韓國컴퓨터情報學會 論文誌
第14卷 第4號, 2009. 4.

2009-14-4-5-6

휴리스틱 탐색을 통한 동적시스템 분석을 위한 모델링 방법과 CRM 위한 인터페이스 설계

전 진 호*, 이 계 성**

A Modeling Methodology for Analysis of Dynamic Systems Using Heuristic Search and Design of Interface for CRM

Jin Ho Jeon*, Gye Sung Lee**

요 약

실세계의 많은 시스템들은 동적이며 복잡한 현상으로 이뤄져 있다. 이러한 특징의 시스템을 이해하는 방법의 하나로 시스템에 대한 모델을 세우고 분석하는 방법이 있다. 본 연구에서는 동적시스템에서 발생되는 시계열데이터들의 분석을 위한 방법론을 제시한다. 시스템 모델링을 통해 사용자들에게 1:1의 맞춤정보를 제공하기 위한 CRM(고객관계관리) 인터페이스를 제안한다. 실험에서 실제의 시계열데이터를 통하여 군집화 하는 과정에서는 유사기반의 방식보다 모델기반 방식이 더 나은 군집화 결과를 산출하였고 각 군집의 모델을 생성한 후 모델을 통하여 일정기간 시계열 데이터를 생산하여 이를 실제 곡선의 운동양태와 비교 분석하였다. 주가와 같은 실제 시계열데이터에 제안된 방법을 적용하여 모델로 생산된 데이터가 실제의 데이터와 비교하여 얼마나 근사한지를 확인하여 제안된 방법의 유효성을 검증하였다.

Abstract

Most real world systems 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 them. A two-step methodology comprised of clustering and then model creation is proposed for the analysis on time series data. An interface is designed for CRM(Customer Relationship Management) that provides user with 1:1 customized information using system modeling. It was confirmed from experiments that better clustering would be derived from model based approach than similarity based one. Clustering is followed by model creation over the clustered groups, by which future direction of time series data

• 제1저자 : 전진호
• 투고일 : 2009. 02. 05, 심사일 : 2009. 02. 11, 게재확정일 : 2009. 04. 27.
* 단국대학교 일반대학원 컴퓨터과학 전공 ** 단국대학교 컴퓨터과학 전공 교수
※ 본 연구는 2008학년도 단국대학교 대학연구비의 지원으로 연구되었습니다.

movement could be predicted. The effectiveness of the method was validated by checking how similarly predicted values from the models move together with real data such as stock prices.

▸ Keyword : 시계열데이터(Time series data), 클러스터링(Clustering), 모델기반(Model-based), 고객관계관리(CRM)

1. 서 론

실세계의 대부분 시스템들, 예를 들면, 경제, 의료, 생산설비, 그리고 공학시스템 등은 시간의 흐름에 따라 순차적으로 생성되는 연속적인 모임인 시계열데이터의 형태로 표현되어 동적특징을 나타난다. 이러한 동적시스템으로부터 실시간으로 발생되는 대용량 시계열데이터들의 과학적인 연구와 분석의 과정은 모델화하는 것에 초점이 맞추어진다. 이러한 모델링 문제에 대한 방법론의 제시는 다양한 지식영역의 예외적인 상황에 대한 대처와 의사결정과 같은 미래에 대한 예측 등의 작업등에 적용이 가능하다. 또한 시스템의 모델링을 통한 사용자들에게 실시간으로 1:1의 적합한 맞춤정보를 제공함으로서 사용자의 충성도를 향상시킬 수 있을 것이다.

본 연구에서는 동적특징을 갖는 시스템에서 발생되는 시계열데이터를 분석, 이해하기 위한 모델링 방법론과 1:1의 적합한 정보제공을 위한 e-CRM Interface를 제시한다. 모델링 방법론은 두 과정으로 이루어진다. 첫 번째 과정은 군집화 과정이다. 시스템에서 실시간으로 발생되는 대용량 시계열데이터들에 대하여 각각의 데이터로 요약하는 것은 과잉적합(Overfitting)문제가 발생한다. 그러므로 전체를 유사한 군집으로 구분하여 분석함으로서 유연성 있는 모델을 생성할 수 있다. 두 번째 과정은 분할된 각 군집을 잘 설명할 수 있는 모델 결정이다. 결정된 모델을 통해 예측 등의 다양한 의사결정에 적용이 가능하다.

본 연구에서는 과거 KOSPI 시장에서 발생된 실제 주가데이터를 제시되는 모델링 방법론에 적용하여 유효성 검증을 한다.

2. 모델링 방법론 제안과 관련연구

동적특징을 갖는 시스템에서 발생되는 시계열데이터를 분석, 이해하기 위하여 제시하는 모델링 방법론은 두 과정으로 이루어진다.

첫 번째 과정은 군집화 과정이다. 과거에 연구된 시계열데이터의 군집화[1][2] 방법론은 두 가지 범주로 구분할 수 있다. 스트링편집거리[3], 해밍거리[4], 그리고 데이터의 차원감소를 통한 특징 표현의 랜드마크기법, 구간상수근사[5] 등의 유사기반 방식과 회귀모형, 신경망기법, 마코프체인기법, 그리고 은닉마코프모델[6] 등의 모델기반 방식이 있다.

표 1. 길이가 동일한 데이터의 유사도 계산
Table 1. Similarity computation with data of same length

시계열데이터의 길이가 동일한 경우	
Sequence 1	1111111111
Sequence 2	1111111222
Sequence 3	1111112112
해밍거리	**마코프체인**
거리(S_1,S_2)=3 거리(S_1,S_3)=2	우도(S_1 λ_2)=0.26 우도(S_1 λ_3)=0.075

표 1을 보면 길이가 동일한 세 개의 시계열데이터에 대하여 유사기반의 해밍거리 방식을 적용한 결과, 첫 번째 시계열데이터(S_1)와 세 번째 시계열데이터(S_3)가 유사한 것으로 나타난다. 하지만 모델기반의 마코프체인 모델을 적용한 결과, 첫 번째(S_1)와 두 번째 시계열데이터(S_2)가 유사한 것으로 나타난다. 여기에서 λ_2는 sequence2의 모델을, λ_3은 sequence3의 모델을 나타낸다. 표 2를 보면 길이가 다른 세 개의 시계열데이터에 대하여, 유사기반의 스트링편집거리를 적용한 경우에는 첫 번째 시계열데이터(S_1)와 세 번째 시계열데이터(S_3)가 유사한 것으로 나타난다. 하지만 마코프체인 방식을 적용한 결과는 첫 번째(S_1)와 두 번째 시계열데이터(S_2)가 유사한 것으로 나타난다.

위의 결과를 통하여, 유사기반 방식은 여러 시계열데이터들에서 공통부분을 찾거나, 거의 정확하게 일치하는 부분을 찾는 경우에 유용하다. 그러므로 시계열데이터들 내에서 부분시퀀스들의 길이가 중요하다. 이러한 특징들로 인하여 유사기반에서 나타날 수 있는 문제점은 두 시계열데이터들 사이에 유사도가 어떤 측정방법에서는 높게 나올 수 있지만, 다른 유사도 측정방법을 이용하면 낮게 나올 수 있다는 점과 시계열데이터에 내재되어 있는 동적특징을 거리함수에 적용하여 얻는 것이 어렵다는 점이다.

표 2. 길이가 다른 데이터의 유사도 계산
Table 2. Similarity computation with data of different length

시계열데이터의 길이가 다른 경우	
Sequence 1	1111111111
Sequence 2	111111
Sequence 3	1111112112
스트링편집거리	**마코프체인**
거리(S_1,S_2)=4 거리(S_1,S_3)=2	우도(S_1 λ_2)=1.0 우도(S_1 λ_3)=0.075

모델기반 방식은 여러 시계열데이터들의 비교에서 데이터의 길이는 중요하지 않으며 시계열데이터들이 내포하는 의미를 찾는 동적패턴을 확인하는 것이 목적이다. 그러므로 이러한 특징에 의해 시계열데이터의 군집화 과정에는 마코프체인 모델과 은닉마코프모델 등의 방식이 적합하다. [7]에서 기술된 것처럼 본 연구에서는 은닉마코프모델 방법론을 적용한다.

두 번째 과정은 모델결정 과정이다. 각 군집에 대한 상태 수를 결정하는 것과 그에 따른 최대 우도값을 주는 매개변수를 추정하는 과정이다.

제안된 두 과정에서 군집 수 결정과 각 군집을 설명하는 모델의 구조, 즉 모델 상태 수 결정에 있어 기존의 연구 [8][9][10]에서는 제약점이 있다. 즉, 미리 정의된 군집 수와 상태 수를 통해 생성된 군집과 모델은 다른 지식영역들의 작업에서는 적용키가 힘들다는 점이다. 그러므로 본 연구에서 제시되는 방법론은 시계열데이터의 특징을 잘 설명할 수 있는 유연한 모델링을 위하여 군집 수와 모델 상태 수 결정에 있어 휴리스틱 탐색 특징을 갖는 베이지안정보기준(BIC)[11]을 적용한다.

2.1. 군집 수 결정과 군집화 과정

최소비용으로 최적군집 수를 결정하기 위하여 휴리스틱 탐색 특성의 베이지안정보기준 함수를 적용한다. 주된 개념은 군집 수를 하나로부터 시작하여 하나씩 증가를 반복 수행하다가 가장 높은 기준함수 값을 갖는 군집 수가 최적군집 수가 되는 것이다. 최적 군집을 이루는 혼합모델을 선택하기 위하여 한계우도의 계산에 베이지안정보기준을 적용한다. K 군집을 갖는 분할에 대하여, 식(1)처럼 분할사후확률이 정의되고,

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

$\hat{\Theta}$: 군집 K의 한계우도 모델 파라미터 구성

P_k: 군집 k의 사전확률

$f(x_i|x_i \in \lambda_k, \hat{\Theta}, \lambda_k)$: 군집 k에 대한 모델이 주어졌을 때 데이터 x_i의 확률

식(1)에 베이지안정보기준의 적용은 식(2)와 같다.

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

$\log P(X|\hat{\Theta}M)$: 데이터에 대한 모델의 우도값

$\frac{d}{2}\log N$: 모델복잡도에 따른 패널티항이다.

최적군집을 이루는 혼합모델은 전체 군집분할의 복잡도와 전체 데이터의 우도의 조화를 이루는 것이다. 군집 수가 결정된 후 군집들에 대한 데이터의 할당은 군집모델에 대한 데이터의 우도뿐만 아니라 군집들에 대한 상태 수와 데이터의 상태 수를 고려하여 할당한다.

2.2. 군집 상태 수 결정과 모델 생성

모델결정 과정은 최적상태 수와 파라미터의 추정을 통하여 군집을 가장 잘 설명할 수 있는 모델을 선택하는 것이다. 베이즈 이론으로부터 식(3)을 알 수 있다.

$$P(M|X) \propto P(X|M) \quad (3)$$

식(3)에서 모델의 파라미터 구성 θ가 주어지면, 데이터의 한계우도는 식(4)와 같다.

$$P(X|M) = \int_{\Theta} P(X|\Theta, M)P(\Theta|M)d\Theta \quad (4)$$

파라미터들이 연속적인 값들을 가질 때 적분계산은 폐형해(closed form solution)를 획득하는 것이 어렵다. 한계우도를 구하기 위한 근사기법[1]으로 베이지안정보기준을 적용한다.

1) 몬테-카를로와 라플라스 근사 등이 있는데 이들은 매우 정확하기는 하나 계산적으로 비용이 많이 든다.

식(4)의 내부의 항에 로그를 취한 것을 $g(\theta)$로 정의하고 $g(\theta)$를 최대화 시키는 파라미터 구성을 $\hat{\theta}$라 할 때 이는 사후확률을 최대화하게 된다. 여기에 2차 테일러 다항 근사법을 적용한 후 e의 지수를 취하고 다시 원식에 대입하여 다음 식을 산출한다.

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

d : 모델에서 파라미터의 수

A : $\hat{\Theta}$에서 계산되는 $g(\Theta)$ 음의 Hessian

식(5)는 다시 데이터의 수인 N에 비례하는 항만 남기고 나머지를 제거함으로 더욱 근사 시켜 식(6)을 유도된다.

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

$\log P(X|M, \hat{\Theta})$: 데이터를 가장 잘 설명할 수 있는 상세한 데이터의 모델을 찾도록 유도하는 성분

$-\frac{d}{2}logN$: 모델 내의 파라미터 개수에 대한 패널티 항

베이지안정보기준은 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 최적의 모델 상태 수가 결정된다.

모델 상태수가 결정되어 모델의 구조가 주어지면, 은닉마코프모델의 매개변수의 추정은 모델 매개변수들 $\overline{\pi}$, $A(a_{ij})$ 그리고 $B(\overline{\mu}, \Sigma)$의 최적화를 유도한다. 매개변수의 추정방법으로서 E-M 알고리즘의 한 형태인 관측열에 대하여 최대확률을 주는 최대우도 기법인 바움-웰치(Baum-Welch) 매개변수 추정기법[6]을 사용한다.

3. 실험

3.1 군집화 과정(유사 vs 모델기반)

실험에서는 과거 KOSPI 시장의 실제 주가데이터를 이용하였다. 군집화 과정의 실험을 위하여 그림 1처럼 같은 기간의 업종별 주가지수 중 유사한 시계열패턴을 갖는 전기전자, 유통, 그리고 건설업종의 데이터를 선택하였다. 세 업종이 갖는 주가지수의 차이를 정규화를 시켜 세 업종이 유사한 시계열패턴임을 보여주고 있다.

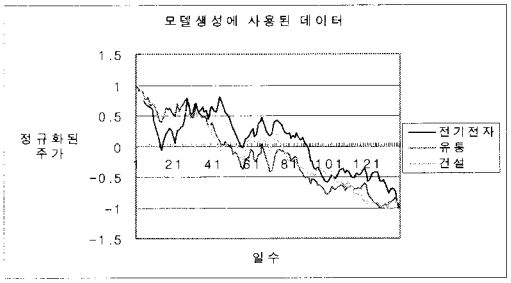


그림 1. 모델생성에 사용된 업종별 데이터
Figure 1. Categorical data used for model creation

모델생성에 사용된 데이터 기간은 2005년 5월 3일부터 2005년 11월 16일이다. 생성된 전기전자, 유통, 건설업종의 모델로부터 각각 6개씩 총 18개의 시퀀스를 랜덤하게 생성하였다.

시계열데이터에 대한 최적군집 수 결정을 위한 베이지안정보기준의 유효성 검증결과는 [7]에 제시되었다. 베이지안정보기준에 의해 추정된 군집 수를 통해 모델기반과 유사기반을 비교한다.

모델기반 군집화에서는 전기전자(시퀀스1-시퀀스6), 유통(시퀀스7-시퀀스12) 두 업종에서 생성된 12개의 시퀀스들에 대하여 우도에 따른 k-means 알고리즘을 적용하였다.

표 3. 전기전자와 유통업종의 우도에 따른 군집 결과
Table 3. Clustering by likelihood over Electrical and Electronic group and distribution industry

시퀀스 No.	전기전자의 우도값	시퀀스 No.	유통업종의 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 3은 각 군집별로 할당된 시퀀스들의 우도값을 보여준다. 이를 통해 시퀀스들이 각 군집에 정확히 군집화된 것을 확인할 수 있다. 그러나 군집들이 다수일 경우, 데이터들이 복수의 군집들에서 유사한 우도값을 갖는 경우가 있다.

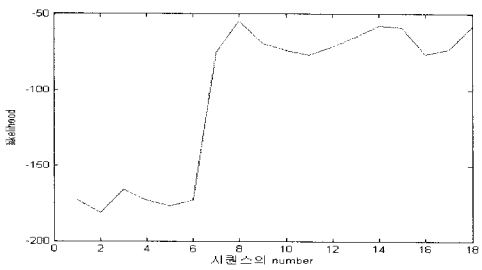


그림 2. 세 군집의 상태 고려 전 우도값
Figure 2. Likelihood before 3 clusters being considered

위의 그림 2를 보면 X축은 시퀀스데이터의 번호를 보여준다. 7번-12번이 시퀀스는 유통업종으로 상태 수가 4개, 13번-18번 시퀀스는 건설업종으로서 상태 수가 3개로서 즉, 서로 다른 상태수를 갖는 모델에서 나온 시퀀스데이터이지만 유사한 우도값을 보여주는 경우도 있다. 이러한 경우, 군집의 모델 상태수와 데이터의 상태수를 비교하여 같은 상태수를 갖는 군집모델에 할당할 경우 더욱 정확한 군집화 결과를 얻을 수 있다. 알고리즘은 표 4와 같다. 표 4의 알고리즘을 세 업종에 적용 한 후 결과는 표 5와 같다.

2행은 세 업종에 우도값만 고려하여 할당한 경우이다. 3행은 우도값과 업종모델들과 데이터들의 상태수를 고려하여 할당한 후의 우도의 합과 분류정확도를 보여주는데 상태수를 고려하기 전보다 높은 우도값과 정확한 군집결과를 보여준다.

표 4. 상태수와 우도를 고려한 군집화 알고리즘
Table 4. Clustering algorithm with number of states and likelihood

알고리즘 상태수와 우도에 따른 K-means 알고리즘

Input : 시계열 데이터들의 집합 $X=x_1,x_2,\ldots,x_N$
Output : 분할모델 $M=\lambda_1,\lambda_2,\ldots,\lambda_K$

// 분할모델의 군집 수 결정 :
BIC 측정에 의한 군집 수 K 결정

$K=1$
repeat
  임의의 선정된 군집 seed에 가장 높은 우도를 주는
  시계열데이터를 통해 모델 생성
  for $K=2$ to K do
  $N-1$의 시계열데이터들 중에서 기존에 생성된 모델에
  가장 거리가 먼(낮은 우도)를 주는 시계열데이터를 통해
모델 생성
  end for
//시계열 데이터들의 상태수를 고려한 우도에 따라 군집에 할당
  Continue ⇐ true
  while Continue do
  가장 높은 우도를 갖는 군집에 할당하고,
  다수의 군집들일 경우에, 우도의 차가 가장 큰 것보다
  작은 차이의 우도의 값을 갖는 경우의
  시계열데이터는 데이터 객체와 모델의 상태의 수를 고려하여
  같은 상태의 수를 주는 군집에 할당
  if 군집 내에 데이터의 움직임이 없으면 then
  continue 〈= false break
  else
  군집에 할당된 시계열데이터를 통해
  군집의 모델 매개변수를 갱신
  end if
  end while
  분할모델의 최대사후확률을 계산
  $K=K+1$
until 현재 분할모델의 최대사후확률 〈 이전 분할의 최대사후확률
Stop

표 5 상태 수 고려 전,후의 군집별 총 우도값의 합과 분류정확도
Table 5. Likelihood and classificatin Accuracy

	각 군집에 대한 우도값	분류 정확도
상태 수 고려 전	(-1044)+(-458.5627) +(-389.9725) = -1892.53	66.6 %
상태 수 고려 후	(-1044)+(-421.9409) +(-389.9725) = -1855.91	100 %

유사기반 군집화에서는 모델기반과 같은 실험데이터를 이용하여 해밍거리와 차원축소를 통해 변형된 데이터들 간의 거리측정을 통해 최소계산을 보증하는 랜드마크 방식과 구간상수근사[5]방식을 적용하였다.

데이터를 변환 후 임의의 하나의 데이터를 선정하여 기준모델로서 정하고 나머지 모든 데이터에 대한 거리측정 결과의 일부분을 표 6에서 보여주고 있다.

표 6에서 1열과 5열은 실험데이터의 전체 시퀀스 수를 나타낸다. 2, 3, 4와 6 ,7, 8열은 1번과 7번과 13번 시퀀스를 기준으로 전체 시퀀스들과의 거리측정의 결과이다.

각 시퀀스에서 1행은 해밍거리에 의한 거리측정을, 2행은 랜드마크에 의한 거리측정을, 3행은 구간상수근사기법에 의한 거리측정값을 나타낸다.

표 6 해밍거리, 랜드마크, 그리고 구간상수근사 의한 거리측정 테이블
Table 6. Distance by Hamming distance, landmarker and approximation of interval

시퀀스 No	1 s	7 s	13 s	시퀀스 No	1 s	7 s	13 s
1 s	0	4	1	10 s	14	12	13
	0	8	5		7	5	6
	0	4.69	2.79		2.93	2.98	2.66
2 s	15	13	14	11 s	17	15	16
	7	9	5		9	6	8

	2.98	4.21	2.90		4.90	2.90	3.29
3 s	17	15	16	12 s	9	9	8
	5	9	7		8	6	6
	2.10	4.07	1.99		5.40	2.00	4.13
4 s	10	14	11	13 s	1	3	0
	5	10	7		5	8	0
	2.62	3.75	2.17		2.79	2.95	0
5 s	12	16	13	14 s	13	13	12
	5	10	8		7	9	6
	3.31	5.08	3.10		3.33	3.76	2.54
6 s	16	20	17	15 s	20	16	20
	5	7	8		5	7	6
	1.96	4.34	3.27		3.83	3.44	3.03
7 s	4	0	3	16 s	15	14	17
	9	0	8		7	5	6
	4.68	0	2.95		4.30	2.10	2.37
8 s	14	12	13	17 s	13	13	14
	7	5	6		8	7	9
	3.79	2.43	2.22		4.68	3.17	3.22
9 s	16	14	17	18 s	13	11	12
	9	7	8		6	8	4
	5.85	3.17	3.91		4.44	2.84	1.01

거리측정의 결과를 살펴보면 수치가 작을수록 유사한 시퀀스를 보여주는데 같은 모델이 아닌 다른 모델에서 생성된 데이터가 유사한 형태를 보여주는 작은 값들이 많이 나타나고 있다.

표 7. 유사기반 적용 분류 정확도
Table 7. Classification accuracy by similarity

기법 \ 모델	전기전자	유통	건설
해밍거리	40%	60%	20%
랜드마크	40%	20%	20%
구간상수근사	55%	50%	44%

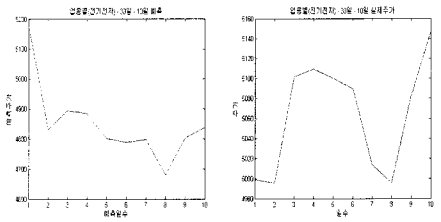
위의 결과들에 따라 군집화 결과를 통해 보면 유사기반 방식의 결과는 군집의 할당에 있어 모델기반 방식보다 표 7처럼 상당히 낮은 정확도를 보여주고 있다.

3.2 모델결정과 예측 적용

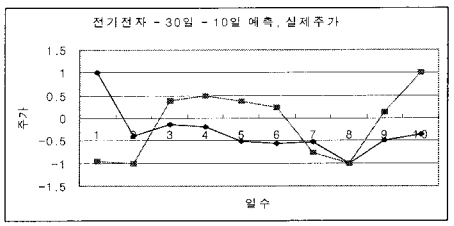
분할된 각 군집에 대하여 은닉마코프모델 결정 시 모델의 상태 수 결정을 위하여 베이지안정보기준을 적용한다. 이에 대한 유효성 검증결과는 [7]에서 제시되었다. 실제의 주가데이터에 적용하여 모델을 생성해보고 이를 예측문제에 적용하여 주가곡선의 패턴을 실질의 주가곡선과 비교하여 확인해 봄으로서 모델이 최적으로 생성된 것이지를 확인한다.

실험에 사용된 데이터는 위에서 군집화 된 결과 중 전기전자업종 군집 데이터와 별도의 KOSPI지수와 개별종목에서 삼성전자와 하이닉스의 데이터를 선정하여 실험하였다. 개별종목으로서 삼성전자는 대표적인 우량종목으로서 주가의 형성에 있어 외적환경요소에 덜 민감할 것으로 생각되어 좀 더 안정적인 예측모델이 가능할 것으로 생각되어 선택하였으며 하이닉스는 중형주로서 예측모델의 예측률이 우량종목의 예측률과 차이가 있는지를 확인함으로서 모델결정의 안정성을 확인하기 위하여 선정하였다.

그림 3에서는 군집 결과 중 전기전자업종의 군집의 시계열데이터를 대상으로 은닉마코프모델 생성을 통한 예측을 한 결과를 보여준다. (a)는 생성된 모델을 토대로 예측을 한 그래프이다. (b)는 모델생성에 적용된 데이터의 기간으로부터 10일 후까지의 실제주가의 그래프이다. (c)는 예측곡선과 실제곡선의 운동양태의 유사성을 확인하기 위하여 두 곡선의 정규화를 시킨 그래프이다.

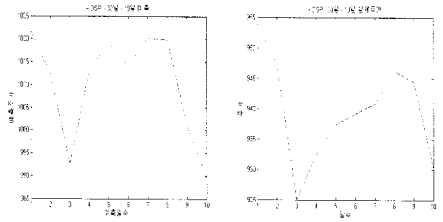


(a) 전기전자-예측 (b) 전기전자-실제

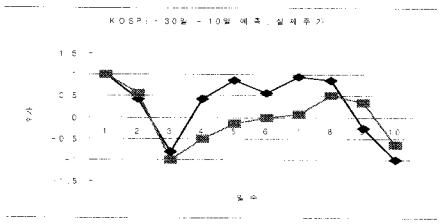


(c) 전기전자-예측과 실제의 정규화
그림 3. 전기전자업종 모델의 예측과 실제
Figure 3. Predication of model for electrical and electronic industry

예측과 실제의 그래프를 보면 1일부터 3일까지는 하락과 상승의 폭의 차이는 있지만 두 그래프 모두 하락 후 상승을 하는 경향을 보여준다. 3일부터 8일까지는 하락하는 패턴을 보여주고 8일부터 10일까지는 상승하는 유사한 운동양태를 보여준다.



(a) KOSPI-예측 (b) KOSPI-실제



(c) KOSPI-예측과 실제의 정규화 [2]

그림 4. KOSPI 지수의 예측과 실제

Figure 4. Predication for KOSPI index

그림 4에서 KOSPI 데이터로부터 생성된 그래프를 보여준다. (a)는 예측을 한 그래프이고 (b)는 실제 데이터의 그래프이다. (c)는 두 그래프를 정규화 한 후의 그래프를 보여준다. 1일부터 3일경까지 하락 후 3일부터 8일경까지 상승을 하다 8일경부터 하락을 하는 매우 유사한 운동양태의 곡선을 보여주고 있다.

본 연구에서는 예측곡선과 실제곡선의 운동양태 유사성을 확인하는 과정으로 두 곡선을 정규화[3]시킨 후, 두 곡선의 각 일별변화의 차이를 통해 두 곡선의 일별변화의 차이가 적을수록 에러율(예측곡선과 실제곡선의 차이)이 적은 것으로서 유사한 곡선임을 확인하는 방법으로 적용하였다. 정규화 시킨 에러율 평균치는 0.3에서 0.9사이에 존재함을 실험을 통해 확인하였다. 중앙값에 해당하는 평균치 임계값을 0.6으로 정하였을 때 에러율값이 임계값보다 낮을수록(0.3에 근접) 두 곡선의 운동양태가 유사함을 보여주었으며 높을수록 (0.9에 근접) 비유사성을 보여주었다. 위의 임계값을 기준으로 낮은 수치 사례에서는 두 곡선 즉, 실제곡선과 예측곡선의 일별 상승, 하락의 패턴이 불일치의 날짜가 10일 중 평균 2일로서 80%의 예측정확도를 확인할 수 있었다.

표 8. 실험의 시행횟수와 유사패턴 수

Table 8. Number of similar patterns

	전기 전자	KOSPI 지수	삼성 전자	하이닉스
시행횟수	20회	20회	20회	20회
유사패턴수	16회	15회	16회	15회
예측정확도	80%	75%	80%	75%

표 8은 전기전자업종 군집과 KOSPI, 개별종목에서 삼성전자와 하이닉스의 주가데이터를 대상으로 시행한 실험 횟수와 유사패턴이 발생된 횟수 그리고 예측정확도를 보여주는 표이다. 모든 데이터에서 유사패턴이 잘 예측되는 것으로 확인할 수 있었으며 개별종목에서 모델결정에 따른 유사패턴의 예측정확도가 대형우량주(삼성전자)와 중소형주(하이닉스)에 있어서 별 차이가 없음을 확인하였다.

3.3 동적시스템의 Interface 설계

동적시스템의 모델이 결정되면 사용자로부터 입력되는 정보를 통하여 맞춤정보의 제공이 가능하다. 동적시스템에서 정보제공의 효과를 높이기 위한 방법으로 그림 5와 같은 Interface를 제안한다.

우선 사용자가 시스템에 접속하면 시스템의 사용자 에이전트는 사용자 DB로부터 Class 정보를 추출하고 이를 정보서버의 서치 에이전트에게 맡긴다. 이때 사용자가 기존의 사용자가 아니고 Class가 정의 되어있지 않은 새 사용자라면 사용자 에이전트는 분석 에이전트에게 이를 알린다.

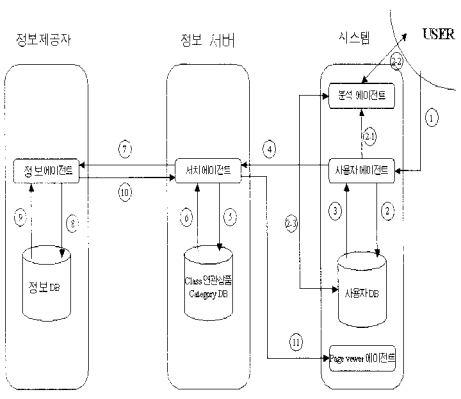


그림 5. 동적시스템 Interface 설계

Figure 5. Design of interface for the dynamic system

2) 2) 그림 4의 데이터 길이는 다음과 같다.
(a)실험데이터:05.03.02-05.04.13
(b)실제데이터:05.04.14-05.04.27

3) 정규화변환을 통해 시퀀스가 갖는 요소값의 절대크기를 무시할 있으며, 이는 요소값의 크기는 다르지만 변화하는 패턴이 유사한 시퀀스들을 파악하는데 유용하다. 정규화는 다음의 식에 따른다.

$$s' = \frac{s[i] - \frac{Max(S) + Min(S)}{2}}{\frac{Max(S) - Min(S)}{2}}$$

분석 에이전트는 새 사용자가 방문하는 page 등을 모니터링하고 사용자 DB의 데이터를 바탕으로 사용자를 Class화하고 이를 사용자 DB에 저장한다. 그럼 다시 사용자 에이전트는 사용자 DB로부터 이 새 사용자에 대한 Class 정보를 추출하여 서치 에이전트에게 보내고 서치 에이전트는 사용자 에이전트에게서 받은 클래스 정보에 해당하는 맞춤정보를 추출하여 정보제공자의 제공자 에이전트에게 각각의 정보에 대한 파일을 요청한다. 정보제공자의 제공자 에이전트는 그에 해당하는 정보파일에 대한 링크정보를 서치 에이전트에게 보내면 서치 에이전트는 이를 다시 시스템의 Page viewer 에이전트에게 보내고 Page Viewer 에이전트는 정보기재 란에 정보에 대한 링크를 연결하게 하여 동적 정보제공을 실현하게 된다.

4. 결론

본 연구에서는 동적시스템에서 발생되는 시계열데이터가 내재하는 현상을 쉽게 이해하고 설명하기 위한 모델링 방법론과 사용자에게 맞춤정보를 제공하기 위한 시스템의 CRM Interface를 제안하였다.

모델링 방법론은 군집화를 통한 모델결정을 이루는 두 과정으로 이루어진다. 실험에서는 실제의 KOSPI 시장에서 발생된 주가데이터를 사용하였다. 대용량 시계열데이터의 군집화 과정에서는 유사기반의 방식보다 모델기반 방식이 정확한 군집결과를 갖는 것을 확인하였다.

군집의 결정 후 생성된 은닉마코프모델을 통하여 일정기간의 일별주가곡선의 운동양태를 예측하였다. 실제의 주가곡선에 적용하여 예측 주가곡선의 운동양태가 실제 주가곡선과 유사함을 보여줌으로서 유효성을 확인하였다. 향 후 예측정확도를 보다 개선시키는 노력과 다양한 업종 및 종목으로 확대 연구가 필요하다.

참고문헌

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

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

[3] 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).

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

[5] J. Lin, E. Keogh, 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.

[6] L. Rabiner., "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. of IEEE 77, pp. 257-286, 1989.

[7] 전진호, 조영희, 이계성., "Temporal 데이터의 최적의 클러스터 수 결정에 관한 연구," 한국콘텐츠학회논문지 제6권 제1호, 23-30쪽, 2006년 1월

[8] T. Kosaka., S. Masunaga and M. Kuraoka., "Speaker-independent phone modeling based on speaker-dependent hmm's composition and clustering," In Proceeding of the Twentieth International Conference on Acoustics, Speech, and Signal Processing, pp. 441-444, 1995.

[9] L. R. Rabiner., C. H. Lee., B. H. Juang and J. C. Wilpon., "Hmm clustering for connected word recognition," In Proceedings of the Fourteenth International Conference on Acoustics, Speech, and Signal Processing, pp. 405-408, 1989.

[10] E. Dermatas and G. Kokkinakis., "Algorithm for clustering continuous density hmm by recognition error," IEEE Transactions on Speech and Audio Processing 4, pp. 231-234, 1996.

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

저 자 소 개

전 진 호(Jin-Ho Jeon)

1998년 : 명지대학교 경영정보학과경영학석사

2007년 2월 : 단국대학교 전자계산학과 이학박사

2003년 3월 ~ 현재 : 관동대학교 경영정보학부 겸임조교수

〈관심분야〉 : 데이터마이닝, CRM

이 계 성(Gye-Sung Lee)

1980년 : 서강대학교 전자공학과(학사)

1982년 : 한국과학기술원 전자계산학과(석사)

1994년 : Vanderbilt University 전자계산학과(공학박사)

1994년 ~ 1996년 : 대구대학교 전산정보학과 전임강사

1996년 ~ 현재 : 단국대학교 컴퓨터과학 전공 교수

〈관심분야〉 : 기계학습, 데이터마이닝