논문 2010-6-21

정보기준과 다중 중심점을 활용한 클러스터별 예측

Prediction on Clusters by using Information Criterion and Multiple Seeds

조영희*, 이계성*

Young-Hee Cho, Gye-Sung Lee

요 약 본 연구에서는 시계열 자료를 베이지안 정보기준을 통해 클러스터링 한다. 보다 안정적인 클러스터를 생산하기 위해 다중 중심점을 모델링한 후 이를 이용하여 클러스터를 생성시킨다. 대상 시계열 자료에 대해 예측할 경우 클러스터에 속한 시계열 자료 중 가장 유사한 시계열 자료를 선택하여 모델링한다. 모델로부터 마코프 규칙을 유도해 내고 이 규칙을 이용해 예측정확도를 측정한다. 시계열 자료를 단독으로 모델링한 후 예측한 결과보다 클러스터에 속한 유사시계열 모델링을 통한 예측정확도가 좀 더 높았음을 확인하였다.

Abstract Bayesian information criterion is used to do clustering for time series data. To acquire more stable clusters, multiple seeds are chosen first for the algorithm. Once clusters being set up, most similar time series data in the cluster to the one under consideration are to be chosen for prediction test. These chosen time series data are used to extract valid Markov rules by which we test the prediction accuracy. We confirmed that clustering with multiple seeds led to better prediction performance.

Key Words : 시계열(time series), 클러스터링 (clustering), 베이지안 정보(Bayesian Information)

Ⅰ. 서 론

실세계에서 발생하는 시계열 자료에 대한 자료 분석은 매우 어려운 분석 가운데 하나로 알려져 있다. 특히 경제, 사회 지표와 같은 시계열 자료에 대한 분석 또는 모델링, 예측 등은 불확실성과 같은 특정 요소로 인해 문제를 더욱 어렵게 만든다. 주가와 같은 금융, 경제 지표 자료를 살펴보면 그 자료는 단순 수치의 연속의 형태를 갖는 단순형 자료이지만 그 속에는 수많은 변수에 의해 독립적 또는 상호 연관적 영향을 서로 미치면서 변화한다. 따라서 이들을 분석하는 방법은 다양하고 복잡할 수밖에 없다. 이들 변인들을 찾아내고 이들로 모델을 구성하는 모델링 방법에는 한계가 있다. 반면에 자료에 기반을 둔 기술 분석과 같은 접근방법을 살펴보면 자료에 집중하는 접근방법을 취한다. 자료를 분석하여 자료의 형태나 패턴을 추출하여 해석하는 방법이다. 예측에도 자료에 기반을 둔 방법을 활용할 수 있다. 자료의 형태나 패턴이 향후 어떤 추세로 움직일 것인지 파악할 수 있다면 향후의 예측도 가능할 것이다. 이런 향후 추세를 판단하는 것은 예측의 문제로 금융경제 분야에서 중요하게 다루는 연구 분야의 하나이다.

시계열 자료의 분석 및 활용에 있어 자주 사용되는 방법으로 모델링 방법이 있다. 복잡하고 다양한 현상을 자료를 중심으로 모델링할 수 있다면 그 현상을 파악하고 이해하는데 많은 도움을 줄 것이다. 시계열 자료 분석에 있어서도 모델을 토대로 시간적 흐름에 대해 어떻게 변화할 것인지를 파악하려는 시도가 계속되고 있다. 이런 분석 및 예측하는 대부분의 연구에서 이뤄지는 방법은

*정회원, 단국대학교 컴퓨터과학과
접수일자 : 2010.11.22, 수정완료일자 : 2010.12.12
게재확정일자 2010.12.15

통상 하나의 시계열 단위로 모델링하는 방법을 취한다. 본 연구에서는 다수 개의 시계열을 중심으로 모델링하는 방법에 대해 연구한다. 사용되는 기본 방법은 클러스터링을 통한 모델링 방법이고 예측하는데 사용하는 모델은 마코프 체인 모델을 기본 모델로 선정하였다. 모델링을 통해 유용한 패턴을 찾아 그 패턴을 이용하여 새로운 상황에 적용하여 향후 추세를 결정할 수 있는 방법을 제안한다. 2장에서는 관련연구를 기술하고, 3장에서는 클러스터링을 이용한 모델링에 대해 기술하고 클러스터별 모델을 이용하여 예측에 활용하는 방안을 제시한다. 4장에서는 이를 실험하여 산출된 결과에 대해 분석하고 그 방법에 대한 타당성을 검토한다. 마지막으로 5장에서는 결론을 기술한다.

Ⅱ. 관련연구

시계열 자료의 클러스터링에는 세 가지 종류의 방법으로 구분된다. 자료 간 근접성을 이용한 클러스터링 방법, 특징에 기반을 둔 클러스터링 방법, 모델에 기반을 둔 클러스터링 방법으로 나눌 수 있다. 근접성에 기반을 둔 방법은 계열 간 상관계수를 구하거나 동적 시간 변형과 같은 방법을 활용하여 계열 간 유사 거리를 측정하여 클러스터링 하는 것이다[1]. 특징기반 방법은 퓨리에 계수나, 구간별 다항 함수를 사용하는 방법이 동원되기도 한다[2].

세 번째 방법인 모델 기반 방법에는 자기회귀 모델과 이동 평균 모델을 비롯하여 베이지안 모델, 마코프 모델[3], 은닉모코프 모델[1,4,5,7,8], 상호정보[9]를 이용한 방법들이 있다. 자기회귀 모델과 이동평균 모델은 자료의 시점 사이의 관계를 추정하는 방법으로 사용된다. 자기 회귀 분석 모델에서 종속변수는 현재 값이고 독립변수는 자기 회귀의 차수로 불리는 N개의 이전 값으로 설정된다. 이 모델의 기본 개념은 시계열 자료의 한 점에서의 값은 그 시점 이전과 이후의 값과 밀접한 관계를 갖는다는 사실에 근거를 둔다. 이 방법은 자연계나 사회적 현상의 다양한 형태를 모델링하고 예측하는데 자주 사용된다. ARMA (ARIMA) 모델은 시계열 자료에서 자기 상관관계나 지속성(persistence)을 모델링하는 수학적 모델로 자주 사용된다. ARMA 모델은 계절성, 비 정지성(non-stationary) 등 다른 요소에 의해 예측에 한정된 역할을 하게 되는 단점이 있다[10]. 상호정보를 포함한 다른 통계적 방법도 항상 고도의 기술적 제한 조건 및 적용 가능한 환경에 대한 조건을 제시하지 않을 경우 모델링은 제한적일 수밖에 없다[6].

시계열 자료에 대한 클러스터링과 모델링에는 은닉마코프 모델이 자주 활용된다[1,5,7,8]. 은닉 마코프 모델은 실세계에서 관측되는 관측 자료에 대한 상태의 확률 함수로 이뤄진다. 관측되는 자료는 다수 개의 숨겨진 상태에서 생성될 수 있기 때문에 이 모델은 특히 이중 내장 추계학적(stochastic) 프로세스라고 불린다[4]. 은닉마코프 모델로 설명되는 동적시스템은 시스템 내부에서 동작되는 동적, 추계학적 프로세스의 징후를 관측 가능한 형태로 관측될 수 있다. 이들 관측 자료는 시간적 운행과정에서 발생하는 동적과정의 결과로 발생되는 단계들의 나열이라고 볼 수 있다. 숨겨져 있는 상태는 동적 프로세스에서 잠재적으로 내재하는 유효한 상태의 집합을 모델링하는데 적합할 수 있다. 이들 상태는 관측될 수 없지만 획득된 자료를 통해 추정하여 자료 속에 내재되어 있는 변화나 정보를 표출시킬 수 있다. 특히, 은닉마코프 모델링은 모델을 구성하는 파라미터를 유도하는데 있어 확률이론적 배경을 잘 갖춘 바움-웰치 방법과 같은 알고리즘을 활용할 수 있다.

은닉마코프 모델을 사용하는 많은 시계열 예측 연구에서는 시계열 자료를 개별로 분석하여 움직임을 예측하는 방법을 취한다[6,7,8]. 만일 유사한 복수개의 시계열 자료가 있다면 이들을 함께 분석하여 예측한다면 보다 개선된 예측이 될 수 있을 것으로 기대된다. 본 연구에서는 은닉마코프의 모델링 능력을 클러스터링 방법에 활용한다. 기본적으로 클러스터링은 은닉마코프 모델링을 통한 베이지안 클러스터링 방법을 사용한다[1,5,7]. 클러스터링의 기준이 되는 측도로 베이지안 정보기준을 사용하는데 이에 대한 내용은 다음과 같다.

1. 베이지안 정보기준

베이지안 정보기준(Bayesian Information Criterion, BIC)는 베이지안 모델 선택 방법에서 유도되었다. 자료, $X=(x_1,\ldots,x_N)$가 주어질 때 모델 λ가 X로부터 생성된다. 베이즈 이론에 의하면 모델의 사후확률 $P(\lambda|X)$는 $P(\lambda|X)=\frac{P(\lambda)P(X|\lambda)}{P(X)}$로 정의된다. $P(X)$와 $P(\lambda)$는 자료와 모델의 사전확률이다. $P(X|\lambda)$는 자료에 대한 한계우도이다. 모델을 구성하는 파라미터 구성 θ가 있을 때

자료의 한계 우도는 식(1)과 같이 정의된다. 이 한계우도는 시계열 자료와 모델의 적합도 계산에 활용된다[11].

$$P(X|\lambda) = \int_{\theta} P(X|\theta,\lambda)p(\theta|\lambda)d\theta \quad (1)$$

자료의 사전확률은 모델 전체에 대해 변하지 않으므로 모델의 사후확률은 $P(\lambda)P(X|\lambda)$에 의해 결정된다.

은닉마코프 모델을 이용한 클러스터링 모델 학습에서 요구되는 모델은 클러스터별 모델로 나눠진다(λ_k). 총 클러스터의 수를 K개로 가정하고, 자료의 수를 N으로 가정할 때 한계우도는 식(2)와 같이 정의된다[5].

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

여기서 $\hat{\Theta}$는 K 클러스터에 대해 한계우도를 최대화하는 모델의 파라미터 구성을 나타낸다. d는 모델의 차원을 의미하고, A는 음의 헤시안이다. 이 중 자료N에 비례하여 값이 변화하는 항만 취하면 식(3)이 구해진다.

$$\log P(X|\lambda) \approx \log P(X|\hat{\Theta},\lambda) - \frac{d}{2}\log N \quad (3)$$

이 식은 자료를 잘 설명할 수 있는 상세한 모델을 선호하는 경향을 갖는 우도 항(첫 번째 항)과 모델을 구성하는 파리미터의 수를 작게 하여 모델을 일반화시키는 두 번째 항의 조화를 통해 모델이 선택되도록 유도하는 식이다. 두 번째 항을 페널티 항이라 부른다. 모델의 구성이 주어진 자료에 너무 맞춰지는 오버피팅 문제를 야기하지 않게 제어하는 역할을 한다. 모델의 설명력과 모델의 간략성이 조화를 이루는 가운데 가장 이상적인 모델이 생성되는 알고리즘이다. 식(3)을 베이지안 정보기준이라 부른다. 이 베이지안 정보기준은 클러스터링의 유효성을 판단하는 기준으로 사용된다.

2. 마코프 체인 모델

마코프 체인 모델은 예측정확도를 측정하는데 사용된다. 대상 시계열 자료가 준비되면 이의 일부는 훈련 집합으로 예측 모델을 만드는데 사용하고 일부는 테스트용으로 사용한다. 마코프 모델은 시계열 자료를 상태 공간으로 구분하여 상태 간 이동을 확률 모델로 만드는 과정에 의해 구성된다[7]. 마코프 모델은 자료를 모델링하는 은닉 마코프 모델보다 모델링하는 데에는 한계를 갖고 있으나 특정 상황이나 패턴이 출현한 이후를 예측하는 모델로 유용하게 사용될 수 있다.

여기서 상태는 자료로부터 정의된다. 시계열 자료의 경우에는 이전 상태에서 다음 상태로 전이하는데 필요한 값의 변화를 측정하면 된다. 값의 변화는 상승하거나 하락하는 두 가지의 상태로 나눌 수 있고 보합을 중간에 포함시키면 세 가지 상태를 갖게 된다. 본 연구에서는 세 개의 상태를 갖도록 자료를 구분하였다.

마코프 체인 모델은 일련의 과거 상태가 현재 상태를 결정하는 모델이므로 과거의 상태 개수를 정하여 특정 패턴으로 정의할 수 있다. 일정 수준의 확률 값 이상을 가질 때 패턴으로 정의된다. 현재 상태를 결정하는데 필요한 과거의 상태 개수에 따라 마코프 모델의 차수가 결정된다. 과거의 상태개수가 많게 되면 전이 확률 값이 기하급수적으로 늘어나 확률 값이 전체적으로 낮아져 유효한 패턴을 찾기가 어렵다. 본 연구에서는 유효패턴을 찾기 위해 현재의 상태와 과거의 2개 상태를 검사한다.

Ⅲ. 본론

시계열 자료를 클러스터링 하기 전 자료에 대한 전처리 과정이 있다. 대상이 되는 주가 자료의 경우 업종별, 종목별 그 크기의 수준이 서로 다르기 때문에 이들을 직접 비교할 수는 없다. 클러스터링은 시계열 간 추세적인 변화가 유사한 것들을 묶는 문제이기 때문에 정규화 된 값으로 전환하여 정규화 시키는 것이 필요하다. 정규화는 각 시계열의 평균값 μ_i와 표준편차 σ_i에 의해서 결정된다. 관측값 o_{ij}에 대한 정규화 값은 $(o_{ij}-\mu_i)/\sigma_i$로 구해진다.

시계열 자료에 대한 모델링은 단일 시계열을 대상으로 생성된 모델이 아니라 대상 시계열과 유사하게 움직이는 다수 개의 시계열 자료를 대상으로 모델링한다. 먼저 하나의 시계열 자료를 선정한 후 이를 모델링한 후 이 모델과 가장 유사한 시계열을 중심점으로 잡는다. 이때 사용하는 유사도 측도는 식(1)과 같은 우도를 사용한다. 복수 개의 중심점 시계열 자료를 복합 모델로 구성한다. 이 모델을 통해 베이지안 클러스터링을 수행한다. 베이지안 클러스터링 알고리즘은 그림 1과 같다. 주어진 입력 시계열 집합, $X=x_1,..x_i.,x_N$에 대해 출력 모델 M과 클러스터 C를 구하는 알고리즘이다. 여기서 $LL()$은 식(1)과 같이 우도를 계산하는 식이다. 클러스터링이 결정된 후에는 마코프 체인 모델을 이용하여 패턴을 구분하고

각 패턴이 출현할 가능성을 전이 확률밀도함수로 찾는다. 그 후 각 패턴의 발생가능성이 가장 큰 패턴을 찾아 그 패턴 이후의 값을 가지고 향후 변화를 예측하는 방법을 취한다.

예측정확도는 테스트용 자료에 대해 예측값과 실제값을 비교하여 비율로 측정한다. 예측정확도를 측정하는데 사용되는 방식은 훈련 자료 구간의 종류에 따라 세 가지로 나눠진다. 훈련 구간을 과거의 특정 기간에 한정하고 테스트용 자료를 사용해 예측정확도를 구하는 방식을 MC 방식이라 부른다. 반면 동일한 기간이지만 테스트가 진행되면서 훈련기간도 하나씩 이동하는 방식은 SD 방식이라 한다. 테스트에 사용되었던 기간을 누적시켜 모델링하는 방식을 CM 방식으로 정한다.

입력 : $X = x_1, \ldots x_i \ldots, x_N$
출력 : $M = \lambda_1, \ldots \lambda_j \ldots, \lambda_K,\ C = C_1, .., C_k, \ldots C_K$
$previousBIC = -\infty,\ K = 1$
do
  for k, $\lambda_k \leftarrow HMM(Center_k)$ 여기서 $Center_k$는 중심점
  $previousLL = -\infty$
  while true do
    for i, j, k, $C_k \leftarrow \max_{y_i} LL(\lambda_k, x_i)$
    for k, $\lambda_k \leftarrow HMM(C_k)$ $currentLL = \sum_{i \in C_k} LL(\lambda_k, x_i)$
    if $currentLL < previousLL$
      break;
    else
      $previousLL = currentLL$
    end
  end while
    $currentBIC = \sum_k BIC(C_k)$
    $K = K + 1$
while $(currentBIC > previousBIC)$

그림 1. 클러스터링 알고리즘
Fig. 1. Clustering Algorithm

예측정확도를 측정하기 위한 규칙 유도는 마코프 모델의 전이 확률값을 기준으로 결정한다. 이전 2개의 상태를 정하고 다음 상태로 이전하는 확률값이 가장 큰 것을 선택하여 전이 규칙을 정하는데 하나가 절대적으로 크다면 규칙으로 선정되는 데에 문제가 없다. 두 개 이상의 전이확률이 큰 편차가 없을 경우에는 유효규칙이 생성되지 않는다. 상위 두 개의 전이확률 편차가 10%이상일 경우에 한해 유효규칙으로 유도된다.

Ⅳ. 실험 및 결과

1. 실험자료

위에서 제시한 클러스터링 알고리즘과 예측 알고리즘을 이용하여 예측정확도를 산출하기 위해 2006년도 코스피 지수 중 업종별 자료를 선정한다. 업종별 코스피 지수에는 총 22가지 업종이 있다. 이들에 부여된 일련 번호를 사용하기로 한다(표 1). 업종별 지수 22개의 자료는 클러스터링을 위해 정규화 과정을 거친다. 자료의 세부적인 값보다는 변화의 방향성에 중심을 두기 위해 20일 이동평균선을 구해 얻어진 자료에 정규화 과정을 적용시킨다.

표 1. 업종별 항목
Table 1. Items by Industrial Category

1.음식료	2.섬유의복	3.종이목재	4.화학
5.의약품	6.비금속광물	7.철강및금속	8.기계
9.전기전자	10.의료정밀	11.운수장비	12.유통업
13.전기가스	14.건설	15.운수창고	16.통신
17.금융	18.은행	19.증권	20.보험
21.서비스	22.제조업		

예측정확도를 테스트하기 위해 원 자료를 로그비로 변환한 후 상태로 구분해야 한다. 로그비로 변환된 자료를 분석해보면 대부분의 값이 0을 중심으로 몰려 있다. 일반적으로 지수의 변동성이 급변하기보다는 대부분 서서히 변화하는 모양을 갖기 때문이다. 더욱이 개별 종목에서 급변하는 날이 발생하더라도 업종에 속한 다른 종목들이 함께 움직이지 않는다면 업종별 지수는 평탄화될 것이다. 상승과 하락으로 구분하여 분석해 보면 상승과 하락 일수의 비율이 비슷한 것으로 조사되었다. 상승, 하락과 보합의 구분을 위한 임계치를 적용할 때 상승, 하락을 약 40%, 보합을 20%의 구성을 갖게 임계치를 조정한다. 이런 비율로 무작위 예측할 때 예측정확도에 대한 이론적인 값은 36%이다.

2. 실험결과

개별 시계열을 이용한 베이지안 클러스터링 방식을 BIC로 부르기로 하고 다중 중심점 계열을 선정한 후 클러스터링 한 방식을 BIC2라 부르기로 정한다. 이 두 가지 방식으로 22개 계열 자료를 클러스터링한 결과가 표 2에 표시되었다. 둘 다 3개의 클러스터를 산출하였다. 그러나

내부 구성요소를 보면 약간의 차이가 있음을 확인할 수 있다.

BIC[7] 방식의 C3에 속한 3개의 계열은 BIC2의 방식에 의해 산출된 C2에 그대로 나타난 것을 볼 수 있다. BIC의 C2에 있던 계열 8 16 20이 BIC2의 C2에 남아 있다. 계열 2, 6, 13은 BIC2의 C1로 이동해 클러스터를 구성하였다. 이 클러스터링 결과에서 유의할만한 점은 클러스터의 분포이다. BIC의 경우 그 분포가 한쪽으로 몰려서 구성됨을 볼 수 있다. 반면 BIC2의 경우 대략 균등하게 분포됨이 비교된다. 아무래도 BIC의 C3과 같이 소규모 집단의 클러스터는 매우 결집도가 높아 다른 클러스터에 비해 서로간의 유사도가 높게 나타난다. 이들 간의 예측정확도를 비교해 보자(BIC의 결과참조[7]).

표 2. 클러스터링 결과
Table 2. Result of Clustering

cluster	BIC	BIC2
C1	1 3 4 5 9 11 12 15 17 18 19 21 22	1 2 5 6 11 13
C2	2 6 8 13 16 20	7 8 10 14 16 17 18 20
C3	7 10 14	3 4 9 12 15 19 22

실험은 기간별 예측의 평균값을 구해 그 평균값의 개선 정도를 분석할 것이다. 예측 기간에 따라 장기, 중기, 단기로 구분하여 예측정확도를 측정한다. 테스트용 자료가 각각 40일, 26일, 16일에 해당한다. 세 가지 방법에 대해서도 구분하여 실험한다(표 3. 결과참조[10]).

표 3. 전체 개별 예측정확도 평균(%)
Table 3. Prediction Accuracy Average(%)

방법	단기	중기	장기	평균
MC	49.99	43.18	41.54	44.90
SD	48.40	43.01	43.59	43.79
CM	47.79	41.30	42.29	45.02

이 표를 보면 45%선의 예측 정확도 평균을 갖는다. SD 방식의 경우는 약간 낮은 정확도를 갖는다. 전체 평균을 구한다면 44.6%가 된다. 개별 시계열에 대해 최대 4개까지 유사 시계열을 구해 이들로 모델을 유도한 후 이를 대상으로 기간별 전체에 대한 예측정확도의 평균을 구한 결과가 표 4에 나타나 있다. 유사 계열 수가 1인 경우는 그 자체로 예측정확도를 구한 것이다. 2 이상에서는 대상 계열과 유사계열이 하나씩 증가하면서 모델이 생성된다. 이 표를 보면 MC의 경우는 유사계열을 추가하여 모델링을 하더라도 정확도에서의 개선을 찾아 볼 수 없으나 SD와 CM의 경우는 각 유사계열 수가 3과 4일 경우에 그 개선 폭은 그리 크지 않을지라도 예측 정확도를 개선시킨 것을 볼 수 있다.

BIC와 BIC2를 이용하여 각 클러스터별 예측정확도 평균을 구했다. 각 클러스터별 예측정확도 평균이 표 5에 표기되었다. BIC 방법의 클러스터 C1의 경우 단독 계열을 통해 예측하는 것이 가장 효과적인 것으로 나온다. 즉, 유사계열을 합해서 모델을 만들 경우 큰 효과를 보지 못하는 것으로 나왔다. C2의 경우, 각 방법에 대해 다중 유사 시계열을 활용하면 개선된 정확도를 얻을 수 있으나 그 값은 전체에 대한 평균값보다 크지 않아 개선된 결과라고 할 수 없다. 클러스터 C3에는 3개의 계열 자료가 있다. 유사한 계열로 이뤄졌다고 볼 수 있는 이 클러스터의 예측정확도는 평균을 훨씬 웃도는 결과를 산출하였다. MC와 CM의 경우에는 50%대까지 개선된 것을 볼 수 있다.

표 4. 방법별 예측정확도
Table 4. Prediction Accuracy by Method

유사계열	MC	SD	CM
1	44.91	43.79	45.02
2	40.93	41.16	40.97
3	41.88	45.52	42.97
4	42.94	43.48	46.63
5	37.97	44.41	40.13

표 5. 예측정확도 비교
Table 5. Prediction Accuracy Comparison

		BIC			BIC2		
		MC	SD	CM	MC	SD	CM
C1	1	47.5	46.3	47.0	40.2	37.0	39.9
	2	41.1	39.4	38.0	40.5	38.2	40.8
	3	40.4	42.8	35.8	41.2	39.5	41.3
	4	43.7	45.0	39.3	42.6	40.7	43.8
	5	45.0	46.9	41.3	34.9	34.0	33.4
C2	1	39.8	38.8	40.0	45.2	45.6	47.2
	2	41.8	44.2	43.2	44.0	44.1	45.0
	3	36.9	38.2	38.3	47.7	49.7	49.4
	4	42.3	35.7	36.3	51.0	52.8	53.2
	5	44.6	41.3	47.4	52.6	50.2	51.1
C3	1	43.8	43.0	46.4	48.2	47.1	46.7
	2	46.4	44.8	48.3	38.9	40.3	38.5
	3	52.2	47.9	51.7	39.6	40.3	39.2
	4				43.5	49.1	47.1
	5				47.6	51.6	44.9

BIC2 방법에서 단독 계열을 이용한 예측정확도는 대부분 평균이하에 해당된다. 클러스터 C1의 경우에는 유사계열 수가 4일 때 최대가 되나 그 값은 평균에 미치지 못한 결과를 나왔다. C2나 C3에서는 단독 시계열을 이용한 예측이 모두 평균을 조금 상회하는 정확도를 나타낸다. 그러나 다중 유사 시계열을 활용할 때는 평균을 훨씬 상회하는 정확도가 구해졌다. C2의 경우 유사시계열 수가 4 또는 5일 때 모두 53%에 해당하는 정확도를 산출하였다. C3에서는 세 가지 방법에 대해 각각 47.6, 51.6, 47.1%의 정확도가 나와 평균을 훨씬 웃도는 결과로 나왔다.

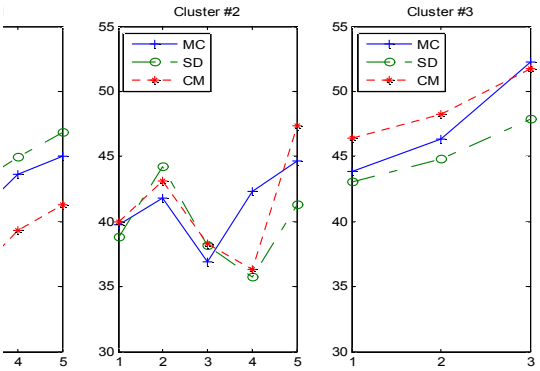

그림 2. BIC의 예측정확도
Fig. 2 Prediction Accuracy for BIC

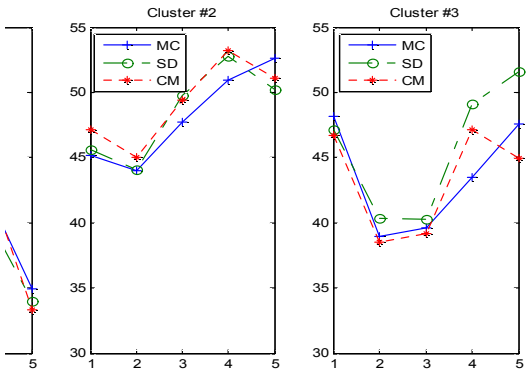

그림 3. BIC2의 예측정확도
Fig. 3. Prediction Accuracy for BIC2

그림 2와 그림 3은 두 방법으로 산출된 정확도 평균을 보여주고 있다. BIC2의 경우 유사 시계열의 수가 증가하면서 값의 변화가 증가하다가 특정 위치에서 정상을 이룬 후 하락하는 모습을 볼 수 있다. 이는 클러스터링으로 구분한 후 클러스터에 속한 유사 시계열을 모아 모델링할 때 더 좋은 정확도를 산출한 것으로 분석된다. 반면 BIC 방법에서는 그런 돌출형 정상을 보이기보다는 아래쪽으로 돌출된 형태를 취하고 있어 유사 시계열의 효과를 크게 볼 수 없다고 할 수 있다. 그라프에서 BIC의 클러스터 C3의 경우를 보면 3개의 시계열 자료를 종합하였을 때 가장 높은 정확도가 산출된 것을 보여준다. 이는 클러스터의 밀집도가 클 경우, 즉, 시계열 자료간 유사도가 매우 높을 경우 다수 개의 유사시계열을 통합한 모델의 예측정확도가 더욱 높아짐을 알 수 있다.

BIC2로 산출된 클러스터 C2에 기간별로 훈련자료 세트의 크기를 변화시키면서 유사계열의 수에 따른 변화를 측정해 보았다. 그 결과가 그림 4에 표시되었다. 좌측 하단 축은 유사시계열 수를 나타내고 우측 하단 축은 테스트 자료의 크기를 표시하였다. 이 그림을 보면 테스트자료의 크기가 작아질수록, 즉, 단기 예측일수록 예측 정확도가 증가한다. 클러스터에 속한 유사계열수가 4일 때 가장 높은 예측정확도를 나타내고 있음을 확인할 수 있다.

각 방법별 예측정확도 평균값을 종합한 결과가 표 6에 표시되었다. Single은 개별 시계열에 대한 결과이고, Sim4는 개별 시계열에 대한 유사시계열을 최대 4개까지 포함시킨 결과이다. BIC나 BIC2의 경우 4개의 유사 시계열을 종합하였을 경우에 가장 큰 예측값을 산출할 수 있었다. 표에서 평균값을 보면 개별로 예측할 때 보다 유사 시계열을 찾아서 예측하는 것이 좀 더 효과적이며, 클러스터로 나눌 경우(BIC, BIC2)에는 예측정확도를 더 개선시킬 수 있었다.

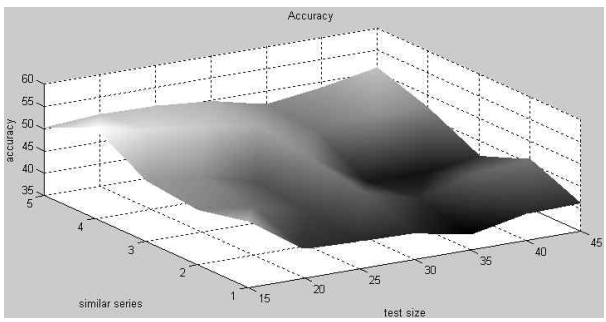

그림 4. 예측정확도 변화
Fig. 4. Change of Prediction Accuracy

표 6. 예측정확도 평균
Table 6. Prediction Accuracy Average

	MC	SC	CM	평균
Single	44.90	43.79	45.02	44.57
Sim4	42.94	45.52	46.63	45.03
BIC	47.31	46.31	46.79	46.81
BIC2	47.60	48.38	48.06	48.01

V. 결 론

본 연구에서는 클러스터링을 통해 유사시계열을 확보한다. 다수 개의 유사 시계열을 확보한 후 단일 시계열을 활용한 모델링에 비해 어느 정도 예측정확도에 개선을 가져왔는지 여부를 조사하였다. 개별 시계열을 통한 예측정확도 평균이 44.6%이다. 개별 시계열에 유사한 시계열을 종합하여 모델링한 후 예측하였을 때 약간의 개선이 있었다. BIC와 BIC2는 그보다 더 큰 개선 효과를 가져왔다. 각각 개별 시계열 모델링보다 예측정확도 평균에 있어 2.2%와 3.4%의 개선을 얻을 수 있었다.

개별 시계열자료로부터 유도해 낼 수 있는 정보보다 다수 개의 시계열을 종합하여 모델링 했을 경우 좀 더 정확한 정보가 추출되었다고 말할 수 있다. 이런 개선에 클러스터가 모두 동등하게 기여하지는 않았다. 실험 결과에서 보았듯이 세 개의 클러스터 중 한 개의 클러스터는 개별 시계열의 예측정확도보다 못한 결과가 나왔다. 그러나 나머지 두 개의 클러스터는 개별 시계열 예측보다 훨씬 좋은 결과를 산출해 내었다. 향후 이런 불균형적인 요소를 해소하기 위해 안정적인 클러스터링 알고리즘에 대해 연구와 예측 방법의 개선에 대해 연구할 예정이다.

참 고 문 헌

[1] 전진호, 이계성, “시계열 데이터의 모델기반 클러스터 결정에 관한 연구”, 한국콘텐츠학회 논문지 제 7권 6호, 22-30쪽, 2007년 6월.

[2] E.S. Ristad & P.N. Yianilog, "Learning strig edit distance," Proc of the 4th Int. Conf. on Machine Learning, pp. 773~779, 1997

[3] Papageorgiou, C. P., "High frequency time series analysis and prediction using Markov models," in Proc. of the conf. on Comp. Intelligence for Finance, pp.182-185, Mar. 1997.

[4] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of the IEEE, vol.77, no.2, pp.557-286, 1989

[5] C. Li, and G. Biswas, "Building models of ecological dynamics using HMM based temporal data clustering," IDA 2001, pp. 53-62. 2001

[6] Duan, J. et al., "A prediction algorithm for time series based on adaptive model selection," Expert Systems with Applications 36, pp. 1308-1314, 2009.

[7] 조영희, 이계성, “마코프 모델에 기반한 시계열 자료의 모델링 및 예측” 한국정보컴퓨터 학회논문지 게재 예정

[8] M.R. Hassan, B. Nath, M. Kirley, "A fusion model of HMM, ANN, and GA for stock market forecasting," Expert Systems with Applications 33, pp. 171-180, 2007.

[9] A. Sorjamaa, et al., “Methodology for long-term prediction of time series," Neurocomputing, pp. 178-186. Elsevier, 2007.

[10] 조영희, 이계성, “다중 유사 시계열 모델링 방법을 통한 예측 정확도 개선에 관한 연구,” 한국인터넷 방송통신학회 논문지, 제10권 6호, 2010.

[11] J.I. Myung, D.J. Navarro, M.A. Pitt, "Model selection by normalized maximum likelihood,", Journal of Mathmatical Psychology, pp.167-179, 2006.

※ 본 연구는 2009학년도 단국대학교 대학연구비 지원으로 연구되었음.

저자 소개

조 영 희(준회원)

- 2000: 단국대학교 이학석사.
- 2008: 단국대학교 이학박사.
- 2008- 현재: 단국대학교 컴퓨터과학과 강사

<주관심분야 : 데이터마이닝, 지능형시스템, 시계열 자료분석>

이 계 성(정회원)

- 1982: 한국과학기술원 이학석사.
- 1994: Vanderbilt 대학 공학박사.
- 1994~: 현재: 단국대학교 컴퓨터과학과 교수

<주관심분야 : 데이터마이닝, 지능형시스템, 시계열 자료분석, 바이오 인포마틱스>