Temporal 데이터의 최적의 클러스터 수 결정에 관한 연구

A Study for Determining the Best Number of Clusters on Temporal Data

전진호, 조영희, 이계성
단국대학교 전자계산학과

Jin-Ho Jeon(jhjy@dankook.ac.kr), Young-Hee Cho(zerowh@daum.net),
Gye-Sung Lee(gslee@dku.edu)

요약

Temporal 데이터의 클러스터링 방법론 중의 하나로 모델기반 방법론이 있다. 이는 각 클러스터에 대하여 오토마타기반의 모델을 가정하는 것이다. 개별 모델을 추출하기 위해서는 먼저 전체 데이터에 대한 적합한 모델을 찾는 것이 필요하다. 전체에 대한 모델은 데이터집합에 대한 최적의 클러스터의 수를 결정함으로 개별 모델 구축의 준비를 완료한다. 본 연구에서는 클러스터 수를 결정하기 위한 기준인 베이지안 정보기준(BIC : Bayesian Information Criterion) 근사법의 활용도를 검증하고 데이터 크기와 BIC 값의 상관관계를 파악함으로 탐색 효율을 높이는 방안을 제안한다.

실험에서는 인위적 모델을 통하여 생성된 인공적인 여러 형태의 데이터집합을 활용하여 BIC 근사 측도의 활용성에 대해 살펴보았다. 실험결과에서 보여주는 것처럼 BIC 근사 측도는 데이터의 크기가 비교적 클 경우에 올바른 파티션의 사이즈를 추정함을 확인하였다.

■ 중심어 : | 시계열 데이터 | 클러스터링 | 베이지안정보기준 | 클러스터 수 | 모델기반 |

Abstract

A clustering method for temporal data takes a model-based approach. This uses automata based model for each cluster. It is necessary to construct global models for a set of data in order to elicit individual models for the cluster. The preparation for building individual models is completed by determining the number of clusters inherent in the data set. In this paper, BIC (Bayesian Information Criterion) approximation is used to determine the number clusters and confirmed its applicability. A search technique to improve efficiency is also suggested by analyzing the relationship between data size and BIC values. A number of experiments have been performed to check its validity using artificially generated data sets. BIC approximation measure has been confirmed that it suggests best number of clusters through experiments provided that the number of data is relatively large.

■ Keyword : | Temporal Data | Clustering | BIC | Number of Clusters | Model-based |

* 본 연구는 2004학년도 단국대학교 대학연구비의 지원으로 연구되었습니다.

접수번호 : #050808-001
접수일자 : 2005년 08월 08일
심사완료일 : 2005년 12월 08일
교신저자 : 전진호, e-mail : jhjy@dankook.ac.kr

I. 서 론

감독분류(supervise classification)와 달리, 무감독분류(unsupervised classification), 또는 클러스터링(clustering)은 클래스 정보가 기록되지 않은 데이터를 가정한다. 목적은 그룹 내에서는 객체 유사도(similarlity) 그리고 그룹들 사이에서는 객체 비유사도(dissimilarity)가 최적이 되도록 구조를 생성하는 것이다. 이러한 구조의 분류와 해석은 각 그룹 내에 특징값 분포들에 의해서 세워진 모델들을 분석함으로서 달성되어진다.

과거 클러스터 분석 기법들은 정적특징[1]들에 의해 묘사된 데이터에 초점을 맞추었다. 즉, 시간에 대해서 특징값의 변화가 없거나 또는 변화를 무시해도 된다. 그러나 많은 실세계의 응용에서 대부분의 시스템들은 동적이며 그것들은 시간적인 특징들에 의해서 묘사되고 그것들의 값들은 관측 기간 동안 의미 있게 변한다.

본 연구에서는 시간적 특징들로 묘사되는 데이터의 클러스터링 방법론을 살펴본다. temporal 데이터의 클러스터링 문제는 정적데이터의 클러스터링의 문제보다 복잡하다. 이유는 첫째, 데이터의 차원이 동적인 경우에서는 정적인 경우보다 의미 있게 크다. 데이터 객체들이 정적특징들로 특징지어지면, 각 특징에 대해 오직 하나의 값이 표현된다. 시간의 특징인 경우에서는, 각 특징은 값들의 시퀀스가 되기 때문이다. 둘째, 클러스터의 정의와 해석의 복잡도는 동적인 데이터가 갖는 차수의 크기에 의해 증가되기 때문이다[2].

본 연구에서는 클러스터링 방법론에서 모델기반의 방법론을 가정한다. 이는 각 클러스터에 대하여 오토마타 기반의 모델을 가정하는 것인데, 클러스터링의 목적은 데이터에 가장 적합한 모델의 집합을 찾는 것으로 본 연구에서 클러스터링 알고리즘은 크게 두 개의 과제로 나누어 볼 수 있다. 첫째, 데이터의 최선의 분할을 표현하는 클러스터의 최적 집합을 찾는 것이다. 둘째, 각각의 클러스터에 가장 적합한 모델을 생성하는 것이다. 즉 궁극적 목적은 동적시스템의 표현을 잘 설명할 수 있으며 정확한 모델을 개발하는 것이다. 본 연구에서는 위의 두 가지 과제 중에서 첫 번째 과제인 최적의 클러스터의 수를 결정하는 과정에 대해서 살펴보고자 한다. 클러스터의 수를 결정할 베이지안 정보기준(BIC) 측정에 대해서 고찰과 실험을 통해 유효성을 살펴본다.

II. 배경 연구

클러스터링의 주요 과정은 크게 세 가지의 과정으로 이루어진다. 첫 번째로 특징선택 및 추출과정이다. 이는 관련된 특징들의 부분집합으로 특징을 나타내는 것과 확인시키는 과정으로서 매우 중요한 과정이다. 둘째, 클러스터링의 유도로서 크게 파라메틱 방법과 비파라메틱 방법이 있다. 파라메틱 방법은 알려진 데이터집합으로부터 형성된 각 클러스터의 확률분포함수(pdf)를 추정한다. 비파라메틱 방법들은 K-means 클러스터링 알고리즘처럼 클러스터 내에 객체 유사도를 최대화하는 파티션을 생성하는 것이 목적이다. 마지막으로서 클러스터의 확인 단계는 생성된 클러스터는 중요하고 의미 있다는 것을 확인시켜 준다.

전통적으로 클러스터 분석 방법들은 정적 특징값들을 갖는 것으로 묘사되는 데이터의 분석에 대하여 표현되어졌다. 정적데이터에 대하여 개발된 이러한 클러스터링 기법들의 대부분은 시간적 데이터의 표현과 모델링 방법들에 적당한 작업의 사용을 통해 시간적 데이터의 클러스터링 문제에 적용되어 왔다. 기법들의 변화는 데이터 객체들의 시간적 특징 값들을 표현과 입력방식 등에 적용될 것이다.

과거에서 현재까지 연구된 temporal 데이터의 클러스터링의 방법론에 상응하는 것은 크게 세 가지의 범주들로 그룹 되어 질 수 있다. 첫째, 쌍의 객체 또는 거리측정을 이용하는 근사기반 방법이다. 클러스터 형성과정은 쌍 객체들 사이의 유사도 또는 거리측정에 의해서 유도된다. 근사기반 방법들은 Correlation measure와 Hemming distance[3], String edit distance[4], 그리고 Longest common sequence measure[5]와 같은 String distance metrics 그리고 Dynamic time warping method[6] 등이 있다. 둘째, 각 데이터 객체들로부터 특색을 이루는 특징의 집합을 추출하여 이용하는 특징기

반이다. 특징기반의 방법들은 Fourier descriptor, Wavlet analysis[7], Piecewise polynomial function with MDL[8] 등이 있다. 그리고 마지막으로 데이터에 가장 적합한 모델의 집합을 찾는 모델기반 방법론으로 나누어 볼 수 있다. 모델기반 방법들은 각 클러스터에 대하여 분석적인 함수 또는 오토마타 기반 모델들을 가정한다. 클러스터링 과정의 목적은 데이터에 가장 적합한 모델들을 찾는 것이다. 위의 기법들은 가장 적합한 모델들에 객체들을 반복적으로 할당하며 새로운 객체 분배와 함께 모델 파라미터를 갱신한다. 이러한 과정은 수렴 시까지 반복되어진다. 모델기반의 방법들은 회귀모델, 시계열분석, 신경망, 그리고 비결정적 유한상태 오토마타인 마코프체인(MC), 은닉마코프모델(HMM) 등이 있다.

모델기반의 방법들의 각 특징을 살펴보면, 회귀모델은 시계열분석과 달리 길지 않은 데이터를 다루므로 시계열분석과 달리 해석이 쉽지 않다. 즉, 회귀모델의 형태는 동적현상의 특성묘사를 나타내기 어렵다.

신경망은 많은 부분에서 temporal 현상을 예측하는 작업에 성공적으로 적용되어 왔으나, 일반적 temporal 데이터의 클러스터링 모델링에는 적합하지 않다. 그 이유는 첫째, 모델의 구조가 알려져 있다는 것이다. 즉, 모델에서 은닉층 수, 노드들에서 사용되는 기준함수뿐만 아니라 각 층에서 노드들의 수가 정해져 있다는 것이다. 둘째, 모델의 해석을 지원하지 않는 것이다. 이는 훈련과정동안, 모델 파라메터 값들의 조정의 목적은 객관적 기준함수에 따라 산출층에 값들을 최적화하는 것이다. 그러므로, 신경망에서 노드들 사이의 연결들과 노드들과 관련된 실질적 의미가 없다는 것이다.

마코프체인 모델은 모델의 단순성 때문에 하나의 이산값을 갖는 temporal 특징으로 묘사되는 temporal 데이터의 표현 모델링에 유용하다[9]. 그러므로 일반적인 temporal 데이터의 클러스터링에 사용될 때 다음과 같은 제한점이 있다 첫째, 연속적인 값을 갖는 temporal 데이터 특징을 묘사하는 데이터모델에 적합하지 않으며 둘째, 다수의 temporal 특징에 의하여 묘사되는 데이터를 표현에 어렵다. 이러한 문제를 해결하기 위하여 각 상태에서 특징들에 대한 적합한 확률함수를 사용하여 연속적인 값을 갖는 temporal 시퀀스를 쉽게 다루며, 다수의 temporal 특징들을 가진 데이터의 묘사가 쉬운 은닉마코프모델을 사용하는 것이 일반적 temporal 데이터의 클러스터링에서는 효과적이라고 할 수 있다. 그러므로 본 연구에서는 은닉마코프모델의 모델기반으로 temporal 데이터의 클러스터링 과정에서 클러스터의 수를 결정짓는 방법과 그에 대한 BIC 측정의 유효성을 살펴본다.

III. Temporal 데이터의 클러스터링

본 연구에서의 초점은 temporal 데이터의 클러스터링에서의 클러스터의 수 결정이다. 먼저 데이터의 형식을 입력 형식으로 변환시키는 방법에 대해 설명하기로 한다.

3.1 입력 데이터 형식

각 temporal 데이터들은 데이터 객체 x_i에 대해서 K temporal 특징들로서 묘사되어지고 각 특징들은 L 길이의 temporal 값들을 갖는 것으로서 표현되어진다. 즉, K X L의 행렬로 다음과 표현되어진다.

$$x_i = \begin{pmatrix} V^i_{11} & V^i_{12} & \cdots & V^i_{1L} \\ V^i_{21} & V^i_{22} & \cdots & V^i_{2L} \\ \cdots & \cdots & V^i_{kl} & \cdots \\ V^i_{K1} & V^i_{K2} & \cdots & V^i_{KL} \end{pmatrix}$$

위의 행렬에서 V^i_{kl}은 l 시간에서 temporal 특징 k의 값을 가리킨다.

temporal 데이터의 표현과 해석에서 간단한 방법은 각 시간에서의 특징값들이 독립이라고 가정하는 것이다. 즉, 하나의 temporal 데이터 객체 x_i는 L개의 정적데이터객체들을 생산한다.

$$x^i_1 = [V^i_{11} \ V^i_{12} \cdots\cdots V^i_{1K}]$$
$$x^i_2 = [V^i_{21} \ V^i_{22} \cdots\cdots V^i_{2K}]$$

$$x_L^i = [V_{L1}^i \quad V_{L2}^i \cdots\cdots\cdots V_{LK}^i]$$

N개의 데이터가 있다고 가정한다면 $N \times L$의 데이터 객체들을 가진 새로운 데이터 집합이 만들어진다. 위에서 설명된 클러스터링 알고리즘이 이러한 데이터의 그룹에 적용된다.

3.2. Bayesian 클러스터링 방법론

N개의 데이터객체를 갖는 데이터의 최선의 분할을 표현하는 최적 클러스터의 수는 1부터 N까지 매우 다양할 수 있다. 최악의 경우 이것은 N번 실행된다. 이것은 계산적으로 매우 큰 비용이 소요된다. 그러므로 본 연구에서의 주된 아이디어는 미리 선택되어 정의된 기준함수에 의해, 최선의 분할사이즈는 하나의 클러스터 수로부터 시작하여 클러스터의 수를 하나씩 증가하여 계속 반복해 나가다가 가장 높은 기준함수의 값을 갖는 클러스터의 수가 최적의 클러스터 수가 된다. 이와 같은 특성은 뒤에 소개될 BIC 측도의 특성의 활용에 대한 근거를 제공한다.

3.2.1. 베이지안 클러스터링(Bayesian Clustering)

모델기반 클러스터링에서, 데이터는 확률분포의 혼합(Mixture)에 의해 생성되어지는 것을 가정한다. 혼합모델 M은 K개의 콤포넌트 모델들에 의해 표현되고 독립적 이산변수인 C로 표현된다. C의 각 값인 i는 λ_i 에 의해 모델 되어지는 클러스터 수를 표현한다. 데이터 $X=(x_1,\ldots,x_N)$이 주어지면, k번째 콤포넌트(k번째 클러스터모델) λ_k에 속하는 객체 x_i인 집합확률을 $f(x_i \mid \theta_k, \lambda_k)$으로 표현한다. 파라메터들은 θ_k로서 표현되어지며, 혼합모델이 주어졌을 때, 데이터의 우도(likelihood)는 식(1)과 같이 표현되어진다.

$$P(X \mid \theta, M) = P(X \mid \theta_1,\ldots,\theta_K,\lambda_1,\ldots,\lambda_K)$$
$$= \prod_{i=1}^{N} P(x_i \mid \theta_1,\ldots,\theta_K,\lambda_1,\ldots,\lambda_K)$$
$$= \prod_{i=1}^{N} \sum_{k=1}^{K} P_k \cdot f(x_i \mid \theta_k, \lambda_k) \qquad (1)$$

위에서 P_k는 콤포넌트모델 λ_k의 사전확률이다. $P_k = P(x_i \in \lambda_k),\ i=1,\ldots,N,\ k=1,\ldots,K.$ 이다.

베이지안 클러스터링은 모델기반 클러스터링 문제를 베이지안모델 선택의 문제 형태로 바꾼다. 서로 다른 콤포넌트 클러스터들을 갖는 분할들이 주어졌을 때, 목적은 가장 큰 사후확률을 갖는 가장 좋은 모델 M을 선택하는 것이다.

3.2.2. 모델선택기준(BIC)

베이즈 이론으로부터, 모델의 사후확률 $P(M \mid X)$은 $P(M \mid X) = \frac{P(M)P(X \mid M)}{P(X)}$에 의해 주어진다. $P(X)$와 $P(M)$은 데이터와 모델의 사전확률을 나타낸다. 그리고 $P(X \mid M)$은 데이터의 한계우도(Marginal Likelihood)이다. 식(2)처럼 모델 비교 목적을 위해 서로 다른 모델들 사이에서 데이터의 사전확률이 변하지 않고 유지되면

$$P(M \mid X) \propto P(M)P(X \mid M) \qquad (2)$$

그리고 모든 모델들이 사전에 같다고 가정하면, 모든 모델들에 대하여 같은 사전확률을 할당할 수 있다. 그러므로

$$P(M \mid X) \propto P(X \mid M) \qquad (3)$$

즉, 식(3)처럼 모델의 사후확률은 데이터의 한계우도(ML)에 비례한다. 그러므로 베이지안 모델 선택의 목적은 가장 큰 한계우도를 주는 모델을 선택하는 것이다.

모델 M의 파라메터 구성 θ가 주어지면, 데이터의 한계우도는 식(4)와 같이 계산되어진다.

$$P(X \mid M) = \int_{\Theta} P(X \mid \theta, M)P(\theta \mid M)d\theta \qquad (4)$$

파라메터들이 연속적인 값들을 가질 때 적분계산은 폐형해(closed form solution)를 획득하는 것은 어렵다. 한계우도를 구하기 위한 근사기법은 다양하다. 몬테-카를로 방법 그리고 라플라스 근사[10] 등이 있는데 이 둘은 매우 정확하기는 하나 계산적으로 비용이 많이 드는 것으로 알려져 있다. 따라서 본 연구에서는 정확성은 좀 떨어지나 더 효율적인 근사기법인 BIC 근사기법을 살펴볼 것이다. BIC는 다량의 데이터가 있을 때 우도함수나 사전 확률이 다변량 가우시안 분포로 근사된다는 점에서 유도된다[2]. 식(4)의 내부의 항에 로그를 취한 것을 $g(\theta)$로 정의하고 $g(\theta)$를 최대화 시키는 파라메터 구성을 $\hat{\theta}$라 할 때 이는 사후확률을 최대화하게 된다. 이를 θ의 최대사후구성(MAP)라 부른다. 여기에 2차 Taylor 다항 근사법을 적용한 후 e의 지수를 취하고 다시 원식에 대입하여 다음 식을 산출한다.

$$\log p(\theta|X, M) \approx \log P(X|\hat{\theta}, M) + \log P(\hat{\theta}|M) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A| \quad (5)$$

여기서 d는 모델에서 파라미터의 수이고, A는 $\hat{\theta}$에서 계산되는 $g(\theta)$ 음의 Hessian이다. 식(5)는 다시 데이터의 수인 N에 비례하는 항만 남기고 나머지를 제거함으로 더욱 간략화 시키면 식(6)이 유도된다.

$$\log P(X|M) \approx \log P(X|M, \hat{\theta}) - \frac{d}{2}\log N \quad (6)$$

위의 식에서 첫 번째 항인 자료의 우도 $\log P(X|M, \hat{\theta})$는 데이터를 가장 잘 설명할 수 있는 상세한 데이터의 모델을 찾도록 유도하는 성분이다. 두 번째 항인 $-\frac{d}{2}logN$은 모델 내의 파라미터 개수에 대한 패널티(penalty) 항으로 볼 수 있다.

BIC[11]는 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 최선의 모델이 구축되는 방안을 제시한다. BIC 근사기법의 장점은 수식의 의미가 직관적이면서 동시에 사전확률을 구할 필요가 없다는 점이다.

3.2.3. 클러스터 분할 선택을 위한 기준

베이지안 구조에서, 최선의 클러스터링 혼합모델 M은 가장 높은 분할사후확률(PPP), $P(M|X)$를 갖는다. 우리는 혼합모델의 한계우도 $P(X|M)$를 가지는 분할사후확률을 근사한다. 여기에서 클러스터 분할선택을 위한 한계우도의 계산에 BIC를 적용한다.

$\lambda_1, ..., \lambda_K$로서 모델된 K 클러스터를 갖는 분할에 대하여, 식(7)처럼 분할사후확률이 정의되고 BIC 근사법을 사용하여 그 값이 계산된다.

$$\log P(X|\hat{\theta}, M) = \prod_{i=1}^{N}\sum_{k=1}^{N} P_k f(x_i|x_i \in \lambda_k, \hat{\theta}, \lambda_k) \quad (7)$$

위에서 $\hat{\theta}$는 클러스터 K의 한계우도 모델 파라미터의 구성을 나타낸다. P_k는 클러스터 K의 사전확률이 되고 $f(x_i|x_i \in \lambda_k, \hat{\theta}, \lambda_k)$은 클러스터 k에 대한 모델이 주어졌을 때 데이터 x_i의 확률을 나타낸 것이다. 데이터 우도가 계산되어질 때, 데이터가 완벽하다는 것을 가정한다. 즉, 각 객체는 분할에서 알려진 하나의 클러스터에 할당된다. 최선의 모델은 전체 클러스터 분할의 복잡도와 전체데이터의 우도의 조화를 이루는 것이다.

IV. 실 험

실험을 통하여, 클러스터의 수를 결정짓는 판단기준으로 사용된 BIC 효용성을 살펴보고 데이터객체 수와 각 객체 특징의 길이 변화에 따라 정확한 클러스터의 수를 추정하는지를 알아본다. 먼저 임의의 선택된 모델로부터 생성된 데이터를 통해 최적의 클러스터의 수를 찾기 위한 클러스터링의 프로세스는 다음과 같다 :

- 선택된 모델로부터 임의의 데이터를 생성
- 병합식(agglomerative) 클러스터링[1]을 통해 파라미터들의 값들을 초기화
- Expection-Maximization(EM) 과정을 거친다.

1 계층적(hierarchical)클러스터링 알고리즘에서 계층구조를 Bottom up 방식으로 만드는 경우를 병합식(agglomerative)라 하며, Topdown 방식이면 분할식(divisive)라 한다.

- 선택된 기준에 수렴시까지 E-M단계를 반복

위의 과정을 거쳐 클러스터의 수를 하나씩 증가하며 반복하여 분할사후확률(PPP)을 BIC를 이용하여 구한다. BIC 곡선의 특징은 분할(Cluster)의 수가 증가할수록 첫 번째 항인 우도의 값이 점진적으로 증가를 하지만 클러스터의 수가 증가함에 따라 모델의 복잡도는 커지기 때문에 두 번째 항인 패널티항의 수치 또한 커지므로 두 항의 합에서는 우도값을 상쇄하는 즉, 어느 점에서 정점을 이루다가 점차 하강하는 곡선으로 표현된다. 즉 최적의 클러스터의 수에서 정점을 이루게 된다.

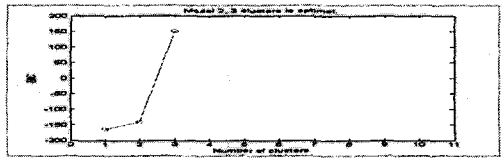
실험을 위한 혼합모델에 대한 데이터를 생성하기 위해 혼합의 파라미터들은 다음과 같은 조건들로 정의하였다. 클러스터의 수 결정 작업에 대해 데이터집합에서 객체의 수와 데이터객체의 특징의 길이에 따라 BIC 측도의 특징을 살피기 위해 생성되는 데이터객체의 집단을 네 개의 집합으로 변화를 시켜 실험을 하였다. 실험 데이터의 생성 조건은 다음과 같다.

Temporal 특징의 수 : 4
클러스터의 수 : 4
각 클러스터의 사전확률 : 0.2, 0.2, 0.1, 0.5
각 클러스터의 *Temporal* 특징별 평균값 : [2 2 2 2], [-2 -2 -2 -2], [-2 2 -2 2], [2 -2 2 -2]
각 클러스터의 공분산 : 1.5, 1, 0.75, 1

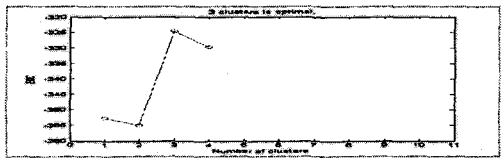
위와 같은 조건을 적용할 실험 집단을 데이터객체 수와 특징의 길이에 따라 각각 4개의 데이터집합을 생성하여 실험이 이루어지도록 구성하였다:

실험 별 데이터 객체의 수 : 10, 20, 30, 40
데이터 객체의 특징의 길이 : 50, 100, 200, 300

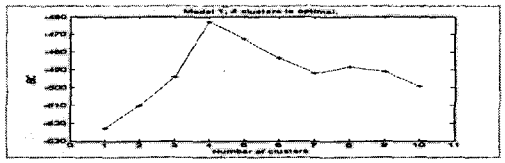
먼저 위의 조건들에 의하여 생성된 데이터 객체의 수에 따른 집합에 대하여, 병합식 클러스터링을 통해 파라미터들을 초기화하여 이를 토대로 Maximization 단계에 의해 갱신된 모델들에 대하여 BIC를 이용한 우도값의 측정의 결과는 [그림 1]과 같다. 위에서 설명한 BIC 곡선을 확인할 수가 있다. 우선 데이터의 객체수가 적은 10, 20개의 데이터객체 실험군은 클러스터의 수를 3개로 추정을 하였으며, 상대적으로 데이터객체가 충분한 30, 40개의 데이터객체를 가진 실험군은 정확히 4개의 클러스터의 수를 추정하는 것을 볼 수 있다.



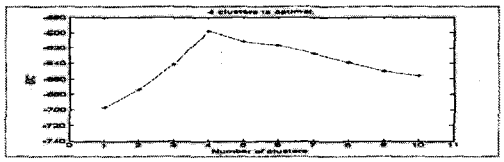
(a) 10 data objects



(b) 20 data objects



(c) 30 data objects



(d) 40 data objects

그림 1. 객체 수에 따른 클러스터의 수 결정 BIC 곡선

즉, 객체의 사이즈의 변화에 따라 클러스터의 수의 추정이 차이가 있는데 데이터객체의 수가 상대적으로 적은 데이터객체에서는 부정확한 클러스터의 수를 추정하며 데이터객체의 수가 충분하면 즉, 30개 이상의 데이터 객체집합들에서는 정확한 클러스터의 수를 추정한다는 것을 알 수 있다.

그림 2. 객체 길이에 따른 클러스터 수 결정 BIC 곡선

두 번째, 데이터객체 특징의 길이가 다른 4개의 데이터 집합에 대하여 BIC 측정의 결과는 [그림 2]와 같다. 특징의 길이에 따라 추정결과가 달라지는 것을 알 수가 있다. 즉, 특징의 길이가 50, 100인 데이터집합에서는 클러스터의 수가 2개에서 BIC의 값이 최고가 되었으며, 길이가 200, 300인 데이터객체의 집합은 정확히 4개의 클러스터에서 BIC의 값이 최고로 추정하는 것을 볼 수 있다. 위의 두 실험집단을 통해 본 결과 데이터객체의 수와 특징의 길이가 충분하다면 정확한 클러스터의 수를 추정함을 알 수가 있었다.

본 실험결과에서 보듯이 자료가 충분히 제공되었을 경우 BIC 값은 초기의 값으로부터 증가하는 방향으로 값이 변화하다가 어느 시점에서는 하강하는 방향으로 진행하게 된다. 따라서 최적의 클러스터 수를 결정하는데 있어 탐색 방법은 BIC 값을 추적하면서 값이 증가하는 방향에서 감소하는 방향으로 변화하는 지점을 최적을 점으로 결정할 수 있다. 물론 데이터의 개수가 적은 경우(예로, [그림 1]의 b) 최적의 클러스터 수 예측이 잘 못되는 것은 물론 BIC 값의 변화에 있어서도 일관성이 없다는 점을 발견할 수 있다. 즉, BIC 값의 증감이 일관적이지 못하여 증감 방향이 무작위 적인 점을 확인할 수 있다.

V. 결 론

일반적으로 클러스터링 문제의 대표적인 문제 중의 하나는 데이터가 고유하게 갖고 있는 클러스터의 수에 대한 추정 문제이다. 본 연구는 모델 기반의 temporal 데이터를 대상으로 한 클러스터링 과정에서 최적의 클러스터의 수를 결정짓는 과정인 BIC 측도에 대하여 고찰했다. 실험결과는 BIC 측도가 일반적으로 클러스터의 수를 정확하게 추정하는 결과를 보여주고 있으나 데이터객체의 수와 특징의 길이에 영향을 받는 것을 확인하였다. 즉, 객체의 수와 특징의 길이가 충분한 데이터가 확보되는 경우 클러스터의 수를 정확하게 예측할 수 있었으나 그렇지 않은 경우 클러스터 수의 추정에 있어 오류가 있을 수 있음을 확인하였다. 이점은 BIC 측도의 유도가 자료의 개수가 많은 경우에 다변량 가우시안 분포로 근사할 수 있다는 점에서 볼 때 당연한 결과로 예측된 것이다.

향 후 연구해야 할 내용은 과연 어는 정도 개수의 데이터가 BIC의 유효한 한계인지를 추정하는 방법에 대한 연구이며, 추정되어진 클러스터 수에 대해 개별 클러스터에 대한 모델을 생성하는 문제이다. 이상의 연구가 이루어진다면 일반적인 용도의 temporal 데이터의 클러스터링과 모델링 방법론을 개발하는 것으로서 이러한 방법론이 복잡하고 동적인 시스템들과 프로세스들을 가진 현상들을 이해하는 것에 도움이 될 것이다. 그러므로 앞으로 각 클러스터에 대한 초기모델의 생성과 이를 통한 실세계의 다양한 데이터에 대하여 적용을 통한 연구를 해야 할 것이다.

참 고 문 헌

[1] P. Cheeseman and J. Stuze, "Bayesian classification(autoclass)," Advanced in Knowledge Discovery and Data Mining, AAAI-MIT press, pp.153-180, 1996.

[2] R. E. Kass and A. E. Raftery, "Bayes factor," Journal of American Statistical Association90, pp.773-795, 1995(6).

[3] A. K. Jain and D. C. Dube, *Algorithms for clustering data,* Prentice Hall, 1988.

[4] T. Okuda, E. Tanara, and T. Kasai, "A method for the correction of garbled words based on the levenshtein metric," IEEE Transaction on Computers C25, 2, pp.172-177, 1976(2).

[5] D. S. Hirschberg, "Algorithms for longest common subsequence problem," Journal of Association of Computer Machine 24, pp.664-675, 1977.

[6] T. Oates, "Identifying distinctive subsequences in multivariate time series by clustering," Proceedings of the Sixteenth International

Conference on Machine Learning, 1999.

[7] Y. Huhtala, J. Karkkinen, H. Toivonen, and N. R, "Mining for similarity in aligned time series using wavlets," Proceedings of SPIE on Data Mining and Knowledge Discover: Theory, Tools, and Technology, 1999.

[8] S. ManGanaris, "Learning to classify sensor data," IJCAI'95 Workshop on Manchine Leanring in Engineering, 1995.

[9] P. Sebastiani, M. Ramoni, P. Cohen, J. Warwick, and J. Davis, "Discovering dynamic using bayesian clustering," Advances in Intelligent Data Analysis, Springer-Verlag, D. J. Hand, J. N. Kok, and M. R. Berthold, Eds. Berlin, Springer-Verlag, pp.199-210, 1999(8).

[10] D. Heckerman, D. Geiger, and D. M. Chikering. "A tutorial on learning with Bayesian Network," Machine Learning, 20, pp.197-243, 1995.

[11] S. S. Chen and P. S. Gopalkrishana, "Speaker, enviroment, and channel change detection and clustering via the Bayesian information criterion," Proceedings of the IEEE Interational Conference on Vol.2, pp.645-648, 1998(5).

저 자 소 개

전 진 호(Jin-Ho Jeon) 정회원

- 1994년 : 관동대학교 경영학과(경영학사)
- 1998년 : 명지대학교 경영정보학과(경영학석사)
- 2003년 2월 : 단국대학교 전자계산학과(박사과정 수료)
- 2003년 3월~현재 : 관동대학교 경영정보학부 겸임조교수

<관심분야> : 기계학습, 데이터마이닝

조 영 희(Young-Hee Cho) 정회원

- 1985년 : 단국대학교 전자계산학과(학사)
- 2000년 : 단국대학교 전자계산학과(석사)
- 2005년 : 단국대학교 전자계산학과(박사과정)

<관심분야> : 기계학습, 데이터마이닝, Ontology

이 계 성(Gye-Sung Lee) 정회원

- 1980년 : 서강대학교 전자공학과(학사)
- 1982년 : 한국과학기술원 전자계산학과(석사)
- 1994년 : Vanderbilt University 전자계산학과(공학박사)
- 1994년~1996년 : 대구대학교 전산정보학과 전임강사
- 1996년~현재 : 단국대학교 컴퓨터과학 전공 부교수

<관심분야> : 기계학습, 데이터마이닝, 바이오인포메틱스, 비디오마이닝