시계열데이터의 모델기반 클러스터 결정

Determining on Model-based Clusters of Time Series Data

전진호, 이계성
단국대학교 전자계산학과

Jin-Ho Jeon(jhjy@dankook.ac.kr), Gye-Sung Lee(gslee@dku.edu)

요약

대부분의 실세계의 시스템들, 즉 경제, 주식시장, 의료분야 등의 많은 시스템들은 동적이며 복잡한 현상을 갖는다. 이러한 특징들의 시스템을 이해하는 전형적인 방법은 시스템행위에 대한 모델을 세우고 분석하는 것이다. 본 연구에서는 실세계의 동적시스템에서 발생되는 시계열데이터들에 대하여 최적의 클러스터를 형성하기 위한 방법을 연구한다. 먼저 클러스터 수를 결정하는 기준으로 베이지안정보기준(BIC : Bayesian Information Criterion)근사법의 활용도를 검증하고 데이터 크기와 베이지안정보기준값의 상관관계를 파악함으로 탐색 효율을 높이는 방안을 제안하며 클러스터링 과정으로 모델기반과 유사기반의 방법론을 비교 확인하여 본다. 실제의 시계열데이터(주가)에 대해 실험을 시행하였고 베이지안정보기준 근사 측도는 데이터의 크기에 따라 파티션의 사이즈를 정확히 추정하는 것을 확인하였으며 또한 유사기반의 방식보다 모델기반의 방법론이 클러스터링에서 더 나은 결과를 갖는 것을 확인하였다.

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

Abstract

Most real word systems such as world economy, stock market, and medical applications, contain a series of dynamic and complex phenomena. One of common methods to understand these systems is to build a model and analyze the behavior of the system. In this paper, we investigated methods for best clustering over time series data. As a first step for clustering, BIC (Bayesian Information Criterion) approximation is used to determine the number of clusters. A search technique to improve clustering efficiency is also suggested by analyzing the relationship between data size and BIC values. For clustering, two methods, model-based and similarity based methods, are analyzed and compared. A number of experiments have been performed to check its validity using real data(stock price). BIC approximation measure has been confirmed that it suggests best number of clusters through experiments provided that the number of data is relatively large. It is also confirmed that the model-based clustering produces more reliable clustering than similarity based ones.

■ Keyword : | Time Series Data | Model-based | Clustering | BIC |

* 본 연구는 2006학년도 단국대학교 대학연구비의 지원으로 연구되었습니다.

접수번호 : #070119-001
접수일자 : 2007년 01월 19일
심사완료일 : 2007년 05월 25일
교신저자 : 전진호, e-mail : jhjy@dankook.ac.kr

I. 서론

대부분의 실세계의 시스템들, 즉 경제분야, 의료분야, 과학 및 공학 현상을 분석하는 시스템들은 동적이며 복잡한 현상을 갖는다. 이러한 특징들의 시스템들을 이해하는 전형적인 접근법은 시스템 행위에 대한 모델을 세우고 분석하는 것이다.

본 연구에서는 최근 들어 폭발적인 증가를 보이고 있는 상업적이거나 과학적인 실세계의 동적인 시스템에서 발생되어 관측기간 동안 의미 있게 변하는 시간적 특징들로 묘사되는 데이터들을 대상으로 연구 분석한다. 주가 데이터, 환율데이터, 기업성장률 데이터, 기온과 같은 날씨 데이터, 기기 측정 데이터 등에 내재하는 현상을 쉽게 이해하는데 필요한 최적 클러스터 집합의 구성을 찾는 모델 기반 클러스터링 방법론을 살펴본다. 대용량의 데이터에서는 각각의 데이터를 요약하는 것보다 전체를 유사한 클러스터로 구분하여, 복잡한 전체 데이터 대신에 클러스터들을 분석함으로서 전체 데이터에 대한 의미 있는 정보를 얻을 수 있다. 이러한 클러스터링 문제는 과거 정적특징[1]들에 의해 묘사된 데이터에 많은 연구가 진행되어 왔다. 그러나 시계열 데이터의 클러스터링은 정적인 데이터의 클러스터링의 문제보다 복잡하다. 이유는 데이터의 차원이 동적인 경우에서는 정적인 경우보다 크다. 따라서 클러스터의 정의와 해석의 복잡도는 동적인 데이터가 갖는 차수의 크기에 의해 증가되기 때문이다[2].

시계열 데이터의 모델기반 클러스터링 알고리즘은 크게 두 가지의 과정으로 나누어 볼 수 있다. 첫째, 데이터에 대하여 최적의 클러스터 수 결정과 클러스터링을 통한 최적 집합을 찾는 것이다. 둘째, 각각의 클러스터에 가장 적합한 모델을 생성하는 것이다. 본 연구에서는 위의 두 가지 과제 중에서 첫 번째 과제인 최적의 클러스터 수를 결정하는 과정에 대해서 살펴보고자 한다.

최적의 클러스터 수를 결정할 베이지안정보기준[3] 측정에 대해서 고찰과 실험을 통해 유효성을 살펴보고 클러스터링 과정으로는 모델 기반과 유사기반의 방법을 비교 확인하여 본다. 결과에서는 베이지안정보기준 근사 측도는 데이터의 크기가 비교적 클 경우에 정확한 클러스터 수를 추정함을 확인하였으며 실제의 주가데이터에 적용하였을 때 모델기반의 클러스터링은 유사기반의 방법론보다 더 나은 클러스터링 결과를 산출하는 것을 확인하였다.

II. 관련연구

클러스터링은 주어진 데이터 집합에 대하여 서로 유사성을 가지는 몇 개의 클러스터로 구분해 나가는 과정으로서, 하나의 클러스터에 속하는 데이터들 간에는 서로 다른 클러스터 내의 데이터와는 구분되는 유사성을 갖는다. 최근에는 다양한 다차원의 데이터를 효율적으로 분류해 나가기 위한 방안으로 연구되고 있다.

과거에서 현재까지 연구된 시계열 데이터의 클러스터링 방법론은 크게 두 가지의 범주로 구분할 수 있다. 유사기반 방식과 모델기반 방식의 접근법이다.

유사기반 방식은 두 객체 간 또는 두 시퀀스간의 데이터 포인트 간의 거리측정을 이용하는 방법이다. 클러스터 형성과정은 두 시퀀스간의 유사도 또는 거리측정에 의해서 유도된다. 유사기반 방법들은 해밍거리(Hamming distance)[4], 스트링편집거리(String edit distance)[5], 유클리디안 거리[6], 구간상수근사(Piecewise constant approximations)(PCA)[7], 기호집합근사(Symbolic aggregate approximation)(SAX)[7] 등이 있다. 해밍거리는 두 시퀀스의 데이터 항목들 사이에서 불일치하는 항목들의 개수를 세는 방식이다. 스트링편집거리는 하나의 시퀀스를 비교 대상의 시퀀스로 변환하는데 드는 총 작업량(삽입, 삭제, 이동 작업 등)을 의미하며, 이를 점수로 나타내어 점수가 낮을수록 시퀀스를 변환하는데 적은 노력이 필요하므로 유사도가 높다고 한다. 가장 일반적으로 사용되어지는 유클리디안 거리는 데이터의 항목 값들에 따라 이진값을 갖는 항목들의 점으로 대응시켜 유사도를 계산한다. 좀 더 진화된 유사기반 방법으로서 계산 복잡도를 줄이기 위하여 데이터 시퀀스에 내재되어 있는 의미를 손실 없이 압축하여 표현하는 방식으로 구간상수근사(PCA),

기호집합근사(SAX)등이 있다. PCA와 SAX방법론은 3.2절에서 자세히 설명한다. 유사기반에서 나타날 수 있는 문제점은 두 시계열 데이터들 사이에 유사도가 어떤 측정 방법에서는 높게 나올 수 있지만, 다른 유사도 측정 방법을 이용하면 낮게 나올 수 있다는 점과 시계열데이터가 내포하는 핵심적 현상은 거리함수를 가지고 얻는 것이 상대적으로 어렵다는 점이다.

모델기반 방법들은 각 클러스터에 대하여 분석적인 함수 또는 오토마타 기반 모델들로 가정한다. 클러스터링 과정의 목적은 데이터에 가장 적합한 모델들을 찾는 것이다. 모델 기반의 방법들은 회귀모델, 신경망, 그리고 비결정적 유한상태 오토마타인 마코프체인(MC), 은닉마코프모델(HMM)[8] 등이 있다. 모델 기반의 방법들의 각 특징을 살펴보면, 회귀모델은 상대적으로 길지 않은 데이터를 다루므로 해석이 쉽지 않기 때문에 동적 현상의 특성을 나타내기 어렵다. 신경망은 많은 부분에서 시계열 현상을 예측하는 작업에 성공적으로 적용되어 왔으나, 일반적 시계열 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 조정목적은 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 템포랄(temporal)특징으로 묘사되는 시계열 데이터의 표현 모델링에 유용하다고 알려져 있다[9]. 그러므로 일반적인 시계열 데이터가 클러스터링에 사용될 때 다음과 같은 제한점이 있다 첫째, 마코프모델은 연속적인 값을 갖는 시계열 데이터의 특징을 묘사하는 데 적합하지 않으며 둘째, 다수의 시계열 특징에 의하여 묘사되는 데이터 표현이 어렵다. 이러한 문제를 해결하기 위하여 각 상태에서 특징들에 대한 적합한 확률함수를 사용하여 연속적인 값을 갖는 시계열 시퀀스를 쉽게 다루며, 다수의 시계열 특징들을 가진 데이터의 묘사가 쉬운 은닉마코프모델을 사용하는 것이 일반적 시계열 데이터의 클러스러링에서는 효과적이라고 할 수 있다. 그러므로 시계열데이터의 클러스터링은 유사기반 방식보다 모델기반 방식이 더욱 적합하며 본 연구에서는 은닉마코프모델의 모델 기반으로 시계열 데이터의 클러스터링[11][12] 과정에서 최적의 클러스터를 결정짓는 방법의 유효성을 살펴본다.

III. 클러스터링 방법론

1. 베이지안(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 베이지안정보기준 함수에 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 베이지안정보기준 측도의 특성의 활용에 대한 근거에 의한 것이다.

1.1 베이지안 클러스터링(Bayesian Clustering)

모델기반 클러스터링에서, 데이터는 확률분포의 혼합(Mixture)에 의해 생성되어지는 것을 가정한다. 혼합모델 M은 K개의 컴포넌트 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 컴포넌트(k번째 클러스터모델) λ_k에 속하는 객체 x_i인 집합확률을 $f(x_i \mid \Theta_k,\lambda_k)$으로 표현한다. 파라미터들은 Θ_k로서 표현되어지며, 혼합모델이 주어졌을 때, 데이터의 우도(likelihood)는 식(1)과 같이 표현되어진다.

$$P(X \mid \Theta, M) = P(X \mid \Theta_1,\ldots,\Theta_K,\lambda_1,\ldots,\lambda_K)$$

$$= \prod_{i=1}^{N} P(x_i \mid \Theta_1, \ldots, \Theta_K, \lambda_1, \ldots, \lambda_K)$$
$$= \prod_{i=1}^{N} \sum_{k=1}^{K} P_k \cdot f(x_i \mid \Theta_k, \lambda_k) \quad (1)$$

위에서 P_k는 컴포넌트모델 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 컴포넌트 클러스터들을 갖는 분할들이 주어 졌을 때, 목적은 가장 큰 사후확률을 갖는 가장 좋은 모델 M을 선택하는 것이다.

1.2 클러스터 평가 위한 베이지안 모델 선택

우리의 목적은 데이터에 대한 최적의 혼합모델 M을 찾는 것이다. 최적의 클러스터링 혼합모델 M은 가장 높은 분할사후확률(PPP), $P(M \mid X)$를 갖는다. 우리는 혼합모델의 한계우도 $P(X \mid M)$를 분할사후확률에 근사시킨다. 여기에서 클러스터 분할선택을 위한 한계우도의 계산에 베이지안정보기준을 적용한다. 베이지안 정보기준은 다량의 데이터가 있을 때 우도함수나 사전확률이 다변량 가우시안 분포로 근사된다는 점에서 유도되어진다[10].

$\lambda_1, \ldots, \lambda_K$로서 모델된 K클러스터를 갖는 분할에 대하여, 식(2)처럼 분할사후확률이 정의되고 베이지안 정보기준 근사법을 사용하여 그 값이 계산된다.

$$\log P(X \mid \hat{\Theta}, M) = \prod_{i=1}^{N} \sum_{k=1}^{N} P_k f(x_i \mid x_i \in \lambda_k, \hat{\Theta}, \lambda_k) \quad (2)$$

위에서 $\hat{\Theta}$는 클러스터 K의 한계우도 모델 파라미터의 구성을 나타낸다. P_k는 클러스터 k의 사전확률이 되고 $f(x_i \mid x_i \in \lambda_k, \hat{\Theta}, \lambda_k)$은 클러스터 k에 대한 모델이 주어졌을 때 데이터 x_i의 확률을 나타낸 것이며 은닉마코프모델의 전향절차에 의하여 계산된다. 베이지안정보기준을 적용하여 K 클러스터들을 갖는 분할에 대한 분할사후확률은 식(3)과 같다.

$$\log P(X \mid M) \approx \log P(X \mid \hat{\Theta} M) - \frac{d}{2} \log N$$
$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \log P_k + \sum_{i=1}^{N} \sum_{k=1}^{K} f(x_i \mid \hat{\Theta}_k, \lambda_k) - \frac{K + \sum_{k=1}^{K} d_k}{2} \log N \quad (3)$$

식(3)에서 첫 번째 항 $\log P(X \mid \hat{\Theta} M)$는 데이터에 대한 모델의 우도값을 나타내며, 두 번째 항 $\frac{d}{2} \log N$는 모델 복잡도에 따른 패널티항이다. d_k는 클러스터내의 의미있는 파라메터의 수를 나타낸다. 데이터우도가 계산되어질 때, 데이터가 완벽하다는 것을 가정한다. 즉, 각 객체는 분할에서 알려진 하나의 클러스터에 할당된다. 최선의 모델은 전체 클러스터 분할의 복잡도와 전체 데이터의 우도의 조화를 이루는 것이다.

2. 유사기반 클러스터링

유사기반 방식은 두 객체간 또는 두 시퀀스간의 데이터포인트 간의 거리측정을 이용하는 방법이다. 클러스터 형성과정은 두 객체간 또는 두 시퀀스간의 거리측정에 의해서 유도된다. 본 절에서는 유클리디안 거리, 구간상수근사기법(PCA), 기호집합근사기법(SAX)방법을 살펴보고자 한다.

시계열 시퀀스들 간의 가장 일반적인 거리측정은 유클리디안 거리이다. 이는 같은 길이의 두 시퀀스 Q와 C가 주어지면 거리측정은 식(4)과 같다. 여기서 q_i, c_i는 두 시퀀스의 데이터 포인트를 가리킨다.

$$D(Q, C) \equiv \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2} \quad (4)$$

위의 방식과 달리, PCA, SAX 방식은 차원축소를 통하여 원래 시퀀스들에 대응하는 변형된 시퀀스들 간의 거리 측정을 통해 최소계산을 보증한다.

PCA방식은 n 길이의 시퀀스 C를 벡터 $\bar{C} = \bar{C}_1, \ldots, \bar{C}_w$ 로 표현한다. $\bar{C}$ 의 i 번째 요소는 다음의 식(5)에 의하여 계산되어진다.

$$\overline{C}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} C_j \qquad (5)$$

데이터는 n차원으로부터 시퀀스의 축소를 통해 같은 사이즈 프레임 w로 나누어지며 각 프레임은 데이터의 평균값으로 표현되어진다. 이 값들은 차원이 축소된 표현이 된다. PCA[7]에 의해 변형된 시퀀스들 간의 거리 측정은 다음 식(6)에 의하여 계산되어진다.

$$D(\overline{Q}, \overline{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (\overline{q_i} - \overline{c_i})^2} \qquad (6)$$

SAX[7] 방식은 제일 먼저 원래의 시퀀스를 PCA 방식을 적용하여 차원축소를 거쳐 일정 프레임으로 구간을 설정한 뒤 시퀀스를 정규분포로 표현하여 임의의 구간으로 나눈 후 기준값에 의해 기호화 하며 이러한 방식으로 변형되어진 시퀀스들 간의 거리측정은 다음의 식(7)에 의하여 계산되어진다.

$$D(\overline{Q}, \overline{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(\overline{q_i} - \overline{c_i}))^2} \qquad (7)$$

IV. 실험

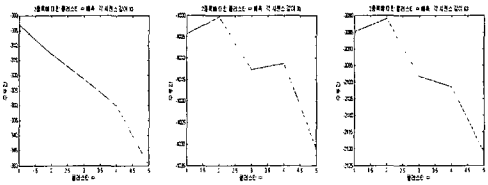
실험을 통하여, 두 가지 요소를 확인한다. 첫 번째 클러스터의 수를 결정짓는 판단기준으로 사용된 베이지안정보기준의 효용성을 살펴보고, 두 번째 추정된 클러스터 수를 통해 모델기반과 유사기반 방법론의 클러스터링 결과를 비교 확인하여 본다.

1. 베이지안정보기준을 이용한 클러스터 수 결정

베이지안정보기준의 효용성을 살펴보기 위한 실험으로 데이터 시퀀스의 수와 각 시퀀스의 길이 변화에 따라 정확한 클러스터의 수를 추정하는지를 확인한다.

실험데이터의 생성은 실제의 주가데이터[1]를 통해 2개의 클러스터 모델을 생성 후[2] 해당 모델로부터 생성된 임의의 여러 데이터 시퀀스들을 대상으로 실험하였다.

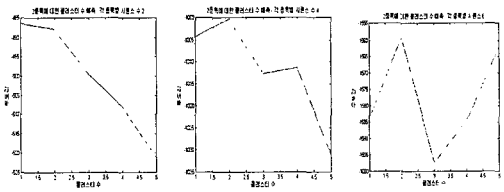
먼저 시퀀스의 길이에 따라 클러스터를 정확히 추정하는지 결과를 살펴보기 위해 두 모델에서 생성된 시퀀스의 길이는 10, 30, 60 그리고 각 시퀀스의 수는 각 모델별로 4개의 시퀀스로 하였다. [그림 1]와 [그림 2]에서의 X축은 클러스터의 수를 나타내며 Y축은 각 클러스터의 수에서의 우도값을 나타낸다.



(a)길이10 (b)길이30 (c)길이60

그림 1. 데이터 시퀀스 길이에 따른 BIC 추정결과

먼저 [그림 1]의 (a)를 보면 데이터 시퀀스의 길이가 10인 경우는 클러스터 수가 1개일 경우에 가장 큰 우도값을 갖기 때문에 클러스터의 수를 1개로 추정을 하였다. [그림 1]의 (b)(c)를 보면 데이터 시퀀스의 길이가 30, 60인 경우에는 클러스터의 수가 2개인 경우에서 가장 큰 우도 값들을 갖는다. 즉, 실험을 위하여 생성한 클러스터 모델의 수인 2개의 클러스터 수를 정확히 추정하는 것을 확인하였다.



(a)시퀀스 2 (b)시퀀스 4 (c)시퀀스 6

그림 2. 데이터 시퀀스 수에 따른 BIC 추정결과

1 주식데이터 중 업종별 주가지수를 선택하였으며 전기전자와 유통업종의 데이터를 사용하였다. 모델생성에 사용된 데이터 기간은 다음과 같다.
전기전자업종:2005.05.03-2005.11.16
유통업종:2005.05.03-2005.11.16

2 두 모델의 생성은 각각의 주가데이터들에 대하여 베이지안정보기준(BIC)을 적용하여 각 데이터들에 대한(즉, 모델에 대한) 상태수를 추정하였으며 추정된 상태수를 기준으로 하여 Baum-Welch 기법을 적용하여 모델의 파라메터(초기, 전이, 방출확률)을 추정함으로서 모델들을 생성하였다.

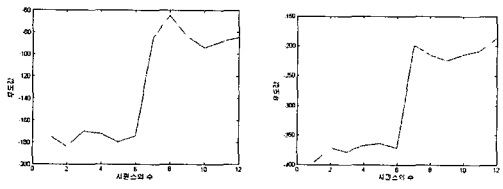
두 번째는 데이터 시퀀스의 길이는 동일하며 수를 달리 한 경우, 즉, 각 클러스터 모델별로 시퀀스의 수를 2, 4, 6인 경우는 [그림 2]의 결과와 같다.

각 모델별 시퀀스가 2개인 경우인 (a)에서는 클러스터의 수가 1개일 때 가장 큰 우도값을 가지므로 클러스터의 수를 1개로 추정한다. 그러나 각 모델별 시퀀스의 수가 4, 6개인 (b),(c)인 경우는 클러스터의 수가 2개인 경우에 가장 큰 우도값을 보여주므로 실험데이터를 생성 위한 클러스터 모델의 수인 2개를 정확히 추정하는 것을 확인하였다.

위의 두 실험을 통해 본 결과 데이터 시퀀스의 수와 길이가 충분하다면 베이지안정보기준 값은 초기의 값으로부터 증가하는 방향으로 값이 변화하다가 어느 시점에서부터는 하강하는 방향으로 진행하는 형태를 가지며 정확한 클러스터의 수를 추정함을 확인하였다.

2. 상태수를 고려한 모델 기반 클러스터링

베이지안정보기준에 의해 추정된 클러스터 수를 통해 모델기반의 클러스터링 결과를 확인한다. 미리 추정된 클러스터 수에 따른 시퀀스들의 할당 방식은 k-means에서와 같이 주어진 클러스터에 대하여 최대의 우도값을 갖는 시퀀스를 할당하는 방식을 적용한다. 두 모델[3]에서 각 모델별로 6개의 시퀀스 즉, 12개의 시퀀스를 생성하였으며 주어진 클러스터들에 대하여 일반적인 k-means 알고리즘을 적용하였다. [그림 3]에서의 X축은 두 모델로부터 생성된 시퀀스들의 번호이며 Y축은 각 시퀀스별로 해당 모델에 대한 우도값을 나타낸다.



(a) 시퀀스 길이 30 (b)시퀀스 길이 60

그림 3. 두 모델의 시퀀스의 길이에 따른 분할결과

3 각주1에서의 실험 데이터와 같은 데이터임. 즉, 전기전자업종의 모델로부터 1번-6번 시퀀스를 생성하였으며 유통업종의 모델로부터 7번-12번 시퀀스를 생성하였다.

[그림 3]의 (a),(b)처럼 각 모델에 대하여 유사한 우도값을 갖는 6개의 시퀀스씩 나누어지는 것을 확인할 수 있다. 즉, 전기전자업종의 모델로부터 생성된 1번-6번 시퀀스들이 유사한 우도값을 갖으며 유통업종의 모델로부터 생성된 7번-12번 시퀀스들이 유사한 우도값을 갖는다. [표 1][표 2]는 각 클러스터별로 할당된 시퀀스들의 우도값을 보여준다. 이를 통해 시퀀스들이 각 클러스터들에 정확히 클러스터링된 것을 확인할 수 있다.

표 1. 길이30인 시퀀스들의 클러스터링 결과

시퀀스 No.	모델1의 우도값	시퀀스 No.	모델2의 우도값
시퀀스 1	-173.8042	시퀀스 7	-84.6596
시퀀스 2	-184.0256	시퀀스 8	-64.6709
시퀀스 3	-169.8895	시퀀스 9	-84.5333
시퀀스 4	-172.2585	시퀀스 10	-93.7357
시퀀스 5	-179.8379	시퀀스 11	-88.5668
시퀀스 6	-174.1618	시퀀스 12	-84.7166

표 2. 길이60인 시퀀스들의 클러스터링 결과

시퀀스 No.	모델1의 우도값	시퀀스 No.	모델2의 우도값
시퀀스 1	-396.1441	시퀀스 7	-198.6770
시퀀스 2	-371.4224	시퀀스 8	-215.5681
시퀀스 3	-378.2508	시퀀스 9	-223.9802
시퀀스 4	-367.7189	시퀀스 10	-214.6081
시퀀스 5	-364.1348	시퀀스 11	-207.8860
시퀀스 6	-372.0818	시퀀스 12	-186.0929

위의 결과들은 두 클러스터일 경우를 보여주며, 클러스터가 다수일 경우에는 시퀀스들이 복수의 클러스터들에서 유사한 우도값을 갖는 경우가 있다. [그림 4]를 보면 각 업종별 3 모델[4]에서 생성되어진 18개의 시퀀스(각 모델별 6개 시퀀스)들의 각 모델에 대한 상태 수를 고려하지 않은 우도값에 따른 일반적 k-means 방식에 의한 클러스터링의 결과를 볼 수 있다. 6개의 시퀀스는 해당 모델에 대하여 유사한 우도값을 보여 정확히 분할되어지지만 나머지 12개의 시퀀스는 유사한 우도값을 보여주며 시퀀스가 생성된 모델이 아닌 다른 모델에 대하여 높은 우도값을 보여주는 경우도 있다. 좀 더 정확한 클러스터

4 전기전자, 유통, 건설업종을 말함.

링 위해 시퀀스가 두 모델에 대하여 유사한 우도값을 가질 경우 클러스터의 모델 상태수와 시퀀스의 상태수를 비교하여 같은 상태수를 같는 모델에 할당할 경우 더욱 정확한 클러스터링 결과를 얻을 수 있다. [그림 5]는 각 업종별 모델에서의 상태의 수를 나타낸다.

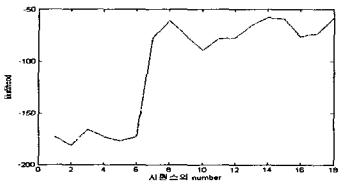
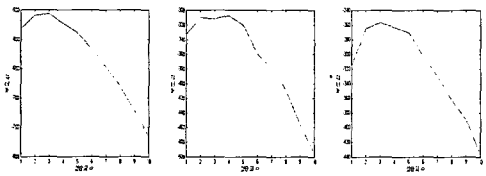


그림 4. 상태수 고려 안한 경우(우도값에 의한 분할)



(a) 전기전자 (b) 유통 (c) 건설

그림 5. 실험에 사용된 업종별 모델의 상태 수

표 3. 상태수 고려 후 클러스터별 우도값 합

업종 모델별 상태의 수 조합	각 모델(클러스터)에 대한 우도값	우도값 합
우도값만 고려 경우	(-1044)+(-458.5627)+(-389.9725)	-1892.5352
우도값+상태수 고려	(-1044)+(-421.9409)+(-389.9725)	-1855.9134

[표 3]에서 2행은 각 시퀀스별로 각 모델들에 우도값만 고려하여 할당한 경우이며 3행은 우도값과 시퀀스들의 상태수를 고려하여 할당한 후의 합을 보여주는데 상태수를 고려하기 전(-1892.5352)보다 높은 우도값(-1855.9134)을 보여주기 때문에 상태수를 고려한 방식이 더 정확한 클러스터링의 결과임을 보여준다.

3. 유사기반 클러스터링

앞 절에서 생성된 같은 실험데이터를 통해 각 시퀀스별로 PCA, SAX방식에 근거하여 시퀀스를 변환 후 임의의 하나의 시퀀스를 선정하여 기준으로 정하고 나머지 모든 시퀀스에 대한 거리측정의 결과의 일부분을 [표 4][5]와 [표 5]에서 보여주고 있다.

표 4. PCA 거리 측정 테이블

시퀀스 No	1 s	7 s	13 s	시퀀스 No	1 s	7 s	13 s
1 s	0	4.69	2.79	10 s	2.93	2.98	2.66
2 s	2.98	4.21	2.90	11 s	4.90	2.90	3.29
3 s	2.10	4.07	1.99	12 s	5.40	2.00	4.13
4 s	2.62	3.75	2.17	13 s	2.79	2.95	0
5 s	3.31	5.08	3.10	14 s	3.33	3.76	2.54
6 s	1.96	4.34	3.27	15 s	3.83	3.44	3.03
7 s	4.68	0	2.95	16 s	4.30	2.10	2.37
8 s	3.79	2.43	2.22	17 s	4.68	3.17	3.22
9 s	5.85	3.17	3.91	18 s	4.44	2.84	1.01

표 5. SAX 거리 측정 테이블

시퀀스 No	1 s	7 s	13 s	시퀀스 No	1 s	7 s	13 s
1 s	0	3.84	1.16	10 s	2.32	2.01	1.16
2 s	1.64	2.84	1.16	11 s	4.64	2.84	2.59
3 s	0	3.84	0	12 s	5.05	1.64	3.48
4 s	1.16	3.48	1.16	13 s	1.16	2.01	0
5 s	2.01	4.78	1.64	14 s	2.84	3.07	1.16
6 s	0	3.48	1.16	15 s	3.66	1.16	2.01
7 s	3.84	0	2.01	16 s	4.18	1.64	1.64
8 s	3.28	1.16	1.16	17 s	4.18	2.01	1.64
9 s	5.05	2.84	2.84	18 s	2.01	3.66	1.16

실험에 사용된 시퀀스는 1번부터 6번이 같은 모델에서 생성되었고 7번부터 12번, 13번부터 18번이 각각 다른 모델에서 생성된 시퀀스들이다. 거리측정의 결과를 살펴보면 수치가 작을수록 유사한 시퀀스를 보여주는데 같은 모델이 아닌 다른 모델에서 생성된 시퀀스가 유사한 형태를 보여주는 작은 값들이 많이 나타나고 있다. 위의 결과들에 따라 실험데이터의 각 시퀀스를 생성한 모델에 클러스터링 결과는 [표 6]과 같다.

5 [표 4][표 5]에서 1과 5의 열은 실험데이터의 시퀀스 수를 나타내며 2,3,4와 6,7,8열은 1번과 7번과 13번의 시퀀스를 기준일 경우 다른 시퀀스들과의 거리측정의 결과이다.

표 6. PCA & SAX 의 분할의 정확도

유사도 방식	1 클러스터 모델	2 클러스터 모델	3 클러스터 모델
PCA	55%	50%	44%
SAX	72%	44%	50%

위의 [표 6]을 통해 보면 유사도 기반 방식의 결과들이 모델기반 방식의 결과보다 상당히 낮은 정확도를 보여주고 있다.

V. 결론

시계열 데이터의 모델 기반 클러스터링 알고리즘은 크게 두 단계의 과정으로 이루어진다. 본 연구에서는 첫 번째 과정인 최적의 클러스터 수를 결정하는 과정에 대해서 살펴보았다. 클러스터 수를 추정함과 클러스터링 방법론으로 유사기반과 모델기반의 방법론을 비교하였다.

최적의 클러스터 수를 결정짓는 과정인 베이지안정보기준 측도에 대하여 고찰한 결과 실험결과는 베이지안정보기준 측도가 일반적으로 클러스터 수를 정확하게 추정하는 결과를 보여주고 있으나 데이터객체의 수와 특징의 길이에 영향을 받는 것을 확인하였다. 이점은 베이지안정보기준 측도의 유도가 자료의 개수가 많은 경우에 다변량 가우시안 분포로 근사할 수 있다는 점에서 볼 때 당연한 결과로 예측된 것이다.

모델기반과 유사기반 방법론의 클러스터링 결과를 비교한 결과 모델기반의 방법론이 정확한 결과가 나타남을 확인하였으며 모델기반 클러스터링에서 적용된 k-means 알고리즘에서는 모델에 대한 시퀀스의 우도값 뿐만 아니라 모델과 시퀀스의 상태수를 고려하여 할당되어진 것이 일반적으로 우도값 만을 고려하여 할당된 것보다 정확한 클러스터링 결과를 보여주었다.

향 후 연구해야 할 내용은 추정되어진 클러스터들에 대해 개별 클러스터에 대한 모델을 생성하는 문제이다. 이상의 연구가 이루어진다면 일반적인 용도의 시계열 데이터의 클러스터링과 모델링 방법론을 개발하는 것으로서 이러한 방법론이 복잡하고 동적인 시스템들과 프로세스들을 가진 현상들을 이해하는 것에 도움이 될 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] 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).

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

[5] 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).

[6] E. Keogh., K. Chakrabarti, M. Pazzani, and Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," Journal of Knowledge and Information Systems, pp.263-286, 2000.

[7] J. Lin, E. Keogh, and P. S. Lonardi, "Finding motifs in time series," In the 2^{nd}Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Canada, 2002.

[8] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. of IEEE77, pp.257-286, 1989.

[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. Chickering, "A tutorial on learning with Bayesian Network," Machine Learning, Vol.20, pp.197-243, 1995.

[11] J. Alon, S. Sclaroff, and G. Kollios, "Discovering cluster in motion time-series data," Proceedings of Computer Vision and Pattern Recognition, 2003.

[12] F. Porrkli, "Clustering variable length sequences by eigenvector decomposition using hmm," International workshop on statistical pattern recognition(SPR 2004), 2004.

저 자 소 개

전 진 호(Jin-Ho Jeon) 정회원

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

<관심분야> : 데이터마이닝, 전산금융, 기계학습

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