논문 2004-41CI-6-4

클러스터 분석을 위한 IRC 기반 클러스터 개수 자동 결정 방법

(Systematic Determination of Number of Clusters Based on Input Representation Coverage)

신 미 영*

(Mi-Young Shin)

요 약

클러스터 분석에 있어 중요한 문제 중의 하나는 주어진 데이터에 내재된 적절한 클러스터의 수를 찾아내는 것이다. 본 논문에서는 이러한 클러스터의 개수를 체계적으로 결정하기 위하여 IRC (Input Representation Coverage) 개념을 새로이 정의하고, 이를 이용하여 주어진 데이터에 적합한 클러스터의 개수를 자동 결정하는 방법을 제시한다. 또한, 이러한 방법의 유용성 및 응용성을 알아보기 위하여 가상 데이터를 가지고 분석 실험을 하였으며, 실험을 통해 데이터에 내재된 실제 클러스터의 개수를 찾아내는 데에 제안된 방법이 매우 유용하게 사용될 수 있음을 보여준다.

Abstract

One of the significant issues in cluster analysis is to identify a proper number of clusters hidden under given data. In this paper we propose a novel approach to systematically determine the number of clusters based on Input Representation Coverage (IRC), which is newly defined as a quantified value of how well original input data in Gaussian feature space can be captured with a certain number of clusters. Furthermore, its usability and applicability is also investigated via experiments with synthetic data. Our experiment results show that the proposed approach is quite useful in approximately finding the *real* number of clusters implicitly contained in the data.

Keywords : cluster analysis, cluster number determination, input representation coverage

I. 서 론

클러스터 분석은 대용량의 데이터에 내재되어 있는 미지의 특성을 파악하기 위한 탐구 기법으로, 다양한 응용분야에서 중요하게 사용되어져 왔다[1]. 클러스터 분석에 있어서 무엇보다 중요하고 어려운 문제 중의 하나는 주어진 데이터에 내재된 클러스터의 개수를 찾아내는 것이다. K-means, SOM 등을 포함하는 많은 클러스터링 알고리즘은 사용자에 의해 입력된 클러스터의 개수를 기반으로 클러스터를 생성한다. 하지만, 실제 현실에서 클러스터 분석을 수행하는 경우는 대개 데이터에 대한 사전 지식이 없는 경우이다. 따라서 주어진 데이터에 적합한 클러스터의 개수를 결정하기 위해, 지금까지 주로 사용되어 온 방법은 임의로 선택된 다수의 클러스터 개수 후보들에 대하여 각각 클러스터를 생성하고 그 결과를 실루엣 계수 (Silhouette coefficient), F-통계치(F-statistic), 파티션 계수 (Partition coefficient) 등을 이용하여 평가함으로써 최적의 클러스터 개수를 결정하는 것이다[2]. 그러나 이와 같은 방법은 후보군이 얼마나 폭넓고 조밀하게 선택되었는가에 따라 최종 결과가 달라질 수 있고, 무엇보다 신뢰성 있는 결과를 얻기 위해서는 가능한 많은 후보군에 대해 시도해보는 것이 필요하다.

* 정회원, 한국전자통신연구원 바이오정보연구팀
(Bioinformatics Research Team, Electronics and Telecommunications Research Institute)
접수일자: 2004년9월1일, 수정완료일: 2004년11월15일

본 논문에서는 클러스터 개수 결정 문제를 보다 체계적인 방법으로 접근하기 위하여 입력데이터에 적합한 클러스터의 개수를 IRC 개념을 이용하여 자동 결정하는 방법을 제안한다. 이를 위해, 제 Ⅱ절에서는 입력 데이타에 대한 가우시안 특징 행렬 및 이에 대한 수학적 특성을 소개하고, 제 Ⅲ절에서는 가우시안 특징 행렬에 대한 IRC의 정의와 이를 이용한 클러스터 개수 결정 방법의 이론적 근거 및 알고리즘을 제시한다. 또한, 제 Ⅳ절에서는 가상데이터를 사용한 클러스터 분석 실험 방법을 구체적으로 기술하고, 제 Ⅴ절과 제 Ⅵ절에서는 실험 결과 분석 및 그 중요성에 대해 각각 논의한다.

Ⅱ. 가우시안 특징 행렬과 이의 수학적 특성

1. 가우시안 특징 행렬의 생성

클러스터 분석을 위해 임의의 데이터 집합 $\mathbf{D}=\{\mathbf{x}_i, i=1,\mathrm{K},n : \mathbf{x}_i=(x_{i1},\mathrm{K},x_{id})\in R^d\}$가 주어졌다고 가정하자. 즉, 각 데이터 $\mathbf{x}_i$는 d-차원의 벡터이고, 주어진 데이터 집합 $\mathbf{D}$는 n개의 데이터 벡터를 포함하고 있다고 하자. 이런 경우, 주어진 데이터 $\mathbf{D}$에 대한 가우시안 특징 행렬 $\tilde{\mathbf{D}}$는 다음과 같이 정의된다.

$$\tilde{\mathbf{D}}=\begin{bmatrix}\tilde{\mathbf{x}}_1\\ \tilde{\mathbf{x}}_2\\ \mathrm{M}\\ \tilde{\mathbf{x}}_n\end{bmatrix}=\begin{bmatrix}\phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \Lambda & \phi_n(\mathbf{x}_1)\\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \Lambda & \phi_n(\mathbf{x}_2)\\ \mathrm{M} & \mathrm{M} & & \mathrm{M}\\ \phi_1(\mathbf{x}_n) & \phi_2(\mathbf{x}_n) & \Lambda & \phi_n(\mathbf{x}_n)\end{bmatrix}$$

즉, 각 입력 데이터벡터 $\mathbf{x}_i$는 n개의 가우시안 함수 $\phi_1(\cdot),\phi_2(\cdot),\mathrm{K},\phi_n(\cdot)$을 통하여 n차원의 가우시안 특징 벡터 $\tilde{\mathbf{x}}_i=(\phi_1(\mathbf{x}_i),\phi_2(\mathbf{x}_i),\mathrm{K},\phi_n(\mathbf{x}_i))$로 변환된다. 가우시안 함수는 $\phi_j(\mathbf{x}_i)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/2\sigma^2)$로서 정의되며, 여기서 $\mathbf{x}_j$와 σ는 각각 가우시안 함수의 중심 벡터와 폭을 결정한다.

가우시안 특징 행렬 $\tilde{\mathbf{D}}$에서 j번째 열벡터 $[\phi_j(\mathbf{x}_1),\phi_j(\mathbf{x}_2),\mathrm{K},\phi_j(\mathbf{x}_n)]^T$는 j번째 데이터 벡터 $\mathbf{x}_j$를 중심벡터로 가지는 가우시안 함수에 의한 변환 결과로서 $\mathbf{x}_j$로부터 다른 데이터 벡터 $\mathbf{x}_i$ ($i=1,\mathrm{K},n$) 까지의 거리에 따라 그 값이 0과 1사이에서 상대적으로 결정된다. 여기서 j번째 열벡터의 각 구성요소 $\phi_j(\mathbf{x}_i)$은 데이터 벡터 $\mathbf{x}_i$와 j번째 가우시안 함수의 중심 벡터 $\mathbf{x}_j$간의 거리가 가까울수록 1에 가까운 값을 가지며, 그 거리가 멀수록 0에 가까운 값을 갖는다 (그림 1 참조).

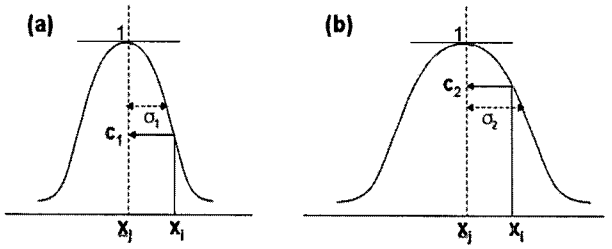


그림 1. 가우시안 폭 (σ)에 따른 두 벡터 간의 상대적 근접도

Fig. 1. Relative closeness between two data vectors for different choice of σ.

한편, 가우시안 함수의 폭을 결정하는 변수인 σ값은 가우시안 함수의 최종 형태를 결정하게 되며, 그림 1에서와 같이, 그 값이 클수록 가우시안 함수의 폭이 넓어지고 그 값이 작을수록 가우시안 함수의 폭이 좁아지는 형태가 된다.

2. 가우시안 특징 행렬의 계수(Rank)와 그 의미

일반적으로 행렬의 계수는 행렬의 하부 구조를 이해하는 데에 중요한 역할을 하는 것으로 알려져 왔다[3]. 구체적으로, 행렬의 계수는 그 행렬에 의해 표현되는 공간의 내재적 차원(intrinsic dimensionality)을 의미하며, 이것은 행렬의 하부 구조를 표현하는 데에 필요한 최소한의 기저 벡터(basis vector)의 수와 동일하다. 행렬의 계수가 그 행렬을 구성하는 행(row) 또는 열(column)의 수보다 작을 경우, 그 행렬을 계수 불충분(rank-deficient)하다고 일컫는다. 만약, 어떤 행렬이 계수 불충분하다면, 이것은 그 행렬을 구성하는 행벡터 또는 열벡터들이 상호 충분히 독립적이지 않음을 암시한다. 즉, 행벡터 또는 열벡터 간에 다소 중복성이 있음을 의미한다.

그러면, 이러한 행렬 계수의 개념을 앞에서 생성된 가우시안 특징 행렬 $\tilde{\mathbf{D}}$에 적용해보자. 가령, $\tilde{\mathbf{D}}\in R^{n\times n}$가 주어졌을 때, 이러한 행렬 $\tilde{\mathbf{D}}$가 계수 불충분하다면, 즉, 계수($\tilde{\mathbf{D}}$)$< n$ 이라면, 이것은 각 열벡터들이 상호 충분히 독립적이지 않음을 말해준다. 가우시안 특징 행렬의 경우, 이러한 상황은 변환에 사용되는 두 가우시안 함수의 중심 벡터들 간에 거리가 매우 가까울 때에 종종 일어난다. 즉, 두 벡터 $\mathbf{x}_i,\mathbf{x}_j$가 서로 가깝다면, 이

를 중심 벡터로 사용하는 두 가우시안 함수 $\phi_i(\cdot), \phi_j(\cdot)$의 변환 결과는 거의 공선적인(collinear) 경향이 있으며, 두 가우시안 함수의 중심 벡터가 서로 충분히 떨어져 있다면, 이 함수들에 의한 변환 결과는 상호 충분히 독립적일 것이다.

다만, 여기서 두 벡터 간의 근접도(closeness)는 절대적인 의미보다는 그림 1에서 보이는 바와 같이 가우시안 함수의 폭을 조절하는 σ 값에 따라 상대적으로 결정되는 것임을 알아야 한다. 즉, σ 값이 작은 경우, 두 데이터 벡터 간의 근접도는 σ 값이 큰 경우에 비해 엄격하게 적용되기 때문에, 동일한 절대적 거리의 두 벡터에 대해서조차도, σ 값이 커질수록 서로 간의 근접도는 상대적으로 커지며, 반면에 σ 값이 작을수록 근접도는 상대적으로 작아진다. 하지만, σ 값이 지나치게 큰 경우, 절대적 거리의 큰 차이에도 불구하고 근접도의 변화는 매우 적은 특징이 있다.

2.3. 가우시안 특징 행렬의 클러스터 분석에의 응용

클러스터링의 목적은 하나의 클러스터에 속한 객체 간의 상호 유사성(homogenity)을 가능한 높이고, 다른 클러스터에 속한 객체 간의 이질성 (seperation)을 가능한 높이는 방식으로 그룹화하는 것이다. 따라서, k-means와 같은 반복적인 방법에 기반한 분할 클러스터링 알고리즘의 경우, 클러스터의 초기 중심점들은 클러스터 간의 이질성을 극대화시킬 수 있도록 가능한 서로 멀리 떨어지도록 선택하는 것이 바람직할 것이다.

앞서 언급하였듯이, 가우시안 특징 행렬은 입력 데이터 집합에 속한 모든 서로 다른 두 데이터 간의 상대적 거리에 대한 정보를 제공한다. 이러한 가우시안 특징 행렬 $\tilde{\mathbf{D}}$의 계수는 입력데이터로부터 추출된 가우시안 특징 공간을 잘 표현할 수 있는 상호 독립적인 벡터의 개수를 의미함을 안다. 따라서 생성될 클러스터의 초기 중심점(centroid)이 충분히 떨어져 있도록 선택되어진다고 가정할 때, 가우시안 특징 행렬의 계수는 데이터에 내재한 클러스터의 개수에 대한 예측값으로 사용될 수 있을 것이다.

Ⅲ. IRC 정의 및 이에 기반한 클러스터 개수 결정 방법

그러면, 가우시안 특징 행렬의 계수는 어떻게 구할 수 있을까? 대개 실제 응용분야에서 얻어지는 데이터는 많은 잡음(noise)을 포함하고 있기 때문에, 일반적으로 행렬의 계수는 잡음을 고려한 계수 추정치(rank estimation)인 유효 계수(effective rank)를 사용하여 왔다. 그러나 이러한 유효 계수를 계산하기 위해서는 잡음 효과(noise effect)를 구체적으로 명시할 필요가 있다[3]. 이를 위해 본 논문에서는 IRC (Input Representation Coverage) 개념을 정의하고, 이를 가우시안 특징 행렬의 유효 계수를 계산하는 데에 이용하기로 한다. 자세한 내용은 아래와 같다.

1. 가우시안 특징 행렬의 유효 계수

행렬의 유효 계수 측정은 특이값 분해 (Singular Value Decomposition, SVD)를 이용한 방법이 주로 사용되어져 왔다[4].

특이값 분해(SVD) 이론에 따르면[3], 가우시안 특징 행렬 $\tilde{\mathbf{D}} \in R^{n\times n}$은, 다음과 같이 $\tilde{\mathbf{D}} = \mathbf{USV}^T$를 만족하는 행렬 U, S, V로 분해될 수 있다. 여기서 $\mathbf{U} \in [\mathbf{u}_1, \mathbf{u}_2, \mathrm{K}, \mathbf{u}_n] \in R^{m\times n}$와 $\mathbf{V} \in [\mathbf{v}_1, \mathbf{v}_2, \mathrm{K}, \mathbf{v}_n] \in R^{n\times n}$는 직교행렬 (orthogonal matrices) 이고, $\mathbf{S} = diag(s_1, s_2, \mathrm{K}, s_n) = R^{n\times n}$, $s_1 \geq s_2 \geq \Lambda \geq s_n \geq 0$은 대각선 행렬이다. 특히, 대각서 행렬 S의 구성요소인 s_i는 행렬 $\tilde{\mathbf{D}}$의 i번째 특이값이라 한다.

이 때, 가우시안 특징 행렬 $\tilde{\mathbf{D}}$의 유효 계수(r_ε)는 $\varepsilon > 0$에 대해 $r_\varepsilon = rank(\tilde{\mathbf{D}}, \varepsilon)$로서 정의되며, 이 때의 r_ε은 다음과 같은 조건

$$s_1 \geq \Lambda \geq s_{r_\varepsilon} > \varepsilon > s_{r_\varepsilon+1} \geq \Lambda \geq s_n \quad (1)$$

을 만족해야 한다. 여기서, ε는 데이터 잡음량을 나타낸다.

2. IRC(Input Representation Coverage)의 정의

IRC는 데이터의 가우시안 함수 변환 결과인 가우시안 특징 행렬에 의해 표현되는 본래의 가우시안 특징 공간과, 행렬의 계수 r ($< n$) 을 가지는 가우시안 특징 행렬에 의해 표현되는 근사(approximate) 가우시안 특징 공간과의 상대적 차이를 정량화하기 위한 장치이다. 구체적으로, 가우시안 특징 행렬 $\tilde{\mathbf{D}} \in R^{n\times n}$에 대하여, 행렬 계수 r이 n보다 작다면, 계수 축소된 (rank-reduced) 가우시안 특징 행렬은 $\tilde{\mathbf{D}}_r = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T$ 으로 표현될 수

있고, 이 때의 행렬 $\tilde{\mathbf{D}}_r$에 대한 IRC는 다음과 같이 정의된다.

$$IRC(\tilde{\mathbf{D}}_r) = 1 - \frac{\| \tilde{\mathbf{D}} - \tilde{\mathbf{D}}_r \|_2}{\| \tilde{\mathbf{D}} \|_2} \tag{2}$$

또한, 상기 수식 (2)는 관련 정리 (참고 [3], Theorem 2.5.3) 와 2-Norm의 성질에 따라, 다음과 같은 수식으로 표현될 수 있다.

$$IRC(\tilde{\mathbf{D}}_r) = 1 - \frac{s_{r+1}}{s_1} \tag{3}$$

여기서 s_1, s_{r+1}는 첫 번째 특이값과 r+1번째 특이값을 각각 나타낸다.

상기 수식 (3)에 따른 IRC 측정치는 0과 1사이의 범위 내의 값을 가지는 특징이 있으며, 그 값이 1에 근접할수록 계수축소된 가우시안 특징 공간과 본래 데이터의 가우시안 특징 공간과 가깝다는 것을 의미한다.

3. 행렬 계수에 따른 IRC의 변화 추이

앞의 수식 (3)을 이용하여, 가우시안 특징 행렬의 계수(*r*)의 변화에 따른 IRC 값들의 변화 추이를 살펴보자. 250개의 가상 데이터로 구성된 실험데이터 (4.1절 참조)를 사용하여 가우시안 행렬 계수 *r=(1:1:250)*에 각각 대응하는 IRC 값을 시뮬레이션한 결과는 그림 2와 같다. 그림 2에 따르면, 가우시안 특징 행렬 계수(*r*)가 증가함에 따라 그에 대응하는 IRC 값은 점차적으로 증가하는 특징을 보인다. 특히, 가우시안 함수 폭을 조절하는 σ의 크기에 따라 다소 차이가 있으나, 전반적으로 초기 단계에서 행렬 계수의 증가는 IRC를 급격하게 높이는 경향이 있으며, 어느 단계를 지나면, 행렬 계수의 증가에도 불구하고 IRC의 변화가 매우 적어지는 양태를 보인다.

4. IRC에 의한 클러스터 개수의 선정

그러면, 상기에서 정의한 IRC를 이용하여 어떻게 클러스터 개수를 선정할 수 있는 지를 살펴보기로 하자.

가령, IRC에 대한 오류 허용율(error allowance)이 δ, $0 < \delta \leq 1$라고 가정해보자. 다시 말해, 우리가 원하는 IRC 기준치가 $1-\delta$ 라면, 주어진 데이터 집합에 대한 적절한 클러스터의 개수는 다음과 같은 조건을 만족하는 최소한의 k로 주어질 수 있다.

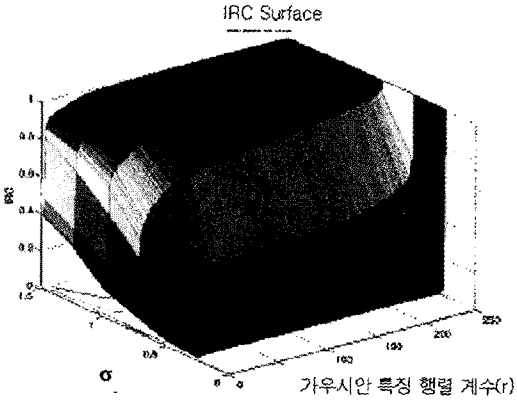


그림 2. 가우시안 특징 행렬 계수(r) 및 폭 조절 변수(σ)의 변화에 따른 IRC 측정값의 변화 곡면.

Fig. 2. Surface of IRC measure over Gaussian feature matrix rank (r) and the width (σ).

$$1 - \frac{s_{k+1}}{s_1} \geq 1 - \delta \tag{4}$$

위의 수식 (4)은 아래와 같이 표현 될 수 있다.

$$\frac{s_{k+1}}{s_1} \leq \delta$$

$$s_{k+1} \leq s_1 \times \delta$$

$$\boxed{s_r > s_1 \times \delta} \tag{5}$$

상기 수식 (5)를 앞에서 기술한 유효 계수에 대한 조건식 (1)에 대응시키면, 이러한 조건을 만족하는 클러스터의 개수 k는 가우시안 특징 행렬 $\tilde{\mathbf{D}}$의 유효 계수 계산식에 잡음 효과 $\varepsilon = s_1 \times \delta$ 을 대체한 것과 같다. 즉, 가우시안 특징 행렬 $\tilde{\mathbf{D}}$와 IRC에 대한 오류 허용율 δ가 주어졌을 때, 이를 만족하는 클러스터의 개수 k는 아래 수식 (6)과 같이 구해질 수 있다.

$$k = rank(\tilde{\mathbf{D}}, \varepsilon) = rank(\tilde{\mathbf{D}}, s_1 \times \delta) \tag{6}$$

그리하여 주어진 데이터 집합 **D**에 대한 클러스터 개수 (k) 선정 알고리즘은 다음과 같이 요약될 수 있다.

1. 데이터 집합 D에 대하여

 1.1 d ← 입력 데이터 벡터의 차원 수

 1.2 $0 < \sigma < \sqrt{d/2}$ 범위내에 속한 후보값 σ'을 선택

2. 각 후보값 σ에 대하여

 2.1 가우시안 특징 행렬 $\tilde{\mathbf{D}}$를 생성한다

2.2 $s_1 \leftarrow \tilde{\mathbf{D}}$의 첫 번째 특이값을 구한다

2.3 클러스터 개수 $k \leftarrow rank(\tilde{\mathbf{D}}, s_1 \times \delta)$를 얻는다.

Ⅳ. 실험 방법

본 논문에서 제안한 IRC에 의한 클러스터 개수 선정 방법의 검증을 위하여 아래와 같이 실험을 수행하였다.

1. 실험 데이터

실험 분석을 위해, 클러스터 분석에 관한 최근 연구[5]에서 사용되었던 다섯 개의 시간열(time-series) 패턴을 미리 설정하고, 이를 내재적으로 포함하는 250개의 데이터 벡터로 구성된 가상 데이터 집합을 생성하였다. 각 데이터 벡터는 10개의 시점에 대한 측정치로 구성된 다섯 개의 시간열 패턴 중의 하나에 가우시안 분포 $N(0,0.5^2)$을 가진 잡음을 추가하여 생성되었고, 각 시간열 패턴에 대한 50개의 서로 다른 데이터 벡터가 생성되었다.

2. 전체 실험 과정

클러스터 분석을 위해, 본 논문에서는 주어진 데이터 집합에 대해 다음과 같은 4가지 단계로 실험을 진행하였다.

(1) 첫째, 주어진 데이터 집합 **D**에 대해 가우시안 특징 행렬 $\tilde{\mathbf{D}}$를 생성한다. 이를 위해, 가우시안 함수 폭 조절 변수 σ는 $0<\sigma<\sqrt{d/2}$ 범위 내에서 선택하였다. (d: 입력 데이터 벡터의 차원 수)

(2) 둘째, 주어진 IRC에 대한 오류 허용율 δ ($0<\delta\leq 1$)가 주어졌을 때, IRC 기준치 $1-\delta$를 만족하는 최소한의 클러스터 개수 k를 결정한다. 이 때, 오류 허용율 δ는 δ=0.01, 0.05, 0.1, 0.2, 0.3에 대해 각각 실험을 진행하였다.

(3) 셋째, 단계 (2)에서 결정된 클러스터 개수 k를 가지고, k-means 방법에 의해 클러스터를 생성한다. 이 때, 데이터 벡터로서 가우시안 특징 벡터를 사용하여 클러스터링을 하였다.

(4) 넷째, 단계 (3)에서 생성된 클러스터 결과를 Adjusted Rand Index를 사용하여 평가한다.

클러스터 생성과 평가 방법에 관한 내용은 각각 4.3절과 4.4절에서 보다 상세하게 설명하기로 한다.

3. 클러스터의 생성

본 논문에서는 클러스터를 생성하기 위하여, 주어진 데이터 $\mathbf{D}=\{\mathbf{x}_i, i=1,\mathrm{K},n : \mathbf{x}_i=(x_{i1},\mathrm{K},x_{id})\in R^d\}$에 대한 가우시안 특징 벡터 $\tilde{\mathbf{x}}_i=(\phi_1(\mathbf{x}_i),\phi_2(\mathbf{x}_i),\mathrm{K},\phi_n(\mathbf{x}_i))$, $i=1...n$ 에 k-means 방법을 적용함으로써 클러스터를 생성한다. 구체적으로, 생성하고자 하는 클러스터의 개수가 $k(<n)$라고 할 때, 주어진 n개의 데이터 벡터 $\mathbf{x}_1,\mathbf{x}_2,\mathrm{K},\mathbf{x}_n$에 대응하는 n개의 가우시안 특징 벡터 $\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\mathrm{K},\tilde{\mathbf{x}}_n$를 대상으로 서로 겹침이 없는 k개의 그룹으로 분할하는 것이다. 이 때, 분할된 그룹 각각은 하나의 클러스터에 해당된다.

k개의 분할된 그룹을 얻기 위해서는 먼저 k개의 임의의 가우시안 특징 벡터를 선택하여 생성될 클러스터의 초기 중심으로 설정하고, 이러한 k개의 클러스터 초기 중심을 기반으로 나머지 가우시안 특징 벡터 각각은 가장 가까운 초기 중심을 가진 클러스터에 할당된다. 일단 초기 할당이 끝나면, 각 클러스터의 중심을 그 클러스터에 현재 할당되어 있는 가우시안 특징 벡터들의 평균으로 재설정하고 상기 과정을 반복한다, 이러한 과정은 목적 함수에 관해 최적화되어 k개의 클러스터 중심이 안정화될 때까지 반복한다. 더 이상 k개의 클러스터 중심에 변화가 없으면, 이 때의 클러스터 멤버쉽이 최종 클러스터 결과가 된다.

클러스터 생성을 위한 두 벡터간의 거리 측정은 유클리드 거리(Euclidean distance)를 사용하여 계산되었으며, k-means 방법에서 클러스터의 초기 중심이 임의로 선정되어 일어나는 우연성을 줄이기 위하여 10번의 반복수행에 대한 평균 결과를 클러스터 평가에 사용하기로 한다.

4. 클러스터의 평가

일반적으로 클러스터를 평가하기 위해서는 주어진 데이터에 내재된 정답 클러스터가 이미 알려져 있는 경우와 그렇지 않은 경우에 각기 다른 방법을 사용할 수 있다. 본 논문에서는 앞서 제안한 클러스터의 개수 설정 방법에 대한 객관적 검증을 위하여 정답 클러스터를 알고 있는 실험 데이터를 사용하였고, 이러한 지식을 클러스터 결과 평가시 활용하였다.

본 실험에 대한 클러스터의 결과 평가는, 최근의 유사한 연구[13]에서 사용되었던, adjusted rand inx (ARI)를 사용하여 이루어졌으며, 이것은 정답 셋에 나타난 분할 그룹과 클러스터링 결과 생성된 분할 그룹이 통계적

으로 얼마나 잘 일치하는 지를 측정함으로써 클러스터의 결과를 평가하는 방법이다.

가령, $U=\{u_1,K,u_R\}$가 정답 셀에 나타난 분할 그룹(이하 클래스)이고, $V=\{v_1,K,v_C\}$가 알고리즘에 의해 생성된 분할 그룹(이하 클러스터)라고 가정해 보자. 이런 경우, 참고 문헌 [6]에 의하면, ARI는 아래 수식 (7)과 같이 정의된다.

$$\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}\right]-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}} \quad (7)$$

여기서 n은 데이터 셀에 속한 전체 데이터의 크기를 나타내며, n_{ij}는 클래스 u_i와 클러스터 v_j 둘 다에 속하는 데이터의 개수를, n_i와 n_j는 클래스 u_i와 클러스터 v_j에 속하는 데이터 개수를 각각 나타낸다.

상기와 같이 계산된 ARI 값은 1에 가까울수록, 비교 대상의 두 분할 그룹이 거의 완전히 일치한다는 것을 의미한다. 본 실험에서는 알고리즘에 의해 생성된 클러스터 결과와 정답 셀에 나타난 분할 그룹간의 일치 정도를 비교하는 것이므로 ARI 측정치가 1에 가깝다면, 이것은 생성된 클러스터 결과가 정답 셀에 매우 가깝다는 것을 의미한다.

V. 실험 결과

1. IRC에 의한 클러스터 개수의 선정

본 논문에서 제안한 IRC 기반 클러스터 개수 자동 설정 방법에 따르면, 주어진 실험 데이터에 대한 클러스터 개수를 선정하기 위해서는, 가우시안 함수 폭 조절 변수 σ 값에 대한 사전 설정 및 IRC에 대한 오류 허용율 δ 에 대한 언급이 필요하다. 먼저, 가우시안 함수 폭 조절 변수 σ 값은 실험 데이터의 각 벡터 차원 수가 d=10이기 때문에 앞에서 언급한 휴리스틱을 따라 $0<\sigma<\sqrt{(10/2)}\approx 2.236$ 범위 내에 속한 σ=(0.25:0.25:2.0) 값들을 사용하였고, IRC에 대한 오류 허용율은 δ=0.01, 0.05, 0.1, 0.2, 0.3에 대해 각각 실험을 진행하였다. 이 결과, 주어진 실험 데이터에 대해 선정된 클러스터 개수 (k)는 표 1과 같다.

표 1에서 나타난 클러스터 개수 선정 결과는 변수 k, σ 그리고 IRC 측정치 간에 존재하는 관계적 특성에 의해 나타난 결과로 볼 수 있다. 이들 간의 관계적 특성을 이해하기 위하여 주어진 실험 데이터에 대해 IRC에 대한 시뮬레이션을 수행하였고, 그림 3와 같은 결과를 얻었다.

표 1. 다양한 오류 허용율 (δ)및 가우시안 함수 폭 조절 변수 (σ)값에 대한 IRC 기반 클러스터 개수 (k)의 선정 결과

Table 1. Results of the number of clusters (k) determined by IRC based method for different δ and σ.

	δ =0.01	δ =0.05	δ =0.1	δ =0.2	δ =0.3
σ =0.25	250	250	250	250	250
σ =0.5	250	250	250	250	250
σ =0.75	250	248	224	104	30
σ =1.0	244	133	53	17	7
σ =1.25	170	48	25	6	5
σ =1.5	100	30	9	5	4
σ =1.75	58	19	7	4	4
σ =2.0	44	11	4	4	4

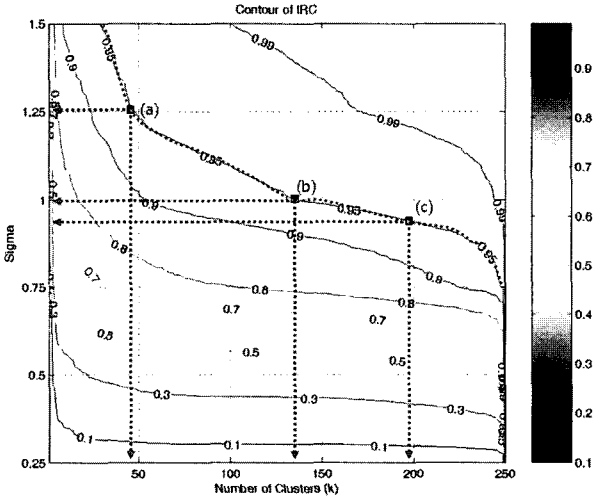


그림 3. 클러스터 개수 (k), 가우시안 함수 폭 조절 변수 (σ), 그리고 IRC 측정치 간의 관계적 특성에 대한 시뮬레이션 결과. 특정 IRC를 충족시키는 많은 (k,σ)조합이 존재함.

Fig. 3. Simulation results for relationship among k, σ and IRC. There are many candidates (k,) satisfying a specific IRC criterion.

앞에서도 언급한 것처럼 IRC 값은 두 가지 변수인 클러스터 개수(k)와 가우시안 함수 폭 조절 변수(σ)에 의해 고유하게 결정된다. 특히, 그림 3에서와 같이, 특정 IRC 값을 충족시키는 많은 (k,σ)조합이 존재하므로, 이는 곧 IRC에 대한 오류 허용율(δ)이 선택될 경우, 이를 충족시키는 많은 (k,σ) 조합이 존재함을 의미한다. 또한, 가우시안 함수 폭 조절 변수(σ) 값을 얼마나 조밀하게 선택하느냐에 따라 최종 선정된 결과가 달라질 수 있다.

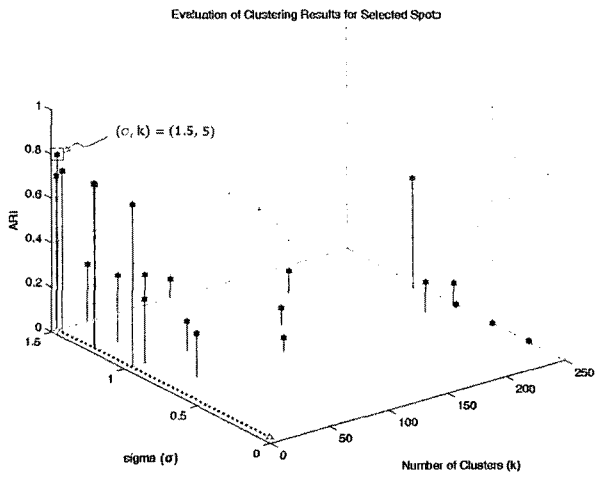


그림 4. 표1에서 선정된 (k,σ)조합에 대한 클러스터 결과 평가치

Fig. 4. Evaluation of clustering results for (k,σ) combinations shown in Table 1.

2. 클러스터의 생성 및 평가

표 1에서와 같이 클러스터 개수(k)가 결정되면, 4.3절에 기술한 바와 같이 주어진 데이터의 가우시안 특징 벡터에 k-means 방법을 적용함으로써 k개의 클러스터를 생성하게 된다. 생성된 클러스터 결과 평가를 위해 수식 (7)에 명시된 대해 ARI 측정치를 계산하고, 표 1에 나타난 (k,σ) 조합들 중 그 값이 가장 큰 클러스터를 최적의 클러스터로 결정한다. 아래 그림 4는 표 1에서 선정된 (k,σ)조합들에 대한 클러스터 평가 결과이다.

그림 4에 따르면, 최적의 클러스터 결과는, (k, σ)=(5,1.5) 조합에 대해 생성된 클러스터이고, 이 때의 ARI 측정치는 0.791이었다. 다시 말해, 본 논문에서 제안한 방법을 통해, 클러스터 개수 k=5인 경우 최상의 클러스터 결과를 얻을 수 있었다.

3. 시뮬레이션에 의한 결과 검증

5.2절에서 얻어진 최적의 클러스터 개수 k=5는 앞서 기술한 IRC 기반 클러스터 개수 자동 결정 방법에 의해 선정된 (k,σ)조합들로부터 최적의 값을 찾아낸 결과이다. 그러면, 실험 데이터에 대한 최적의 클러스터의 개수로 선택되어진 k=5가 실제로 전체 파라미터 공간에서 최적의 결과인지를 검증해 볼 필요가 있다. 이를 위해 클러스터 개수의 범위 k=(5:5:250)와 가우시안 함수 폭 조절 변수 σ=(0.25:0.25:2.0)로부터 구성된 파라미터 공간에서 가능한 모든 (k,σ)조합에 대해 클러스터를 생성하고 각각의 결과에 대한 ARI 측정치를 계산함으로써 전체 ARI surface를 시뮬레이션하였다 (그림 5 참조). 이 때, k-means방법에 의한 클러스터 생성시 나타나는 무작위적 초기치 설정에 따른 우연성을 줄이기 위해 3번의 독립된 실험을 진행하였고, 이들의 ARI 측정치의 평균값을 최종 평가를 위해 사용하였다.

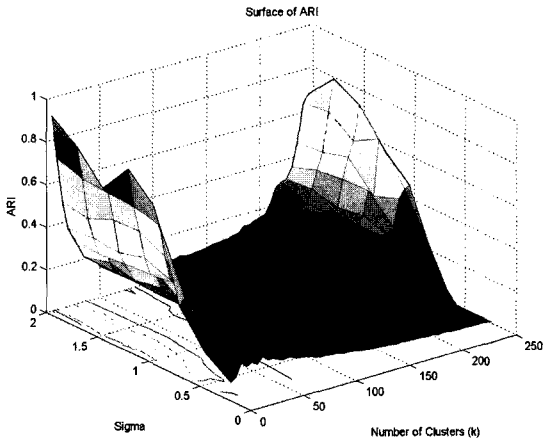


그림 5. 전체 ARI surface 시뮬레이션 결과

Fig. 5. Simulation of entire ARI surface.

그림 5에 나타난 전체 ARI surface에서도 최적의 결과는 클러스터 개수 k=5일 때였고, 클러스터의 개수(k)가 5를 넘을 때 점차 그 값이 감소하다가 클러스터의 개수(k)가 전체 데이터 크기에 가까워질 때 그 값이 다시 증가하는 추세를 보였다. 이것은, 앞서 제안한 IRC 기반 클러스터 개수 결정 방법이 주어진 오류 허용율에 대해 자동으로 선정된 후보 조합 (k,σ)들에 대한 클러스터링 시도만으로 전체 공간에서 최적의 (또는 최적에 가까운) 결과를 지니는 클러스터의 개수(k)를 적절히 찾아낼 수 있음을 보여주고 있다.

VI. 결 론

지금까지 본 논문에서는 IRC 개념을 이용한 클러스터 개수 자동 설정 방법에 대해 살펴보았다. 이를 위해, IRC 개념을 주어진 데이터의 가우시안 특징 공간이 특정한 클러스터 개수로 표현되었을 때 본래의 가우시안 특징 공간과 얼마나 유사한지를 정량화하고, 이를 이용하여 주어진 데이터에 적합한 클러스터의 개수를 결정하였다. IRC 기반 클러스터 개수 자동 결정 방법의 실제 활용을 위해서는 IRC 오류 허용율 (δ)에 대한 적절한 값의 설정이 필요하다. 그러나 이전의 다른 방법에서와 달리, 이러한 변수의 값이 최종 결과에 미치는 영

향을 미리 예측할 수 있기 때문에 무작위적인 단순 반복 실험이 아닌, 체계적이고 신뢰성 있는 실험 진행이 가능하다는 장점이 있다. 실험 결과에서 보듯이, IRC 기반 클러스터 자동 개수 설정 방법은 실제 데이터에 내재된 클러스터의 개수에 매우 근접한 수치를 쉽게 찾을 수 있도록 해준다. 게다가 수학적 이론에 근거하고 있기 때문에 결과에 대한 신뢰성이 높아 클러스터 분석을 필요로 하는 실사용자들에게는 매우 매력적인 방식일 것으로 판단된다.

참 고 문 헌

[1] A.K. Jain, M.N. Murty and P.J Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 3, No. 3, 264-323, 1999

[2] P. Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc, 2002

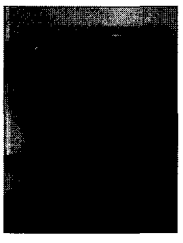
[3] Golub, G.H. and Van Loan, C.F., *Matrix Computation (3rd edition)*, The Johns Hopkins University Press (1996)

[4] D.C. Lay, *Linear Algebra and Its Applicastions,* Addison Wesley Longman, Inc., 2nd edition, 1997.

[5] J. Quackenbush, "Computational Analysis of Microarray Data", Nature Reviews Genetics, vol. 2, 418-422, June 2001.

[6] K.Y. Yeung, D.R. Haynor and W. L. Ruzzo, "Validating Clustering for Gene Expression Data", *Bioinformatics* (2001), Vol. 17(4), 309-318.

저 자 소 개



신 미 영(정회원)
1991년 연세대학교 전산과학과 학사 졸업.
1993년 연세대학교 대학원 전산과학과 석사 졸업.
1998년 미국 Syracuse Univ. 전산학 박사 졸업.
1999년~현재 한국전자통신연구원 바이오정보 연구팀 선임연구원
<주관심분야: 패턴인식, 데이터마이닝, 바이오인포매틱스>