응용통계연구 제 18권 1호, 2005년, pp. 137–146

시계열 예측의 변형된 ENSEMBLE ALGORITHM

김연형 [1] 김재훈 [2]

요 약

신경망은 전통적인 시계열 기법들에 비해 대체적으로 예측성능의 우수함이 입증되었으나 계절성과 추세성을 갖는 시계열자료에 대해 예측력이 떨어지는 단점을 가지고 있다. 최근에는 Ensemble 기법인 Bagging Algorithm과 신경망의 혼합모형인 Bagging Neural Network이 개발되었다. 이 기법은 분산과 편향을 많이 줄여줌으로써 더 좋은 예측을 할 수 있는 것으로 나타났다. 그러나 Ensemble 기법을 이용한 예측모형은 시계열자료를 적합시키는데 있어 초기부여확률 및 예측자 선정시의 문제점을 가지고 있다. 이에 본 연구에서는 이러한 문제점을 해결하고 더불어 예측력을 향상시키기 위한 방법으로 초기부여확률이 균일분포가 아닌 순차적인 형태의 분포를 제시하고 신경망을 예측자로 활용한 변형된 Ensemble Algorithm을 제안한다. 또한 예측모형의 평가를 위해 실제자료를 가지고 기존 예측모형들과 제안한 방법을 이용하여 예측하고 각 MSE의 비교를 통하여 예측정확도를 알아보고자 한다.

주요용어: 신경망, 초기부여확률, 순차적 초기부여확률, 변형된 Ensemble Algorithm

1. 서론

Box-Jenkins기법을 이용한 시계열 예측시 시계열자료에 적합한 모형은 최소 모수의 원리에 따라 도출하여야 한다. 도출된 모형이 불필요하게 복잡하면, 이 모형에 포함된 모수의 수가 많아지므로 모수추정에 따르는 오차가 커지게 되어 예측을 정확하게 할 수 없다. 또한 Box-Jenkins의 방법은 검진의 단계에서 모형이 타당하지 않았다고 판정되었을 경우에 식별, 추정, 그리고 검진의 한 주기에서 얻은 모든 정보가 무용해지는 단점을 가지고 있었으며 이에 다른 방법과의 연계를 연구하게 되었다. 이에 따라 전통적인 시계열 예측보다 좋은 방법론들이 등장하게 되었다. 대표적으로 신경망이 있는데 신경망은 다양한 시계열자료의 예측에 많이 이용됐다. Box-Jenkins 방법과의 비교에서도 장기예측의 경우는 신경망의 예측이 높다는 것이 실증되었다(Lapedes 1987, Zhang 1998). 특정 시계열자료가 분기나 월별일 경우 전통적인 시계열 기법에 비해 신경망이 예측 성능이 우수하며(Hill et al. 1996), 신경망을 이용한 방법이 윈터스법과 회귀분석에 비해 예측 성능이 뛰어남을 보였다(Alon et al. 2001). 하지만 다양한 응용 분야에서 ARIMA가 평균절대백분율오차 측면에서 신경망보다 같거나 좋다고 보였다(Kang 1991). 따라서 신경망은 적용 대상 시계열에

1) (560-759) 전주시 완산구 효자동 1200번지, 전주대학교 데이터정보학과, 교수
E-mail: yhkim@jj.ac.kr
2) (560-759) 전주시 완산구 효자동 1200번지, 전주대학교 교양학부 강의전담
E-mail: muggeby@jj.ac.kr

따라 성능이 다름을 알 수 있다. 이는 신경망 모형이 실제 적용시 구축의 노력과 시간이 많이 들며 계절성과 추세에 약하다는 단점이 나타났다. 최근에는 Ensemble 기법인 Bagging Algorithm에 혼합되어 사용되어지고 있다. 이 기법은 전체적인 분산과 편향을 줄여주는 것으로 알려져 있다. 또한 전통적인 시계열 예측 기법들에 비해 우수한 예측성능이 나타났으며, 변동이 큰 시계열 자료에는 시간이 경과한 후 신경망의 재학습을 통한 예측이 가능한 장점을 나타냈다(Freund & Schapire 1996). 그러나 Ensemble 기법은 예측자 및 초기부여확률이 시계열 자료에 적합시키기에는 여러 가지 제약을 가지고 있다. 특히 초기부여확률을 균일분포로 제안하고 있다는 점이다. 이는 시계열자료의 특성상 최근의 자료가 미래시점의 예측에 더 많은 영향을 준다는 점을 간과한 것으로 시계열자료에는 적합하지 않다. 따라서 본 연구에서는 시계열 자료에 적합하고 예측력을 향상시킬 수 있는 방법으로 변형된 Ensemble Algorithm에 대해 제안하고자 한다.

2. 변형된 Ensemble Algorithm

Ensemble Algorithm의 초기 분포 형태가 균일분포가 아닌 순차적인 분포형태와 예측자를 신경망 모형으로 적용시켜 시계열 모형에 적합시키고자 한다. Ensemble Algorithm들은 초기 분포를 $p_i = 1/T,\ i = 1, \cdots, T$인 균일분포로 설정한다. 즉 원자료의 추출확률을 동일하게 둔다는 의미이다. 그러나 시계열 자료는 특성상 최근자료가 과거 먼시점의 자료에 비해 예측값에 영향을 많이 끼치므로 초기 확률값을 동일하게 배분하여 표본을 추출하는데에는 최근자료의 가중값이 전혀 고려되지 않고 있다. 따라서 본 연구에서는 초기부여확률을 순차적인 형태인 $p_i = \frac{2i}{T(T+1)},\ i = 1, \cdots, T$로 놓아 최근자료에 높은 가중값을 두어 추출확률을 높였다. 이 초기 분포값은 이미 알려진 분포가 있다면 매우 유용하겠지만 일반적으로 시계열 자료의 모집단의 분포는 가정에 의해서만 존재하며 실제 자료들은 정규분포를 하지 않는다. 제안한 분포값은 순차적인 함수의 형태인 확률함수이며 함수의 조건을 만족한다. 재표본을 통한 확률값의 갱신으로 인해 최근자료의 추출 확률은 더욱 더 높아질 것이다.

시계열 자료를 신경망에 입력하여 학습을 시킬 때 원 자료 그대로 신경망에 입력할 수 없다. 일반적인 원자료는 전이함수인 시그모이드 함수의 범위를 넘어서기 때문에 최소값이 0이고 최대값이 1인 시그모이드 함수를 통과하기 위해서는 원 자료를 0-1 사이의 값으로 변환해야 한다. 이 과정이 표준화 변환이다. 통상적으로 사용되는 방법은 단순 선형 변환과 평균 및 표준편차에 의한 변환이 있으나 변환이 간단한 단순 선형 변환방식이 더 많이 사용되고 있으며, 단순 선형 변환식은 다음과 같다.

$$x_i = f_{\min} + (f_{\max} - f_{\min}) \times \frac{(t_i - \min(t_i))}{(\max(t_i) - \min(t_i))}$$

여기에서 x_i는 표준화된 값이며, $f_{\min}$과 $f_{\max}$는 전이함수의 상대 최소값과 최대값이다. t_i는 관측값이고 $\min(t_i)$는 전체관측값의 최소값이며 $(\max(t_i) - \min(t_i))$는 관측값의 범위이다. 이렇게 변환된 자료는 3개의 집합인 훈련집합(training set), 검증집합(testing set), 평가집합(validation set) 으로 나눈다.

훈련집합은 신경망을 학습을 위한 자료이고, 검사집합은 신경망이 얼마나 학습이 잘되었는지를 검사한다. 통상적으로 학습자료의 10-30%를 사용하며 또는 전체 학습자료를 나누어 구성하는 경우도 있고 전체 학습자료에서 표본추출하는 경우도 있다. 그리고 평가집합은 신경망의 학습 및 검사시 제외된 자료들로 학습된 신경망의 예측력을 평가하기 위해 필요한 자료집합이다. 각 자료의 크기는 평가집합의 경우 신경망의 입력층내의 처리요소의 수, 즉, 입력벡터의 크기 만큼이다. 예를 들어 향후 10기간을 예측하려고 한다면 평가집합의 요소의 수는 10이 된다. 다른 방법으로 시계열을 학습집합과 평가집합 2개로 나누어 사용하기도 한다.

실증분석시 단순선형변환을 통한 일반화 처리를 하였으며 자료의 형태가 월별자료이므로 평가용 집합을 최근의 12개월의 예측을 위해 12개로 선정하였으며 나머지 자료는 훈련집합과 검증집합으로 나누었다. 원자료를 훈련집합, 검증집합, 평가집합으로 나눈 후 훈련집합을 다음과 같이 R개의 재표본 훈련 입력쌍으로 나타낸다.

$$
\begin{aligned}
&(\mathbf{x}_1, \mathbf{p}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{p}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_R, \mathbf{p}_R, \mathbf{y}_R) \\
&\text{입력벡터 } \mathbf{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jN})^t, \\
&\text{목표벡터 } \mathbf{y}_j = (y_{j1}, y_{j2}, \cdots, y_{jN})^t, \\
&\text{초기부여확률벡터 } \mathbf{p}_j = (p_{j1}, p_{j2}, \cdots, p_{jN})^t.
\end{aligned}
$$

여기에서 j는 재표본수에 관한 첨자이며 N개의 훈련 자료를 추출한 것이다. 예측자로는 시계열 자료가 일반적으로 비선형 관계를 가지고 있기 때문에 비선형문제에 적합한 신경망을 사용한다. 추출된 첫 번째 입력벡터와 초기부여확률벡터를 가지고 신경망을 통한 0에서 1사이의 실수값을 예측하는 학습을 실시한다. 일반적인 신경망의 초기 가중값은 임의의 값을 선택하지만 여기에서는 신경망의 초기가중값을 각각의 초기부여확률을 초기 가중값으로 선택하게 하였다. 신경망을 통해 예측값을 구한 후 검증집합을 통한 오차와 학습률을 계산하고 다음과 같이 확률분포를 계산한 후 원 자료의 확률을 갱신시킨다.

$$
\begin{aligned}
\epsilon_j &= y - \hat{y}_j, \qquad \alpha = \frac{1}{2}\ln((1-\epsilon_j)/\epsilon_j) \\
p^u_{jk} &= (p_{jk}\exp(-\alpha))/Z_j, \qquad \text{여기서, } Z_j = \sum_{k=1}^{N} p_{jk}\exp(-\alpha).
\end{aligned}
$$

여기에서 ϵ_j은 추출된 훈련집합의 j번째 입력쌍의 예측오차이며 α는 학습률이다. 첨자 "u"는 변경된 후를 지칭한다. 이 변경되어진 p^u_{jk}값을 원 자료의 p_{jk}값에 갱신시켜 다시 전체의 확률값을 다음과 같이 계산하여 준다.

$$
p_{jk} = p^u_{jk} / \sum_{k=1}^{N} (p^o_{jk} + p^u_{jk})
$$

여기에서 첨자 "o"는 변경되기 전을 지칭한다. 원자료의 초기부여확률값의 갱신이 이루어지면 위와 같은 과정을 반복한다. 이제 재표본을 통하여 각각 예측값을 구하며 최종 예측

값 및 MSE는 다음과 같이 구해진다.

$$\hat{y} = \left(\sum_{j=1}^{R} \hat{y}_j\right)/R, \quad R: \text{ 재표본수, } \hat{y}_i: \text{ 각각의 예측값}$$

$$MSE = \frac{\sum(y-\hat{y}_j)^2}{N}$$

여기에서 최종예측값들과 MSE를 기존의 Box-Jenkins 및 신경망 방법과의 비교를 통해 제안된 알고리즘이 시계열 자료에 우수한 예측력을 가질 수 있는 모형임을 보이고자 한다.

3. 실증분석

이 절에서는 여러 가지의 시계열 자료를 이용하여 Box-Jenkins, Neural Network과 제안된 방법과를 비교하고자 한다. 신경망이 시계열의 계절성이 존재하면 계절성 때문에 모형의 적합에 문제가 되므로 계절성을 제거하고 또한 추세성분이 있으면 추세제거를 위한 1차 차분 상태의 자료로 분석을 실시하였으며 Box-Jenkins 방법으로 AR(1), MA(1), ARIMA(1,1,0)으로 식별되는 모형들을 가지고 모의실험을 실시하였다.

첫 번째 자료인 우리나라 실업률 자료는 1982년 1월부터 2004년 2월까지의 자료로서 그림 3.1에서 1998년 이전까지는 평균을 중심으로 약간 비정상 시계열로 보이나 특히 1998년 이후 IMF로 인한 실업률의 급상승으로 불안한 형태가 나타나 비정상 시계열로 나타나고 있다. Box-Jenkins방법으로 2003년 3월부터 2004월 2월까지의 자료의 예측시 $\log(y_t)$ 변환과 추세성분 및 계절성을 제거하기 위한 차분을 한 후 ARIMA(1,1,0)모형으로 식별 및 예측이 되는 시계열 자료이다.

두 번째 자료는 우리나라 1990년 1월부터 1994년 12월까지 치즈소비량을 나타내는 데이터이다. 그림 3.2의 시계열 도표로 확인한 결과 관측된 시계열은 시간이 흐름에 따라 큰 변화를 볼 수 없어 정상시계열임을 알 수 있으며, Box-Jenkins방법으로 식별한 결과 자기상관함수는 시차가 증가함에 따라 지수적으로 감소하고 부분자기상관함수는 시차 1에서만 양으로 유의한 값을 가지며 나머지 시차에서는 신뢰한계내에서 나타나 시차 1이후에 절단되어 있는 형태인 AR(1) 모형 $y_t = 0.56y_{t-1} + a_t$으로 식별되어진 시계열 자료이다.

세 번째 자료는 1993년 1월부터 2000년 11월까지의 입직률을 나타낸 자료이며 그림 3.3에서 보는 바와 같이 비정상시계열이며 계절성은 나타나지 않았다. 대수변환과 1차 차분을 한 후 모형을 식별한 결과 ARIMA(0,1,1) 모형으로 식별이 되어지는 자료이다.

네 번째 자료는 그림 3.4의 형태로 신경망을 이용한 시계열 예측을 위한 대표적인 벤치마크 자료 베이스인 Makridakis 경쟁대회의 자료중 하나이며 1991년 9월부터 1998년 10월까지의 Panter로 명명된 자료(http://www.ms.ic.ac.uk/iif/data/m2comp/m2comp.htm)로서 신경망의 예측이 Box-Jenkins 방법보다 더 좋은 예측을 하는 것으로 알려진 자료이다. 이 자료는 비정상 시계열이어서 대수변환과 차분을 실시한 후 ARIMA(1,1,0)으로 식별이 되어진다.

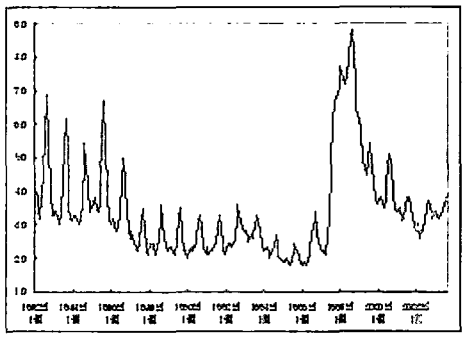

그림 3.1: 실업률

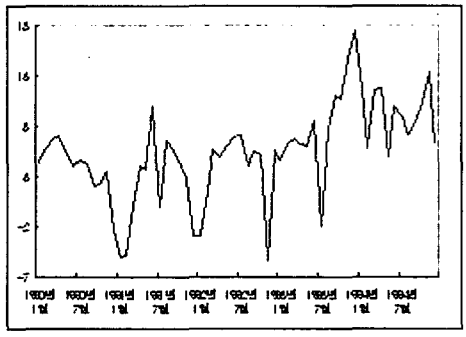

그림 3.2: 치즈 소비량

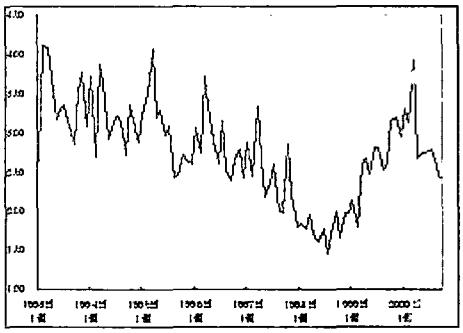

그림 3.3: 입직률

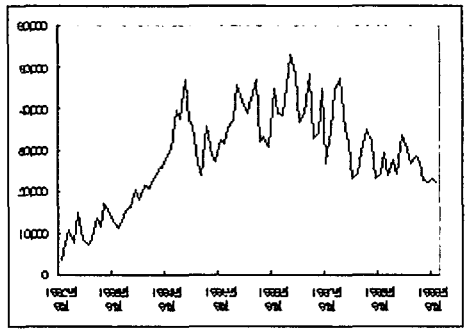

그림 3.4: Makridakis 경쟁자료 중 Panter 자료

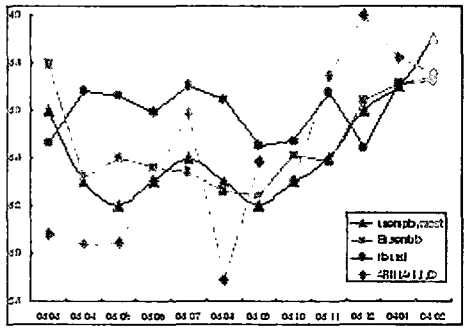

그림 3.5: 실업률 자료에 대한 여러 방법의 예측비교

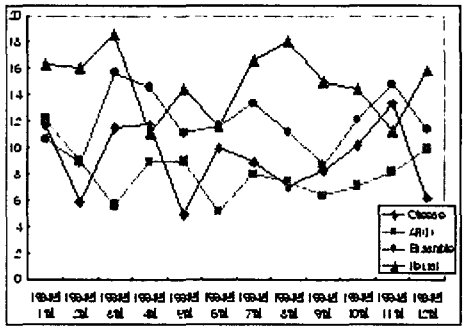

그림 3.6: 치즈 소비량 자료에 대한 여러 방법의 예측비교

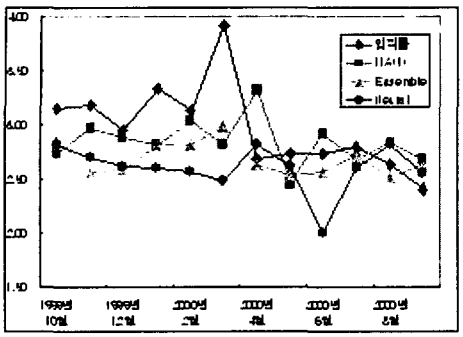

그림 3.7: 입직률 자료에 대한 여러 방법의 예측비교

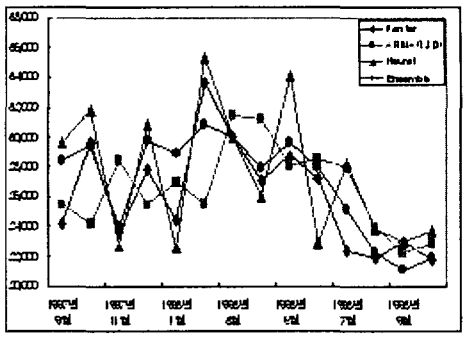

그림 3.8: Panter 자료에 대한 여러 방법의 예측비교

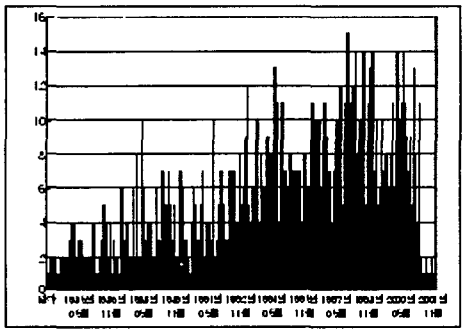

그림 3.9: 실업률 자료의 100번의 재표본시의 추출 빈도

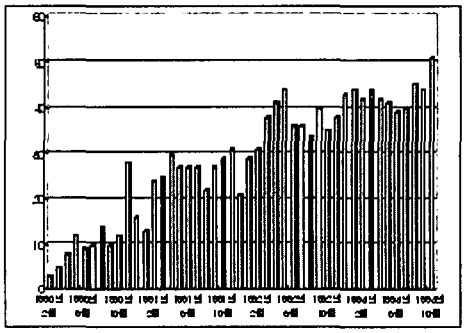

그림 3.10: 치즈 자료의 100번의 재표본시의 추출 빈도

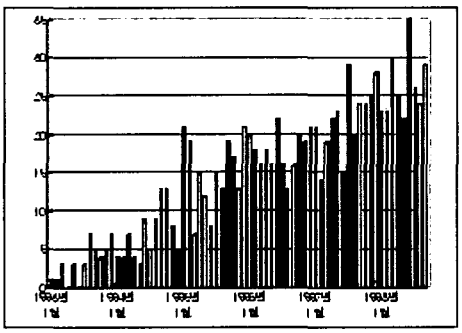

그림 3.11: 입직률 자료의 100번의 재표본시의 추출 빈도

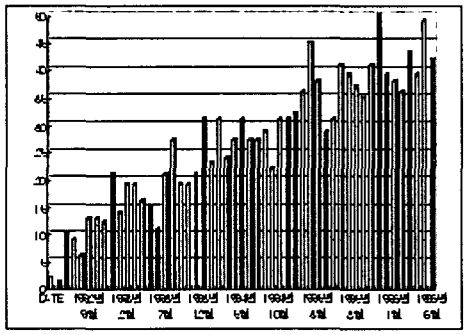

그림 3.12: Panter 자료의 100번의 재표본시의 추출 빈도

표 3.1: 4가지 자료에 대한 여러 방법의 MSE비교

실업률				치즈소비량			
날 짜	ARIMA(1,1,0)	Neural	Ensemble	날 짜	AR(1)	Neural	Ensemble
2003/3	3.08	3.46	3.79	1994/1	12.199	16.290	10.641
4	3.03	3.68	3.32	2	8.964	16.130	8.800
5	3.04	3.66	3.40	3	5.616	18.620	15.692
6	3.30	3.58	3.36	4	8.851	11.107	14.548
7	3.58	3.70	3.34	5	8.964	14.436	11.090
8	2.88	3.64	3.26	6	5.162	11.587	11.672
9	3.38	3.45	3.24	7	8.000	16.640	13.365
10	3.30	3.46	3.41	8	7.375	18.100	11.184
11	3.74	3.67	3.38	9	6.354	15.014	8.740
12	4.00	3.44	3.64	10	7.035	14.428	12.098
2004/1	3.82	3.71	3.71	11	8.113	11.256	14.820
2	3.75	3.72	3.74	12	9.872	15.898	11.350
MSE	0.076	0.070	0.011	MSE	12.265	51.040	12.097
입직률				Panter			
날 짜	ARIMA(0,1,1)	Neural	Ensemble	날 짜	AR(1,1,0)	Neural	Ensemble
1999/10	0.1644	0.1052	0.0713	1997/9	25,502	29,591	28,432
11	0.0436	0.2350	0.3944	10	24,143	31,780	29,348
12	0.0045	0.1082	0.1339	11	28,476	22,723	23,917
2000/1	0.2680	0.5254	0.2539	12	25,456	30,894	29,730
2	0.0105	0.3311	0.1049	1998/1	26,983	22,565	28,878
3	1.1878	2.0704	0.8606	2	25,564	35,296	30,766
4	0.4134	0.0196	0.0017	3	31,450	29,950	30,054
5	0.0796	0.0173	0.0447	4	31,225	26,045	27,951
6	0.0323	0.5471	0.0300	5	28,084	34,094	29,654
7	0.0123	0.0371	0.0050	6	28,567	22,940	27,957
8	0.0377	0.0313	0.0156	7	27,876	28,206	25,190
9	0.0739	0.0193	0.0499	8	23,912	23,891	22,291
MSE	2.3280	4.0471	1.9659	9	22,283	22,943	21,091
				10	22,942	23,739	21,884
				MSE	13,576,638	9,946,807	4,525,010

모의 실험 결과 그림 3.5에서 그림 3.8까지의 도표는 평가집단을 이용한 예측을 한 것이다. 모의실험은 각 자료의 식별된 모형에 의해 예측을 하였으며 신경망과 제안한 방법은 한 번 학습시 하나의 추정값을 얻기 때문에 각각의 시점을 예측하였다. 표 3.1은 각각의 자료에 대해 분석방법들의 MSE값 비교이다. 4가지 자료에 대해서 살펴보면 제안된 방법이 가장 효율적인 것으로 나타나고 있다. 100번의 재표본시에도 그림 3.9에서 그림 3.12까지의

그림상에서 최근자료의 추출률이 확연히 높게 나타나고 있음을 알 수 있다. 이는 초기확률 부여시 시계열 자료의 특성에 따라 최근의 자료가 더 자주 추출되도록 가중값의 부여에 따른 것으로 반복수가 많을 수록 더 자주 추출 되어 예측에 영향을 주고 있음을 알 수 있다.

4. 결 론

시계열 자료의 선정, 자료의 변환, 신경망 훈련을 위한 훈련집합 자료의 선정, 신경망의 구조, 재표본 추출을 통한 예측 등에 대해 연구하고 또한 구현을 통해 제안한 방법이 기존의 방법보다 우수하다는 것를 증명하였다. 제안한 방법에서 훈련집합 자료의 수는 자료의 특성에 따라 12-14개로 선정했으며 이를 통하여 재표본 추출을 100번 하였을 경우 초기부여확률의 효과로 최근의 자료가 더 많은 추출이 되었다. 제안한 방법은 기존의 신경망과는 달리 시계열 자료의 계절성, 추세성과 상관없이 예측이 가능하였으며 다른 분석방법에 비해 향상된 예측력을 나타났다. 제안한 방법은 반복적인 학습 및 추출확률의 갱신을 통하여 예측을 하기 때문에 새로운 관측값이 입력되면 새로운 학습을 통해 새롭게 예측을 함으로 모형의 식별과정이 없다는 장점을 가지게 된다.

기존의 여러 분석 방법과의 예측정확도의 비교 판단 기준으로 MSE를 이용하여 실증분석한 결과 ARIMA(1,1,0), AR(1), ARIMA(0,1,1)로 각각 식별되는 모형의 자료들에 대해서 제안한 방법이 다른 방법들보다 예측력이 높게 나타나고 있다. 또한 추출되는 표본의 분포형태가 최근자료의 추출빈도가 많았음을 알 수가 있었다. 이는 초기부여 확률의 영향으로 더 많은 최근의 값들이 미래 시점에 영향을 주고 있어 예측력을 향상시키고 있음을 알 수 있다.

앞으로 재표본 추출 횟수에 대한 추후 연구가 더 필요할 것으로 생각되며, 단변량시계열의 다른 모형과 다변량 시계열로의 확장이 필요하며 예측자로서 신경망의 구조형태의 변형 및 훈련집합 자료와 검증집합 자료의 수에 대한 연구가 더 필요할 것이다.

참고문헌

김연형 (2001). 시계열예측, 형설출판사, 서울.

송규문 (2001). 시계열자료에 대한 분석방법의 비교, <수리과학논집>, 21호, 49-61.

지원철 (1999). 신경망을 이용한 시계열의 분해분석, <대한산업공학회지>, 25권 1호, 111-124.

Alon, I., Qi, M. and Sadowski, R. J. (2001). Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods, *Journal of Retailing and Consumer Services* **8**, 147-156.

Drucker, H. (1999). *Boosting Using Neural Networks : Combining Artificial Neural Nets : Ensemble and Modular Learning*, Springer Verlag, 51-77.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of the 13th International Conference on Machine Learning.*

Hill, T., O'Connor, M. and Remus, W. (1996). Neural network models for time series forecasts, *Management Science* **42**, 1082-1092.

Kang, S. (1991). *An Investigation of the Use of Feedforward Neural Networks for Forecasting*, Ph.D. Dissertation, Kent State University. Kent Ohio

Lepedes, A. and Farber, R. (1987). Nonlinear signal processing using neural networks : prediction and system modeling, *Los Almos National Laboratory Report*, LA-UR-87-2662.

[2004년 8월 접수, 2004년 11월 채택]

Time Series Forecasting Based on Modified Ensemble Algorithm

Yon Hyong Kim [1] Jae Hoon Kim [2]

ABSTRACT

Neural network is one of the most notable technique. It usually provides more powerful forecasting models than the traditional time series techniques. Employing the Ensemble technique in forecasting model, one should provide a initial distribution. Usually the uniform distribution is assumed so that the initialization is noninformative. However, it would be expected a sequential informative initialization based on data rather than the uniform initialization gives further reduction in forecasting error. In this note, a modified Ensemble algorithm using sequential initial probability is developed. The sequential distribution is designed to have much weight on the recent data.

Keywords: Neural network, Initial distribution, Sequential initial, Modified ensemble algorithm

1) Professor, Department of Data information, Jeonju University, 1200, 3-ga Hyoja-dong, Wansan-gu, Jeonju, Jeonbuk, 560-759, Korea
E-mail: yhkim@jj.ac.kr

2) Full-time instructor, Department of Liberal Arts, Jeonju University, 1200, 3-ga Hyoja-dong, Wansan-gu, Jeonju, Jeonbuk, 560-759, Korea
E-mail: muggeby@jj.ac.kr