유사추론 기반 예측모형

장용식[a], 최윤정[b]

한신대학교 경상대학 e-비즈니스학과
경기도 오산시 양산동 411번지
Tel: +82-31-970-6421, Fax: +82-31-372-3343, E-mail:{[a]yschang@hs.ac.kr, [b]skywalkway@hanmail.net}

초록

본 연구는 비선형적인 시계열 자료로부터 최신 데이터와 유사한 사례를 탐색하여 미래를 예측하기 위하여 유사추론 기법을 이용한 예측 알고리즘을 제안한다. 기존의 연구들이 최신 데이터와 과거 사례와의 유사성을 비교하기 위해 유클리디언 거리 또는 평균제곱에러 등을 이용하나, 추세의 유사성을 고려하지는 않는다. 본 연구는 사례 구간 크기, 예측 오차, 평균차이 검증, 사례간 추세의 유사성 등 다차원적 유사추론 요인을 이용한 예측방법과 그 효과를 제시한다.

키워드:

유사추론, 사례, 예측

1. 서론

효과적 경영관리를 위한 주요업무 중의 하나는 과거 경영데이터를 분석하여 현재 경영활동에 반영하고, 미래에 대비하는 것이다. 데이터의 양이 방대하고 고려해야 할 변수들이 많을수록 미래예측을 위한 효과적 예측모형이 요구된다. 그러나, 다양한 내외부 환경변수를 예측모형에 반영하는 것은 쉬운 일이 아니다.

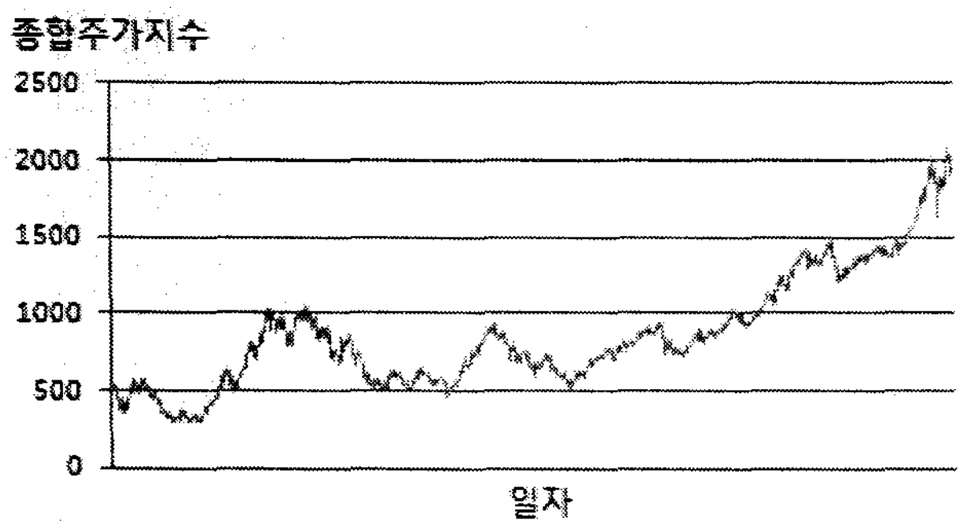

그림 1 - 종합주가지수의 변화(1997.11.1 ~ 2007.10.31)

선형 데이터에 대한 예측모형으로는 회귀모형 등이 유용하나, 설명이 복잡한 비선형 시계열 데이터의 예측을 위한 모형으로 ARIMA, 신경회로망, 사례기반 추론 등이 이용되고 있다.

그림 1은 최근 10년간의 종합주가지수의 변화를 나타내는 시계열 데이터를 나타낸 것이다. 많은 경제지수 및 기업경영 시계열 데이터들은 특정 기간 내에 유사한 패턴이 반복되는 경우가 많다. 특히, 비구조적이며 유사패턴이 반복되는 데이터의 경우, 경험적인 방법은 예측을 위한 문제 해결 대안이 되고 있다. 그러나, 이는 유사패턴을 발견하기 힘들기 때문에 사례 기반 추론은 유사사례 발견 및 예측에 잘 활용될 수 있다. 그러나, 최적 유사사례를 효과적으로 판별하는 것은 쉽지 않다.

기존의 사례 기반 연구들은 주로 유클리디언 거리 또는 평균제곱에러 등을 이용하여 각 비교 데이터들간의 오차를 구하고, 그로부터 유사성을 분석하였다. 그러나, 유사성은 다양한 관점에서 살펴볼 필요가 있다.

본 연구는 사례 기반의 일종으로 사례간 Feature mapping을 통한 유사추론을 위해 사례 데이터의 크기, 사례간 오차 및 평균 차이의 검증, 그리고 추세분석을 위한 기울기의 변화에 대한 유사성 등 다양한 척도에 관한 접근을 통해 보다 효과적인 유사사례를 검색하는 방법을 제시하고자 한다.

이를 위해 2장에서는 사례 기반 관련 문헌을 살펴보고, 3장에서는 유사추론 기반 예측 알고리즘을 제시하며, 4장에서는 제시한 알고리즘의 효과를 검증하기 위해 실험결과를 제시한다. 마지막으로 논문의 유용성과 한계를 살펴보면서 결론을 맺는다.

2. 문헌 연구

사례 기반 추론(CBR: Case-based Reasoning)은 사례의 표현, 색인, 사례의 검색, 유사성, 사례의 적용과 같은 주요 단계로 구성되며, 각 분야별로 많은 연구가 있다[6].

CBR은 의사결정을 지원하는 모델링 분야[3] 등 다양한 분야에 응용되고 있으며, 최근 과거의 유사사례로부터 미래를 예측하는 사례 기반 예측 분야에 많이 응용되고 있다.

CBR에 기초한 예측은 사례 기반 예측시스템을

구축하여 그 유용성과 회귀모형에 대한 우수성을 보인 연구가 있다[4]. 또한, 예측력을 높이기 위해 최적 결합 사례의 수를 결정하는 방법으로 최적화 수리모형을 이용하거나[2], 유전자 알고리즘 기법을 이용한 방법[1]이 있다.

유사추론(Analogical Reasoning)은 A와 B와의 관계는 C와 D와의 관계와 같다는 접근으로 A, B, C 사이의 특성으로부터 D를 결정해 가는 과정이다[5]. 유사추론은 CBR의 한 분야로서, Feature mapping을 통한 사례로부터 예측치를 유추할 수 있는 방법으로 이용될 수 있다.

기존의 사례 기반 예측은 유사도 측정을 위해 NN(Nearest Neighbour) 매칭 기능을 이용하고 있으며, NN 매칭을 위한 방법으로 대부분 유클리디언 거리, 평균제곱에러 등을 이용하고 있다. 보다 효과적인 예측을 위해서는 최적의 유사사례를 검색할 수 있는 다차원적인 관점에서의 유사도 측정 방법이 필요하다.

3. 유사추론 기반 예측알고리즘

새로운 사례(*NC*)와 과거 사례(OC_i)와의 Feature mapping을 통한 최적 유사사례 검색을 위해서 다차원적인 관점에서 유사도 측정 요인을 보면, 사례 구간의 크기, 비교 사례간 오차의 크기, 두 사례의 평균차이 비교, 추세의 유사성을 고려할 수 있다.

비교 사례간 오차의 크기를 재는 방법으로는 유클리디언 거리, 평균제곱에러(MSE: Mean Squared Error), 평균절대백분위예측오차(MAPE: Mean Absolute Percentage prediction Error) 등이 있으며, 본 논문에서는 MAPE를 이용하기로 한다. MAPE는 (1)식과 같다. MAPE가 유사도 범위 $MAPE_C$내에 있는 사례를 유사사례 후보로 고려할 수 있다.

$$\text{MAPE}= \frac{100}{m} \times \sum_j \frac{NC(j)-OC_i(j)}{NC(j)} \quad (1)$$

두 사례의 평균차이 비교는 Paired t-Test로 검증한다. *NC*와 OC_i의 평균차이 검증에서 T값이 t(m, 0.05)보다 작으면 95% 유의수준에서 귀무가설 H_0: $\mu_{NC}=\mu_{OCi}$가 채택되어, 평균차이가 없다고 볼 수 있다.

j 시점에서의 기울기는 j-1 시점 값 대비 j 시점 값과 j-1 시점 값의 차이에 대한 비율을 의미한다. j 시점에서 *NC*와 OC_i의 변화 기울기 $\theta(NC(j))$, $\theta(OC_i(j))$는 각각 (2), (3)식과 같다.

$$\theta(NC(j)) = \tan^{-1} \frac{NC(j)-NC(j-1)}{NC(j-1)} \quad (2)$$

$$\theta(OC_i(j)) = \tan^{-1} \frac{OC_i(j)-OC_i(j-1)}{OC_i(j-1)} \quad (3)$$

추세의 유사성은 각 사례의 기울기 차 $\theta(NC(j), OC_i(j))$로 판단할 수 있으며, (4)식과 같이 표현할 수 있다.

$$\theta(NC(j), OC_i(j)) = \text{Max}\{\theta(NC(j)), \theta(OC_i(j))\} - \text{Min}\{\theta(NC(j)), \theta(OC_i(j))\} \quad (4)$$

비교 사례간 추세의 허용 가능 범위를 θ_C로 하고 $\theta \le \theta_C$인 개수를 s_i라 하면, s_i/m가 유사도 범위 s_c 이내에 있는 사례를 유사사례 후보로 고려할 수 있다.

Feature mapping 요인을 이용한 유사추론 예측 알고리즘은 다음과 같이 네 단계로 이루어져 있다.

Step 1: *NC* 구간(m) 설정
Step 2: *NC*와 각 OC_i간의 유사성 비교
  Step 2.1: OC_i의 변환(TOC_i)
  Step 2.2: TOC_i의 변화구간 탐색
  Step 2.3: 변화구간 내 *NC*와 TOC_i의 유사성 비교
  Step 2.4: 변화구간 내 최적 TOC_i ($TOC_{i,H0}$)탐색
Step 3: 각 m별 유사성을 만족하는 다수 $TOC_{i,H0}$ 선정
Step 4: 모든 m에 대한 최적 $TOC_{i,H0}$ 선정 및 예측

■ Step 1: 새로운 사례(*NC*) 구간 설정

t 시점 이후의 예측을 위해 t 시점부터 t-(m-1)시점까지의 *NC* 구간의 크기(m)을 설정한다. 그림 2는 최근 10년간의 일별 종합주가에 대한 *NC*의 설정 예를 보여주고 있다.

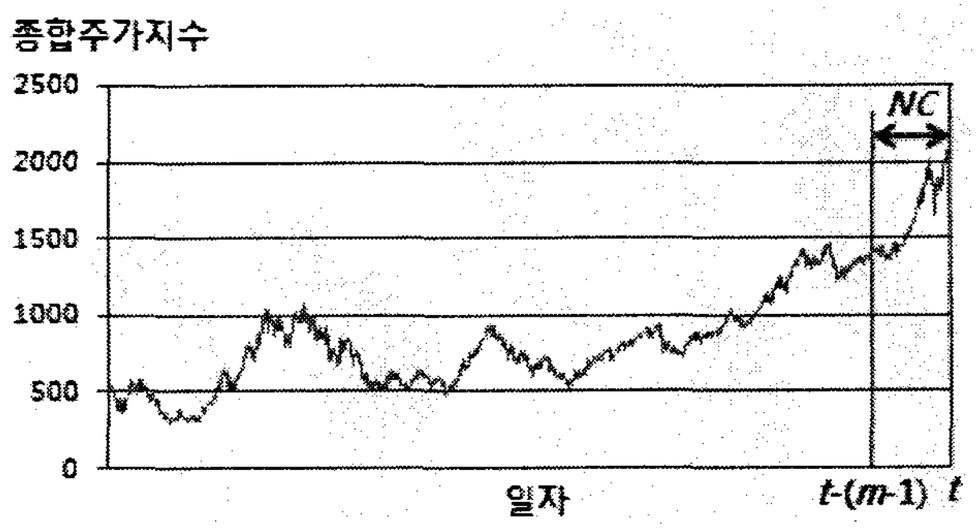

그림 2 – NC 구간 설정

■ Step 2: *NC*와 각 OC_i간의 유사성 비교

*NC*와 OC_i와의 유사성 비교는 다음 네 단계로 이루어진다.

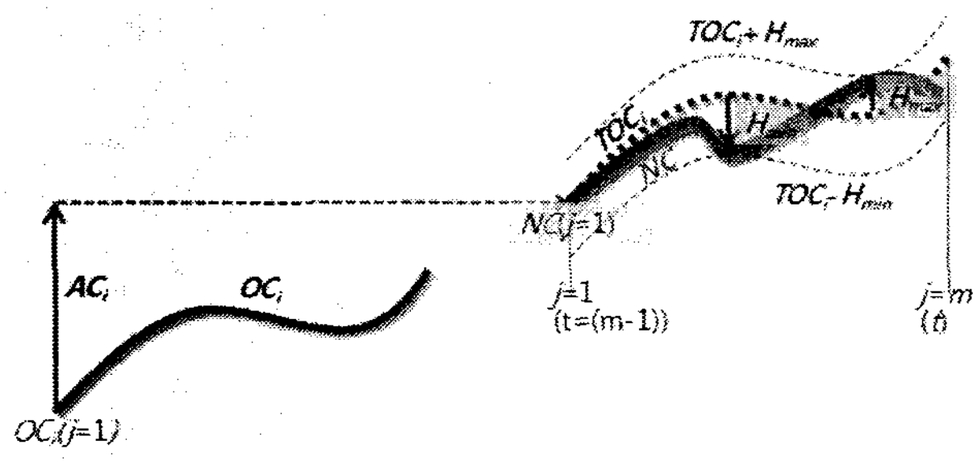

그림 3 - NC와 각 OC_i의 비교

Step 2.1: OC_i의 변환(TOC_i)

OC_i의 첫번째 비교 데이터 $OC_i(j=1)$를 $NC(j=1)$의 크기와 일치시키는 1차 조정계수 AC_i를 구하고, 1부터 m 구간 내의 모든 j번째 데이터에 대하여 AC_i를 곱하여 과거 사례 OC_i를 변환한다. AC_i는 (5)식, 변환된 과거 사례 TOC_i는 (6)식과 같다.

$$AC_i = \frac{NC(j=1)}{OC_i(j=1)} \tag{5}$$

$$TOC(j) = OC(j) \times AC_i \text{ for all } j \tag{6}$$

Step 2.2: TOC_i의 변화구간 탐색

TOC_i를 상하로 이동하면서 NC와 MAPE를 비교하기 위하여 2차 조정계수 H를 구한다. H의 상하한 구간은 TOC_i를 위와 아래로 이동하면서 NC와 교차점이 없이 접점 만을 가질 때까지 이며, 위로는 H_{max}, 아래로는 H_{min} 범위 내에 있다.

Step 2.3: NC와 변화구간 내 TOC_i와의 유사성 비교

변화구간 내에서 조정된 TOC_i 즉, $TOC_i+H(-H_{min} \le H \le H_{max})$와 NC에 대해 MAPE, Paired t-Test, 추세의 유사성을 계산한다.

Step 2.4: 변화구간 내 최적 TOC_i ($TOC_{i,H0}$)탐색

변화구간 내에서 모든 조정된 TOC_i 중에서 최소 MAPE를 갖는 최적 TOC_i인 $TOC_{i,H0}$를 탐색한다.

■ **Step 3:** 유사성을 만족하는 $TOC_{i,H0}$ 선정

모든 OC_i에 대해 유사성을 만족하는 $TOC_{i,H0}$ 탐색한다. 그림 4는 m=20인 비교 구간을 가진 NC에 대해 최소 MAPE를 갖는 3개 $TOC_{i,H0}$에 대한 OC_i의 예이다.

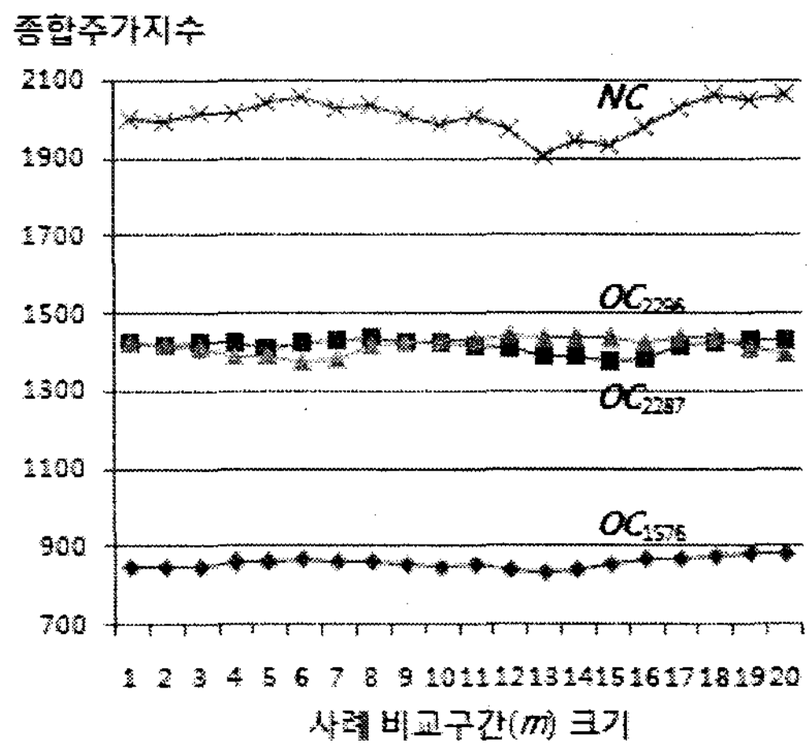

그림 4 - 유사 $TOC_{i,H0}$ 탐색

■ **Step 4:** 최적 $TOC_{i,H0}$ 선정 및 예측

Step 3에서 검색된 하나 이상의 유사사례를 이용하여 t+1 시점부터 $t+n$ 시점까지의 값을 예측한다. 그림 5는 하나의 $TOC_{i,H0}$를 이용하여 예측한 결과를 나타내고 있다. 예측 값은 $TOC_{i,H0}$에 대응하는 OC_i의 비교구간 이후부터 n개의 데이터를 이용하여 1차 조정계수 AC_i와 2차 조정계수 H_0의 값을 반영한 결과이다.

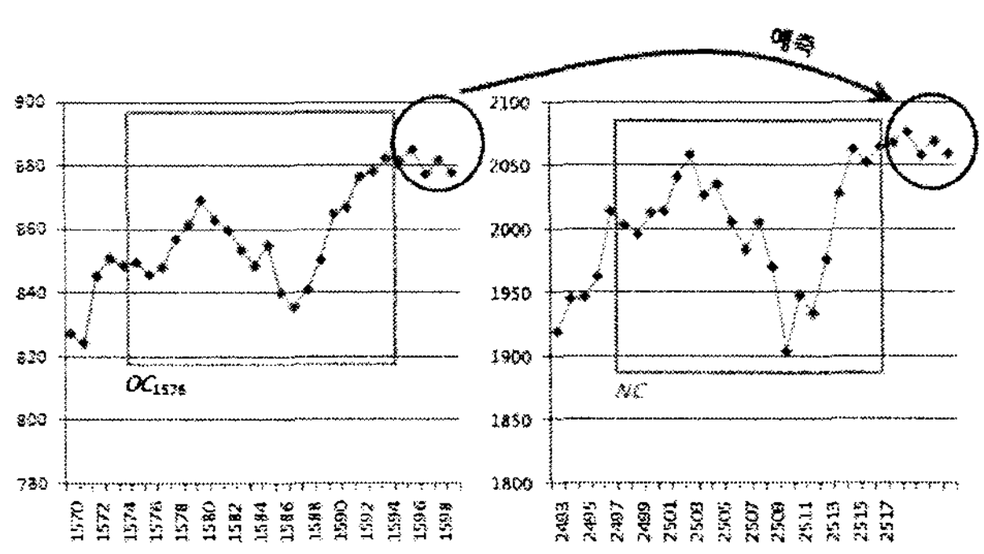

그림 5 – 최적 유사사례를 이용한 예측

4. 유사추론 기반 예측 알고리즘의 효과검증

유사추론 예측알고리즘에 대한 효과를 평가하기 위하여 자바 기반의 유사추론 예측알고리즘 평가시스템을 구현하였다. 테스트 데이터로는 1997년 11월 1일부터 2007년 10월 31일까지의 일별 한국 종합주가지수를 사용하였다. 최근 5개 데이터를 예측 구간으로, 그리고 그 이전 m개의 데이터를 NC의 구간으로 정하였다. m은 11부터 30개 구간으로 변화하였다. i번째 과거사례 OC_i는 1997년 11월 1일자부터 i번째에 있는 데이터에서 시작하여 m개의 구간으로 구성된다.

■ OC_i에 대한 MAPE와 Paired t-Test

그림 6은 m=20에 대한 NC와 총 2473개의 OC_i에 MAPE와 T값을 나타내고 있다. 사례간 MAPE가 감소하더라도 예측치 MAPE와 T값의 변화는 일정하지 않음을 보여주고 있다. 즉, 사례간 MAPE가 작다고 해서 예측치 MAPE 역시 낮은 것은 아님을 보여주고 있다.

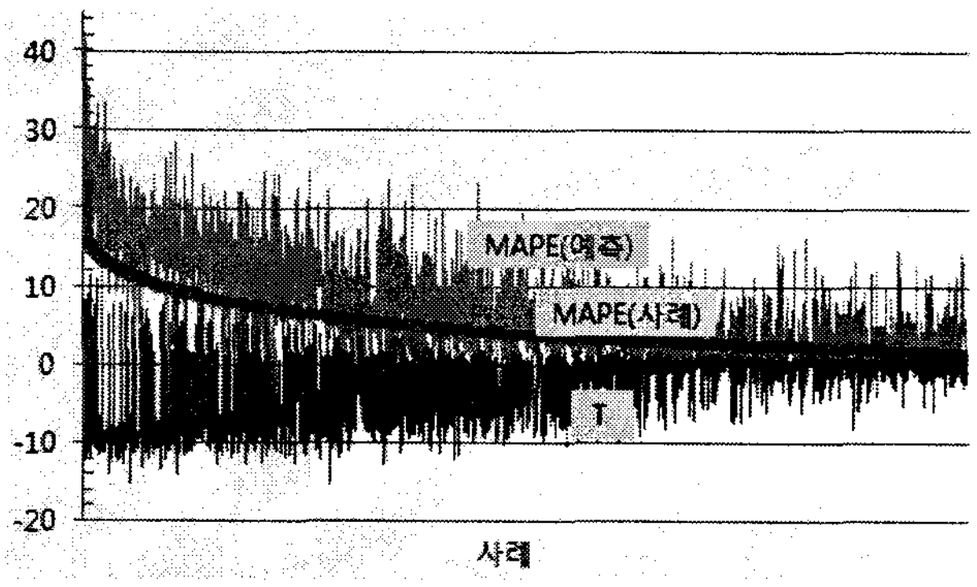

그림 6 - OC_i에 대한 MAPE와 Paired t-Test(사례 MAPE 역순)

그림 7은 사례간 $MAPE_C$가 2 이하, Paired t-Test 결과 95% 유의수준에서 평균 차이가 없는 총 259개 OC_i에 대해 예측치의 MAPE가 줄어드는 순으로 정렬한 것이다.

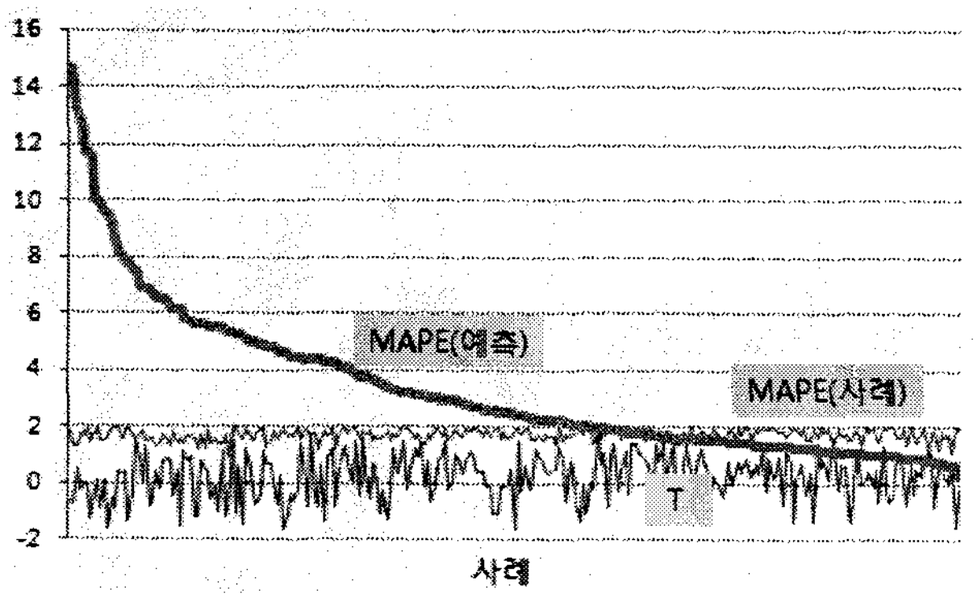

그림 7 – 유사성 범위 내의 OC_i에 대한 MAPE와 Paired t-Test(예측치 MAPE 역순)

■ 비교구간 m의 크기 변화에 따른 MAPE

비교구간 크기 m의 변화에 따른 MAPE의 변화를 보기 위하여 m을 11부터 30까지 변화하여 MAPE의 변화를 그래프로 나타내면 그림 8과 같다. m이 증가할수록 NC와 OC_i간의 사례 MAPE가 증가함을 보여주고 있으나, 예측치 MAPE도 같은 추세를 보이는 것은 아니다.
일정 허용범위내의 사례 중에서 예측결과의 MAPE가 낮은 사례를 최적 사례로 선택하는 것이 의미가 있음을 보여주고 있다. 그림 8에서는 OC_{1562}가 우선적으로 선정될 수 있다.

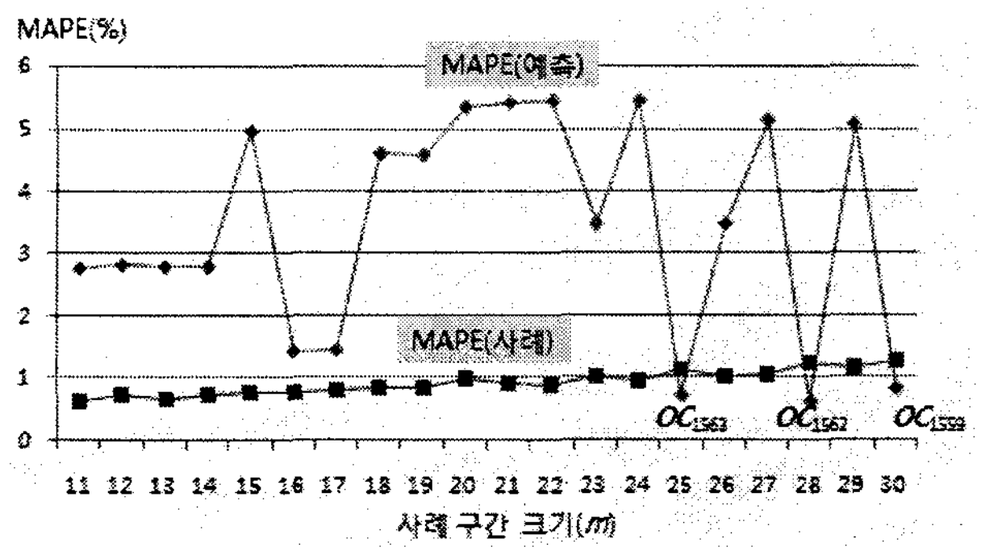

그림 8 – 비교구간 m의 크기 변화에 따른 MAPE

■ 추세의 유사성에 따른 MAPE

사례간 추세의 유사성을 나타내는 $\theta(NC(j), OC_i(j))$의 유사 개수에 따른 MAPE의 변화는 그림 9와 같다. m=16일 때 추세의 유사 개수가 14개를 넘으면 유사성 범위를 벗어나게 된다. 유사 개수가 1부터 14까지 증가할 때 사례간 MAPE는 증가하는 추세를 보이는 반면, 예측치 MAPE는 그렇지 않다. 사례 MAPE가 허용 가능한 범위 내에서 추세의 유사성이 높은 구간에서 예측치의 MAPE가 낮게 나타나는 경우가 있음을 알 수 있다. 따라서, 추세의 유사도가 높으면서 예측 MAPE가 낮은 사례를 유사사례로 선택하는 것이 바람직하다.

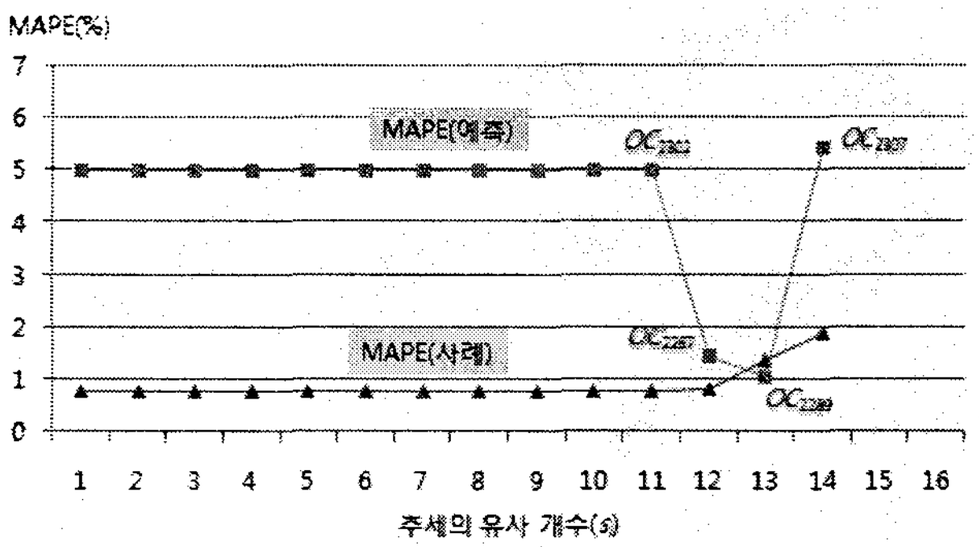

그림 9 – 추세의 유사성에 따른 MAPE

7. 결론

본 연구는 시계열 자료를 바탕으로 미래를 예측하는 방법으로 유사추론 알고리즘을 제안하였다. 유사추론에 의한 유사사례 검색은 Feature mapping 요인인 사례 비교구간의 크기, MAPE, 평균차이 비교에 대한 가설검증, 그리고 사례간 추세의 유사성을 계산하여 유사정도를 분석하고, 유사사례 허용 범위 내의 사례 중에 최적의 유사 사례를 선정해서 예측에 활용한다.

향후, 다차원적 유사도 요인을 통합한 종합지표가 필요하다. 또한, 경제지표는 물론, 생산계획, 판매, 수요예측 등을 위한 제조 및 유통 등 다양한 산업분야의 시계열 자료에 활용하여 예측력의 효과를 분석할 필요가 있으며, 비전문가의 활용을 도와줄 수 있는 유사추론 기반 예측시스템의 개발이 요구된다. 한편, ARIMA, 신경회로망, 유전자 알고리즘, SVM 등의 다양한 기법과 비교하여 예측력의 우수성을 보일 필요가 있다.

본 연구는 내외부 환경변수를 반영하지 못한 한계가 있으며, 예측력을 높이기 위해서는 환경변수의 값을 추세에 반영하는 혼합적 연구방법이 필요하다.

감사의 글

본 연구는 한국과학재단 특정기초연구(R01-2006-000-10303-0(2006)) 지원으로 수행되었음.

참고문헌

[1] 김경재, 안현철, 한인구. (2006). "유전자 알고리즘을 이용한 사례 기반 추론 시스템의 최적화: 주식시장에의 응용", *경영정보학연구*, 제16권, 제1호, pp.71-134.

[2] 이훈영, 박기남. (1999). "사례기반예측시스템의 정확한 예측을 위한 최적 결합 사례개수결정방법에 관한 연구", *경영학연구*, 제27권 제5호, pp. 1239-1252.

[3] Chang, Y.S., and Lee, J.K. (2004), "Case-based Modification for Optimization Agents: AGENT-OPT," *Decision Support Systems*, Vol. 36, No. 4, pp.355-370.

[4] Lee, H.Y. (1994). "A Case-based Forecasting System," *Journal of the Korean OR/MS Society*, Vol. 19, No. 2, pp. 199-215.

[5] Liang, T-P., and Konsynski, B.R. (1993). "Modeling by Analogy," *Decision Support Systems*, Vol. 9, pp.113-125.

[6] Marir, F., and Watson, I. (1994). "Case-based Reasoning: a Categorized Bibliography," *The Knowledge Engineering Review*, Vol. 9, No. 4, pp.355-381.