 성능평가 기준

3.1 평가 기준

클러스터링의 결과를 비교하는 방법으로 N개의 패턴 $\mathbf{X} = \{ \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$이 있다고 가정한다.

이 패턴을 Y라는 방법으로 클러스터링 한 결과를 $\mathbf{Y}=\{ \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{k1}\}$라 하고, Y′이라고 하는 방법으로 클러스터링 한 결과를 $\mathbf{Y}'=\{\mathbf{y}'_1, \mathbf{y}'_2 ,..., \mathbf{y}'_{k2}\}$ 라고 했을 때 Y와 Y′에 있는 패턴 중 두개의 패턴 $\mathbf{x}_i$, $\mathbf{x}_j$ $(1 \le i \le N , 1 \le j \le N , i \neq j)$을 추출한 다음 이 두개의 패턴이 Y와 Y′에서 모두 한 클러스터 안에 속해 있을 경우에는 Y와 Y′은 $\mathbf{x}_i$와 $\mathbf{x}_j$를 유사한 패턴으로 간주한다. 또한 서로 다른 클러스터에 속해 있을 경우에는 Y와 Y′은 $\mathbf{x}_i$와 $\mathbf{x}_j$를 서로 유사하지 않은 것으로 간주하기 때문에 이 두 가지 경우가 Y와 Y′의 유사성을 결정하는 척도가 된다. 그러나 $\mathbf{x}_i$, $\mathbf{x}_j$가 Y에서는 같은 클러스터에 속해 있어서 서로 유사하다는 결과로 나왔으나 Y′에서는 서로 다른 클러스터에 분리되어 속해 있어서 서로 유사하지 않다는 결과가 나왔거나 아니면 이와 반대의 결과가 나왔다면 Y와 Y′의 클러스터링 방법에는 차이가 있다고 할 수 있다. 따라서 N개의 패턴중 두개의 패턴을 추출한 경우의 수를 모두 비교한다면 Y와 Y′의 클러스터링의 유사성을 비교할 수가 있다. 즉, N개의 패턴에서 두개의 패턴 $\mathbf{x}_i$와 $\mathbf{x}_j$를 추출하는 경우의 수는 ${}_NC_2$이고, $\mathbf{x}_i$, $\mathbf{x}_j$가 서로 같은 클러스터에 있는 경우의 수와 서로 다른 클러스터에 분리되어 포함된 경우의 수를 더한 값의 비를 계산한다면 Y와 Y′의 유사성을 비교할 수가 있다. 이를 식으로 표현하면 식 (3.1)과 같다.

$$c(Y, Y') = \frac{\left[\binom{N}{2} - \left[\frac{1}{2}\{\sum_i(\sum_j n_{ij})^2 + \sum_j(\sum_i n_{ij})^2\} - \sum_i\sum_j n_{ij}^2\right]\right]}{\binom{N}{2}} \quad (3.1)$$

여기서, n_{ij}는 Y와 Y′으로 구성한 클러스터링 결과 중 Y의 i번째 클러스터와 Y′의 j번째 클러스터 안에 모두 포함되어 있는 동일한 패턴의 수를 의미한다. 식 (3.1)에서 C=0이면 Y와 Y′의 결과는 전혀 유사하지 않은 것이며, C=1일 경우에는 두개의 클러스터의 결과가 동일한 것임을 나타낸다. 위 식에서 표현된 c 통계량의 특징으로는 첫째, C 통계량은 두 클러스터의 유사성을 나타내고, 둘째, (1-C)는 두 클러스터의 비유사성을 의미하며, 셋째, X가 어느 특정 분포를 갖고 있을 경우에 C는 확률변수이다.

3.2 데이타 선정

3.2.1 데이타

본 논문에서 이용한 자료는 크게 두 가지로 나눌 수가 있다. 첫째, Fisher의 IRIS데이타를 이용하였는 데 이 자료는 Pal et. al.[9]이 신경망 클러스터링인 LVQ와 GLVQ클러스터링 알고리즘의 성능을 비교하는 데 이용하였다. 둘째, 대학교 재학중인 100명의 학생을 무작위로 추출하여 0-9의 숫자 중 하나씩 쓰게 하여 수집된 100개의 패턴을 스캐너로 그 이미지를 입력받아서 특징추출(feature extraction)을 하였고, 이 특징추출한 자료를 이용하여 클러스터링을 하였다. 비교분석의 편의상 0-9까지의 숫자를 쓸 때 각 숫자마다 학생 10명씩을 할당하여 쓰게 하였다.

3.2.2 특징추출

16×16 격자에 숫자에 대하여 이진화(binarized)된 원시 자료를 받고, 이 숫자를 좌,우,상,하로 여백을 제거한 다음 이 패턴을 10×10으로 스케일링을 한다. 그 다음 좌,우,상,하에 한 행 또는 열을 추가시켜 12×12로 만든다. 이 12×12로 스케일링한 자료를 좌측 상단에 있는 격자부터 한 격자씩 탐색해 가면서 '1'이 있을 경우에 그 격자를 중심으로 좌, 우, 상, 하, 좌상, 좌하, 우상,우하의 방향에 '1'이 있으면 1을 증가시키는 방법이다. 따라서 모든 방향에 '1'이 있을 경우에는 최고 9의 값을 갖는다. 12×12의 격자가 모두 탐색되고 나면 <그림 3.1c>와 같은 자료를 얻을 수 있다. 이 자료를 가로,세로 각각 3개씩의 격자를 묶어서 하나의 블럭으로 만들고, 이 블럭 안의 값을 모두 더한 다음에 10으로 나누어 준 자료가 입력패턴벡터가 된다.

<그림 3.1>는 100개의 패턴 중 숫자 '0'에 대한 특징추출 과정에서 얻은 자료의 예를 보여주고 있다.

```
.....1111.
....11..1.
...11...11
..11.....1
.11......1
.1......11
11......11
1......11.
1......11.
1.....11..
11..111...
.11111....
```

(a) 원시자료

```
............
......1111..
.....11..1..
...111...11.
..11.....1.
..1......11.
.11......11.
.1......11..
.1.....111..
.11..111....
..11111.....
............
```

(b) 12×12 스케일링

0	0	0	0	0	1	2	3	3	2	1	0
0	0	0	0	1	3	4	4	4	3	2	0
0	0	1	2	4	5	5	4	5	5	4	1
0	1	3	4	5	4	3	1	2	4	4	2
0	2	4	5	4	2	1	0	2	5	5	3
1	4	5	4	1	0	0	0	2	5	5	3
2	4	4	2	0	0	0	1	4	6	5	2
3	4	4	1	0	0	1	3	6	6	4	1
3	4	4	1	1	2	4	5	6	4	2	0
2	4	5	4	4	5	6	5	4	2	1	0
1	3	4	4	4	5	5	3	1	0	0	0
0	1	2	3	3	3	2	1	0	0	0	0

(c) 12×12 누적 자료

0.1	1.6	3.4	1.8
2.0	2.9	1.1	3.6
3.2	0.7	3.0	3.0
2.2	3.5	2.7	0.3

(d) 4×4 입력자료

<그림 3.1> 16×16 자료를 4×4로 특징추출 예 : 숫자 '0'

4. 성능평가

이 장에서는 Rand 비교 방법에 의해서 신경망 클러스터링과 통계적방법의클러스터링의 성능을 평가한다. Rand 방법에 의한 클러스터링의 성능 평가는 최적의 클러스터 개수를 결정하는 것은 배제되었기 때문에 본 논문에서 이용한 자료 중 IRIS데이타는 클러스터의 수를 3개로 하고, 숫자자료는 클러스터의 수를 10개로 하여 비교하였다.

4.1 "Natural" 클러스터와의 비교

본 절에서는 "Natural"클러스터를 알고 있다는 가정 하에 비교하는 방법이기 때문에 IRIS데이타와 0-9의 숫자 패턴 자료를 이용하여 앞절에서의 C 통계량을 계산하였다.

IRIS데이타에서 "Natural"클러스터를 비교하기 위해서 Y는 세 종류의 붓꽃중 setosa를 0-49까지, versicolor를 50-99까지, verginica를 100-149까지로 하여 세개의 클러스터의 집합으로 하였고, Y'은 GLVQ, k-means방법으로 구성한 클러스터의 집합이라고 하였다.

$$\mathbf{Y}_{IRIS} = \{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{49}), (\mathbf{x}_{50}, \mathbf{x}_{51}, \ldots, \mathbf{x}_{99}), (\mathbf{x}_{100}, \mathbf{x}_{101}, \ldots, \mathbf{x}_{149})\}$$

여기서, $\mathbf{x}_i \in R^4$,
iris setosa $(0 \le i \le 49)$,
iris versicolor $(50 \le i \le 99)$,
iris virginica $(100 \le i \le 149)$

$$\mathbf{Y}'_{IRIS} = \{(\mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \ldots, \mathbf{x}'_{1,n_1}), (\mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \ldots, \mathbf{x}'_{2,n_2}), (\mathbf{x}'_{3,1}, \mathbf{x}'_{3,2}, \ldots, \mathbf{x}'_{3,n_3})\}$$

여기서, $\sum_{i=1}^{3} n_i = 150$,

$\mathbf{x}'_{i,j}$: IRIS데이타를 어느 특정 클러스터링 방법으로 클러스터를 구성하였을 때 i번째 클러스터에 포함된 j번째 패턴 벡터.

$Y'_{IRIS,GLVQ}$와 $Y'_{IRIS,k\text{-}means}$를 각각 GLVQ와 k-means방법으로 3개의 클러스터를 구성한 집합이라고 했을 때, $c(Y_{IRIS}, Y'_{IRIS,GLVQ})$, $c(Y_{IRIS}, Y'_{IRIS,k\text{-}means})$, 를 계산하여 그 유사성을 비교하였다.

0-9까지의 숫자 자료를 비교하는 데 있어서는 실제 Y를 자료의 성질에 따라 구분되는 기준 클러스터로 총 100개의 패턴을 각 숫자별로 구분하여 구성한 클러스터이다. 본 논문에서는 분석의 편의상 각 숫자별로 10명의 필기자를 할당하여 쓰게 하였으므로 10개의 숫자 '0'은 클러스터 0에, 10개의 숫자 '1'은 클러스터 1에 포함시키고, 마지막으로 숫자 '9'는 클러스터 9에 포함시킨 10개의 클러스터 집합이다.

$$Y_{NUM} = \{(\mathbf{x}_{0,0}, \mathbf{x}_{1,0}, \mathbf{x}_{2,0}, \ldots, \mathbf{x}_{9,0}), (\mathbf{x}_{0,1}, \mathbf{x}_{1,1}, \mathbf{x}_{2,1}, \ldots, \mathbf{x}_{9,1}), \ldots, \ldots, (\mathbf{x}_{0,9}, \mathbf{x}_{1,9}, \mathbf{x}_{2,9}, \ldots, \mathbf{x}_{9,9})\}$$

여기서, $\mathbf{x}_{i,j} \in R^p$: 각 숫자별 i번째$(0 \le i \le 9)$ 필기자가 쓴 숫자 j.

$$Y'_{NUM} = \{(\mathbf{x}'_{0,1}, \mathbf{x}'_{0,2}, \ldots, \mathbf{x}'_{0,n_1}), (\mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \ldots, \mathbf{x}'_{1,n_2}), \ldots, \ldots, (\mathbf{x}'_{9,1}, \mathbf{x}'_{9,2}, \ldots, \mathbf{x}'_{9,n_9})\}$$

여기서, $\sum_{i=0}^{9} n_i = 100$,

$\mathbf{x}'_{i,j}$: 0-9 숫자자료를 어느 특정 클러스터링 방법으로 클러스터를 구성하였을 때 i번째 클러스터에 포함된 j번째 패턴 벡터.

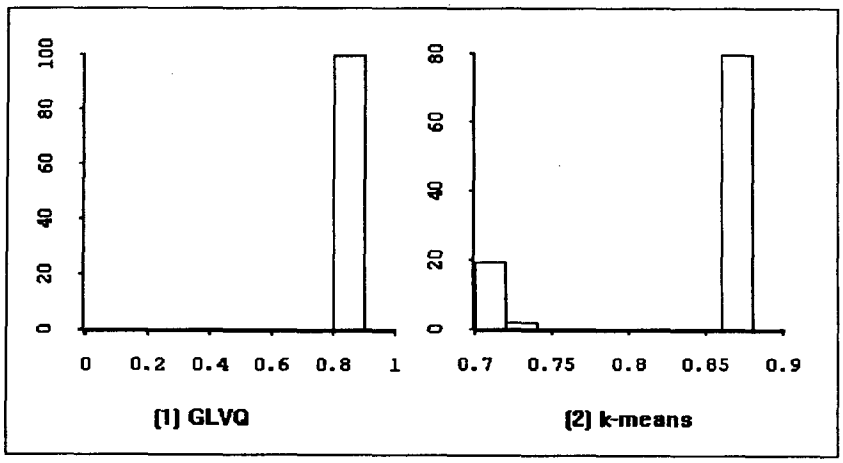
Y'은 GLVQ와 k-means클러스터링방법으로 구성한 10개의 클러스터로써 $Y'_{NUM,GLVQ}$, $Y'_{NUM,k\text{-}means}$을 각각 GLVQ, k-means알고리즘으로 구성한 집합이라고 한다. 두 자료를 GLVQ에 의해서 클러스터링을 하였을 경우 학습률 α_0는 0.4, 반복횟수는 2500번으로 하여 클러스터링을 하였다.

<표 1>은 IRIS데이타와 3.2절에서 특징추출한 자료를 각각의 방법으로 클러스터링 하여 "Natural" 클러스터와 비교한 결과인 데, 이때 GLVQ방법과 k-means 방법은 100번 시행하여 C 통계량의 평균과 표준편차를 비교한 결과이다. <그림 4.1>와 <그림 4.2>은 각 특징추출방법에 의하여 얻은 자료를 GLVQ와 k-means알고리즘으로 클러스터링 결과에 대한 C 통계량의 분포를 히스토그램으로 표현한 것이다.

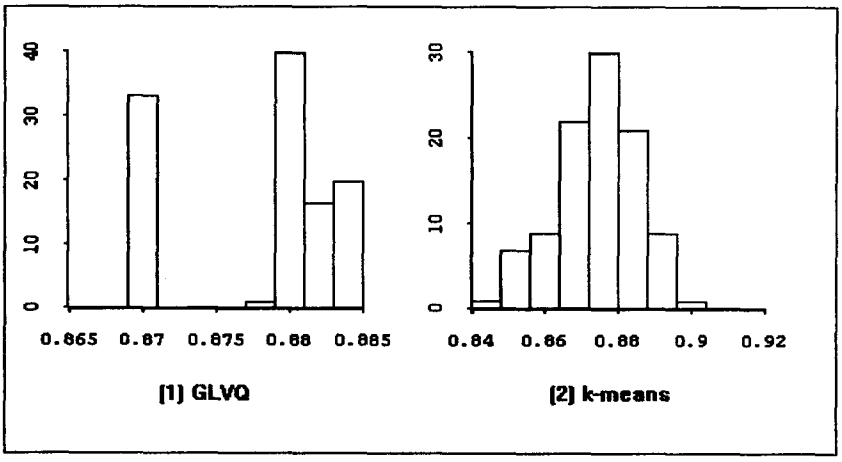
<표 1> "Natural" 클러스터 : GLVQ와 k-means 클러스터링의 C 통계량 비교

클러스터링 방법		IRIS	4×4
c의 평균	GLVQ	0.879731	0.878133
	k-means	0.843904	0.873909
c의 표준편차	GLVQ	0.000000	0.000035
	k-means	0.004147	0.000126

주) 4×4 : 16×16 자료를 4×4로 특징추출한 자료



<그림 4.1> "Natural" 클러스터와의 비교 : IRIS데이타의 C 통계량 히스토그램



<그림 4.2> "Natural" 클러스터와의 비교 : 16×16 자료를 4×4로 특징추출한 자료에 대한 C 통계량의 히스토그램

4.2 자료에 오류가 있는 클러스터와의 비교

본 절에서는 자료의 오류 또는 잡음이 들어 왔을 때 어느 클러스터링 방법이 잡음의 영향에 민감하게 반응하는가 하는 정도로써 그 성능평가를 하는 방법이다.

IRIS데이타를 이용할 경우에는 150개의 자료의 각 변수에 난수를 발생시킨 오류를 더하여 클러스터링을 비교하였다. IRIS데이타를 이용할 경우에 Y는 4.1절에서 각 클러스터링 방법으로 구성한 Y'집합이라고 하고, Y'은 원시 자료에 N(0,1)의 난수를 발생시켜 IRIS데이타 각 변수에 더하여 GLVQ방법과 k-means방법으로 클러스터링 한 집합이다.

$Y_{IRIS} = Y'_{Natural\ IRIS}$: "Natural"클러스터를 비교하기 위하여 각 클러스터링 방법에 의해서 구성된 클러스터의 집합.

$$Y'_{IRIS} = \{(\mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \ldots, \mathbf{x}'_{1,n_1}), (\mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \ldots, \mathbf{x}'_{2,n_2}), (\mathbf{x}'_{3,1}, \mathbf{x}'_{3,2}, \ldots, \mathbf{x}'_{3,n_3})\}$$

여기서, $\sum_{i=1}^{3} n_i = 150$,

$\mathbf{x}' = \mathbf{x} + \mathbf{r}$, $\mathbf{r} \sim \mathbf{N(0,1)}$를 갖는 난수.

0-9까지의 숫자 자료에서도 역시 Y를 4.1절에서 각 클러스터링 방법으로 구성한 Y'집합이라고 하고, Y'은 원시 자료에 10%의 잡음을 추가하여 3.2.2절의 특징추출 방법을 이용하여 100개의 패턴을 새로 구성하여 GLVQ와 k-means방법으로 클러스터링을 하였다. 여기서 10%의 잡음은 원시 자료에서 '1'의 수를 모두 합한 총수의 10%에 해당하는 수(=n이라고 한다)를 의미하며 이 '1'을 원래 패턴에 임의의 셀 위치에 n개 추가하였다.

Y_{NUM} = $Y'_{Natural\ NUM}$: "Natural"클러스터를 비교하기 위하여 각 클러스터링방법에 의해서 구성된 클러스터의 집합.

$$Y'_{NUM} = \{(\mathbf{x}'_{0,1}, \mathbf{x}'_{0,2}, \ldots, \mathbf{x}'_{0,n_1}), (\mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \ldots, \mathbf{x}'_{1,n_2}), \ldots, (\mathbf{x}'_{9,1}, \mathbf{x}'_{9,2}, \ldots, \mathbf{x}'_{9,n_9})\}$$

여기서, $\sum_{i=0}^{9} n_i = 100$,

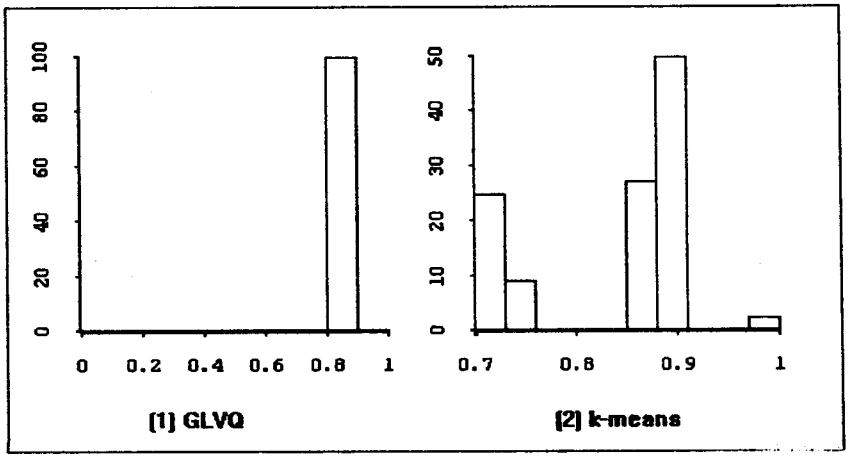
$\mathbf{x}'$: 10%의 잡음을 추가하여 특징추출한 패턴 벡터.

IRIS데이타와 각 특징추출 별로 얻은 자료에 대한 비교 결과는 <표 2>와 같으며, C 통계량에 대한 히스토그램은 <그림 4.3>과 <그림 4.4>에 나와 있다.

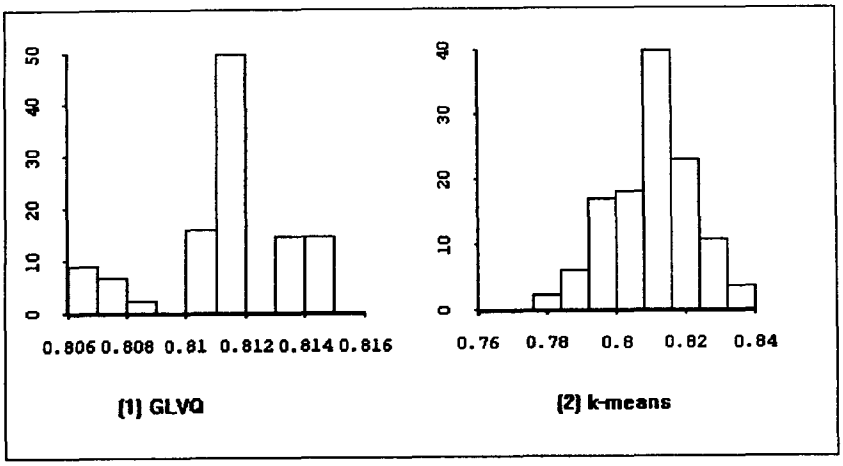
<표 2> 자료에 오류추가 : 신경망과 통계적 클러스터링의 C 통계량 비교

클러스터링 방법		IRIS	4×4
c의 평균	GLVQ	0.882148	0.811303
	k-means	0.833284	0.810192
c의 표준편차	GLVQ	0.000000	0.000005
	k-means	0.005432	0.000140

주) 4×4 : 16×16 자료를 4×4로 특징추출한 자료



<그림 4.3> 자료에 오류가 있는 클러스터의 비교 :
IRIS데이타에 대한 C 통계량의 히스토그램



<그림 4.4> 자료에 오류가 있는 클러스터의 비교 :
16×16 자료를 4×4로 특징추출한 자료에 대한 C 통계량 히스토그램

5. 결 론

본 논문은 클러스터링 방법 중 신경망에 의한 방법과 통계적인 방법을 비교하여 성능을 평가하는 데 그 목적이 있다. IRIS데이타와 필기체 숫자에 대한 자료를 신경망에 의한 클러스터링과 통계적인 방법에 의하여 클러스터링을 한 후 Rand C 통계량을 적용하여 비교한 결과에 국한시켰을 때 신경망 클러스터링의 GLVQ알고리즘이 통계적 방법의 k-means방법보다는 매 반복시행을 할 때마다 오분류되는 패턴이 약간 적게 나타났다. 또한 매 반복 시행을 할 때마다 오분류되는 패턴의 편차도 상대적으로 크게 나타나지 않았다. 이는 GLVQ알고리즘이 클러스터 전체에 연결된 연결강도를 수정하는 학습을 하기 때문에 초기 클러스터 중심값에 크게 영향을 받지 않고 매 반복 실행 때마다 거의 일정한 중심값을 계산할 수 있는 반면에 k-means 알고리즘은 초기에 실제 클러스터의 중심과 거리가 먼 패턴벡터가 주어졌을 때 그 클러스터에 의해서 다른 클러스터에까지 영향을 미치기 때문에 초기 중심벡터에 의존적이 되기 때문이다.

참 고 문 헌

[1] 김대수(1994), "신경망 이론과 응용 (I), (II)", 하이테크정보.

[2] 이성환(1994), "패턴인식의 원리 (I), (II)", 홍릉과학출판사.

[3] Byron J.T. Morgan (1984), *"Elements of Simulation"*, *Greate Britain at the Univ. Press Cambridge.*

[4] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, E. Ostrowski(1994), *"A Handbook of Small Data Sets"*, *Chapman & Hall.*

[5] Helge Ritter, Thomas Martinnets, Klaus Schulten (1992), *"Neural Computation and Self-Organizing Maps"*, *Addison-Wesley Publishing Co.*

[6] John Hertz, Anders Krogh, Richard G. Palmer (1991), *"Introduction to the Theory Neural Computation"*, *Addison-Wesley Publishing Co.*

[7] M.A. Ismail and M.S. Kamel(1989), "Multidimensional Data Clustering Utilizing Hybrid Search Strateges", *Pattern Recognition,* Vol. 22, No. 1, pp.75-89.

[8] Mark S. Aldenderfer, Roger K. Blashfield (1984), *"Cluster Analysis"*, Sage Publications Inc.

[9] Nikhil R. Pal, James C. Bezdek, Etric C.K. Tsao (1993), "Generalized Clustering Networks and Kohonen's Self-Organizing Scheme", *IEEE Trans. on Neural Networks,* Vol. 4, No. 4, pp.549-557.

[10] Robert Schalkkoff (1992), *"Pattern Recognition -statistical structureal and approaches"*, *John and Wiley & Sons, Inc.*

[11] Shunichi Shimoji, Sukhan Lee (1994), "Data Clustering with Entropical Sheduling", *International Joint Conference on Neural Network, Vol. 4,* pp.2423-2428.

[12] T. Kohonen (1990), "The Self-Organizing Map", *Proc. IEEE,* Vol. 78, No. 9, pp. 1464-1480.

[13] Terence D. Sanger (1989), "Optimal Unsupervised Learning in a Single-Layer Linear Feed forward Neural Network", *Neural Networks,* Vol. 2, p.459.

[14] Yoh-Han Pao(1989), *"Adaptive Pattern Recognition and Neural Network"*, *Addison-Wesley Publishing Co. Inc.*