주가예측을 위한 비선형 회귀모형의 구성

Building Nonlinear Regression Models for Stock Price Prediction

저자 (Authors)
김동규, 길이만
Dongkyu Kim, Rhee M. Kil

출처 (Source)
한국정보과학회 학술발표논문집 , 2013.6, 1559-1561(3 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02217541

APA Style
김동규, 길이만 (2013). 주가예측을 위한 비선형 회귀모형의 구성. 한국정보과학회 학술발표논문집, 1559-1561

주가예측을 위한 비선형 회귀모형의 구성

김동규○ 길이만
성균관대학교 정보통신대학
goldencrosser@gmail.com, rmkil@skku.edu

Building Nonlinear Regression Models for Stock Price Prediction

Dongkyu Kim○ Rhee M. Kil
Sungkyunkwan University College of Information & Communication Engineering

요 약

주가 시계열은 일반적으로 비선형성 및 변동성의 동역학에 의하여 발생된다고 알려져 있다. 이러한 주가 시계열을 예측하기 위하여 기존의 선형회귀모형은 적절하지 않으며 주가 시계열에 따라 적절한 비선형 회귀모형의 구조 결정 및 회귀모형의 학습변수를 비모수적으로 추정하여야한다. 본 논문에서는 KOSPI 200 주가 시계열의 동역학적 구조 파악을 위하여 시계열의 위상공간(phase space) 분석을 실행하였고, 이를 바탕으로 인공신경망의 한 종류로서 비선형 비모수추정이 가능한 RBFN(Radial Basis Function Network)의 구조를 결정하였다. 특히, RBFN의 최적화를 위하여 본 논문에서는 주가 시계열의 잡음분산을 추정하였고, 이를 바탕으로 RBFN의 최적화, 즉 여기에서는 커널의 개수를 최적으로 결정하는 방법론을 새롭게 제안하였다. 그 결과 제안된 방법으로 KOSPI 200 주가 시계열예측을 매우 높은 정확도로 예측 가능하다는 것을 보였다.

1. 서 론

오늘날 금융 시장의 불확실성으로 주가예측의 필요성이 매우 높아지게 되었다. 주가와 같은 시계열의 예측방법으로 기존의 통계학적 시계열 예측 방법론으로 주로 사용된 방법은 ARMA(Autoregressive Moving Average) 모형으로 시계열 발생을 선형적인 동역학으로 설명하고자 하였다. 그러나 주가의 시계열은 수많은 참가자의 위험에 대해 다른 대처 방안, 투자 기간의 차이, 새로운 정보에 대한 다른 대응이나 동기 유발에 따라 매우 복잡한 양상을 보인다. 이와 같은 복잡한 경우에 이 모든 상호작용들을 단순히 선형 모형으로 설명할 수는 없을 것이다. 비선형 모형은 겉보기에 마구잡이로 보이는 경제 시계열의 움직임을 외부 잡음에 의해서가 아니라 자생적인 비선형 과정에 의해 일어난다고 설명할 수 있다[1]. 비선형 결정론적 관점인 카오스모형은 미래 상태는 현재의 상태와 그 동역학 법칙에 의해 완전히 결정된다는 것으로 보는 것을 말한다. 즉, 카오스는 무작위하게 움직이는 것처럼 보이는 경제시계열이 사실은 숨은 질서가 존재하는 비선형 결정론적인 과정을 따른다고 보는 것이다. 이와 관련하여 주가 시계열의 실증적 분석을 통해 한국증권시장은 비선형 결정론적인 시스템으로 카오스적 특성이 존재하는 것으로 보이는 연구결과를 도출하였다[2, 3].

본 논문에서는 주가 시계열을 비선형적 동역학으로 설명할 수 있다는 관점 하에 인공신경망의 한 종류로서 비선형 비모수 추정이 가능한 RBFN(Radial Basis Function Network)모형을 선택하였다. 여기서 주어진 문제는 주어진 주가 시계열에 따라 먼저 예측 가능한 동역학의 구조를 설정하고, 주어진 구조에 따라 최적의 비선형 회귀모형을 구축하는 것이다. 이를 위하여 본 논문에서는 주가 시계열의 위상공간(phase space) 분석을 통하여 가능한 동역학 구조를 설정하였고, 주어진 구조에 따라 시계열 예측과 관련된 RBFN의 구조를 설정하고 주어진 시계열을 학습하였다. 이와 관련하여 중요한 문제는 시계열 예측의 정확도를 높이기 위하여 RBFN 예측모형을 최적화 하여야 하는데, 이를 위하여 시계열에 내재된 잡음을 추정하고, 이를 바탕으로 RBFN 예측 모형의 최적화를 시도하였다. 본 방법은 시계열 예측 모형의 새로운 최적화 방법으로 다른 시계열 예측 모형에도 적용이 가능하리라고 판단된다. 제안된 방법은 4년간의 KOSPI 200 주가 시계열(1003개, 2008.06.13. ~ 2012.06.14.)에 적용하였고, 그 결과 매우 높은 정확도로 예측이 가능하다는 것을 보였다.

2. 주가 시계열의 위상공간 분석

시계열은 일련의 시간상의 값 $x(t), t=0,1,2,\cdots$ 으로 표시할 수 있다. 이러한 시계열의 위상공간은 지연시간 τ 및 내재차원 E로 표시된 벡터로 표시가 가능하다. 즉, $\{x(t), x(t-\tau), \cdots, x(t-(E-1)\tau)\}$와 같다. 시계열의 예측은 이러한 위상공간상에서 시계열 자료들이 사상(mapping) 가능한 함수를 구성할 수 있는가의 문제로 귀착된다. 그리고 사상 가능하다면 될 수 있는 데로 예측모형의 학습을 위하여 좀 더 유연성(smoothness)이 높은 함수가 적절하다. 이러한 관점으로 본 논문에서는 예측모형의 구조를 설정하였다.

먼저 본 연구의 목표인 주가예측모형의 설계를 위하여 다음과 같은 예측모형과 관련된 함수를 정의 한다 :

$$x(t+P) = f(x(t), x(t-\tau), \cdots, x(t-(E-1)\tau)) \quad (1)$$

여기서 P는 예측하고자 하는 예측시간(prediction time)이고, f는 예측함수를 의미한다.

시계열의 특성에 맞는 위상공간에 재구성하기 위하여 먼저 적절한 지연시간과 내재차원을 결정하여야 한다. 특히, 지연시간과 내재차원을 결정하는 것은 예측모형의 성능에 지대한 영향을 주는 중요한 요소라고 할 수 있다. 본 연구에서는 사상의 유연성에 기반을 둔 방법을 사용하여 지연시간과 내재차원을 결정한다. 여기서 사상의 유연성 측도 $S(\tau, E)$는 다음과 같이 정의 한다[4]:

$$S(\tau, E) = 1 - \frac{1}{N-E+1} \sum_{n=E}^{N} \left| \Delta x^{r}_{\tau, E}(n) \right| \quad (2)$$

위의 식에서 N은 자료의 개수 그리고 $\Delta x^{r}_{\tau, E}(n)$은 (1)에서 정의한 함수 f에서 위상공간상의 한 벡터 $\mathbf{x}_{\tau, E}(n)$와 $\mathbf{x}_{\tau, E}(n)$와 r번째로 가까운 벡터 $\mathbf{x}^{r}_{\tau, E}(n)$사이의 기울기를 의미한다.

계산한 사상의 유연성 측도를 3차원 궤도와 궤도의 등고선을 통해 보았을 때 기복이 심한 부분은 사상에 필요

한 차원보다 내재차원이 낮을 경우 나타나며, 반대로 사상에 필요한 차원보다 내재차원이 높을 경우에는 기복없이 평평한 모습을 보인다. 따라서 기복이 심한 부분에서 평평한 모습을 보이는 부분으로 변하는 경계, 즉 유연성의 측도가 작은 값에서 급격하게 큰 값으로 올라가는 부분에서 지연시간과 내재차원을 선택한다.

분석대상인 1003개의 KOSPI 200 시계열의 사상의 유연성 측도를 계산한 결과의 3차원 궤도는 아래의 그림 1, 궤도의 등고선은 그림 2로 표시된다. 기복이 심한 부분에서 평평한 부분으로 변하는 경계주변에서 일차적으로 내재차원을 정하고, 같은 내재차원에서 사상의 유연성 측도가 작은 값에서 지연시간을 선택하는 방법을 택하였다. 이 기준에 따라 적절한 지점을 그림 2에 점 $(\tau=4, E=3)$으로 찍어 표시하였다. 그 결과, 다음날 예측(one-step prediction)의 경우 예측모형의 구조는 다음과 같이 결정 된다 :

$$x(t+1) = f(x(t), x(t-4), x(t-8)) \quad (3)$$

따라서 본 분석에서 KOSPI 200의 다음날 예측의 경우 8일전 3개의 자료가 예측에 중요한 정보임을 나타낸다.

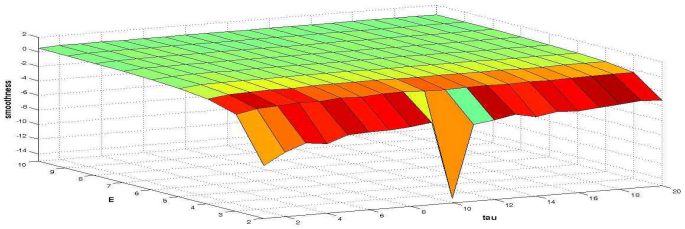

그림 1 : 사상 평활도의 3차원 궤도

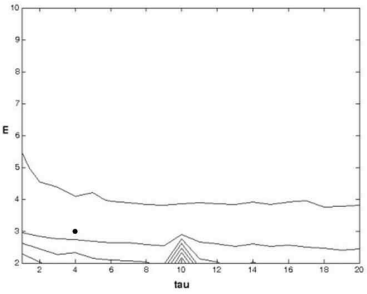

그림 2 : 3차원 궤도의 등고선

3. 비선형 주가예측 모형의 구성

주가 시계열예측을 위하여 시계열 자료로부터 예측모형 (3)을 학습하여야 하는데, 본 연구에서는 다음과 같이 비선형 비모수 추정이 가능한 RBFN으로 예측함수 f의 추정함수 $\hat{f}$을 정의하였다 :

$$\hat{f}(x) = \sum_{i=1}^{m} w_i \psi_i(x), \quad \psi_i(x) = e^{-(x-\mu_i)^2/2\sigma_i^2} \quad (4)$$

위의 식에서 m, μ_i, σ_i, w_i 는 각각 커널의 개수, 중심, 너비, 그리고 커널과 출력단과의 가중치를 의미한다.

RBFN 학습에 있어서 핵심적인 문제는 주어진 학습 자료를 이용하여 커널의 개수(m), 중심(μ_i), 커널의 너비(σ_i), 커널과 출력 층 사이의 가중치(w_i)를 예측의 정확도를 높일 수 있도록 최적화하는 것이다. RBFN의 변수추정 방법 [5]는 커널과 관련된 변수들을 MSE(Mean Square Error)를 줄이는 방법으로 매우 효과적인 추정을 제시하지만 최적의 커널 개수(m)는 제시하지 못한다. 여기서 최적의 커널 개수를 결정하는 것이 매우 중요한데, 그 이유는 커널 개수가 너무 많을 경우 과 적합(overfitting)이 일어나고 그 반대로 커널 개수가 너무 작은 경우 저 적합(underfitting)이 일어나, 학습되어지지 않은 패턴에 대해서는 성능이 떨어지는 일반화능력(generalization capacity)이 저하되는 현상이 일어난다. 그런데 일반적으로 시계열 $x(t)$는 다음과 같은 확률변수(random variable)로 설명할 수 있다 :

$$x(t) = f(t) + \epsilon \quad (5)$$

여기서 f는 시계열 x의 발생과 관련된 함수이고, ϵ은 평균이 0이고 분산이 σ^2인 잡음의 확률변수를 의미한다. 이러한 시계열 x에 대하여 추정함수 $\hat{f}$를 정의하면, 추정값 $\hat{x}$는 $\hat{x}(t) = \hat{f}(t)$로 주어진다. 그러면 시계열 x와 추정값 $\hat{x}$ 사이의 MSE는 다음과 같다 :

$$E[(x-\hat{x})^2] = E[(f-\hat{f})^2] + Var(\epsilon) \quad (6)$$

따라서 일반적인 오차(generalization error) $E[(x-\hat{x})^2]$는 회귀모형 오차(regression error) $E[(f-\hat{f})^2]$와 잡음의 분산(variance of noise)의 합으로 주어진다. 이러한 관점에서 본 연구에서는 잡음의 분산을 추정하고, 이를 바탕으로 커널의 개수를 추정하는 방법을 고려하였다. 여기서 커널 개수 결정의 기본적인 방향은 커널 개수를 증가시키면서 학습오차가 잡음의 분산보다 작을 경우 식 (6)에 근거하여 과도한 학습으로 잡음이 학습 되었다고 해석할 수 있음으로, 학습오차와 잡음의 분산을 비교하여 적절한 수준에서 학습을 종료하는 것이 예측의 정확도 면에서 유리한 방법론이라 할 수 있다.

시계열 잡음의 분산 추정을 위하여 본 논문에서는 다음과 같이 Rice가 제안한 방법 [6]을 고려하였다 :

$$\widehat{\sigma^2} = \frac{1}{2(n-1)} \sum_{i=2}^{n} (x_i - x_{i-1})^2 \quad (7)$$

위의 식에서 n은 시계열의 개수 그리고 x_i, x_{i-1}은 각각 i번째와 $i-1$번째 시계열을 의미한다. 여기서 잡음 ϵ이 정규분포를 따른다고 하면 이의 분산의 추정은 다음과 같이 카이스퀘어(chi-square) 분포를 따르게 된다 :

$$\frac{(n-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-1} \quad (8)$$

그러면 신뢰수준 $100(1-\alpha)\%$ 에서 실제 분산 σ^2의 범위는 다음과 같다 :

$$\frac{(n-1)\widehat{\sigma^2}}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)\widehat{\sigma^2}}{\chi^2_{1-\frac{\alpha}{2},n-1}} \quad (9)$$

주가 시계열의 예측을 위하여 앞서 제안된 방법론을 기초로 비선형 주가 예측모형의 구성 및 예측 과정은 다음의 순서와 같다 :

1. 먼저 주어진 주가 시계열 및 예측시간으로 부터 사상의 유연성 측도 (2)를 구하고, 이를 바탕으로 사상의 유연성 측도가 높으면서(일반적으로 -1.0 이상) 내재차원이 작은 값에서 내재차원 E와 지연시간 τ를 결정한다.
2. 순서 1에서 구한 결과에 따라 예측모형의 구조 (1)을 결정한다.
3. 주가 시계열에 내재된 잡음의 분산을 식 (7)의 방법에 따라 추정한다.
4. 예측모형의 구조 (1)에 따라 RBFN 추정망 $\hat{f}$를 결정하고, 커널의 개수를 증가시키면서 학습오차 $R_{emp}(\hat{f})$와 추정된 잡음의 분산 값 $\widehat{\sigma^2}$을 다음과 같이 비교 한다 :

$$R_{emp}(\hat{f}) \leq \frac{(n-1)\hat{\sigma^2}}{\chi^2_{1-\alpha/2,n-1}} \quad (10)$$

여기서 위의 식을 만족하면 과 적합의 가능성이 높기 때문에 학습을 중단한다. 아니면 커널의 개수를 증가시키면서 계속적으로 학습을 진행한다.

5. 순서 4에서 결정된 비선형 회귀모형으로 주가예측을 실시한다.

이러한 방법론의 실제 예로 KOSPI 200 주가 시계열(1003개, 2008.06.13.～2012.06.14.)에서 앞선 80% 자료를 학습 시계열로 사용하였고 나머지 20%를 검증 시계열로 비선형 회귀모형의 성능을 추정하였다. 여기서 추정된 학습패턴의 잡음분산은 $\hat{\sigma^2} = 0.000127$로 주어졌다. 그러므로 잡음분산의 신뢰구간은 식 (9)에 따라 신뢰도 99%에서 $0.000115 < \hat{\sigma^2} < 0.000159$로 구할 수 있다. 이를 기준으로 하여 MSE로 주어진 학습오차, 검증오차, 그리고 잡음분산의 신뢰구간을 비교하면 그림 3과 같다.

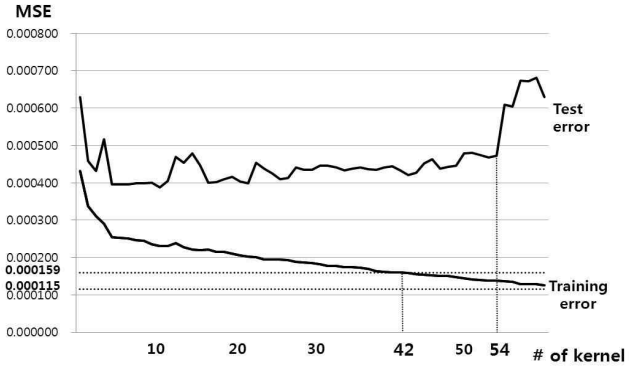

그림 3 : 커널의 개수에 따른 학습오차, 검증오차

그림 3에서 커널의 개수가 54개 이상일 경우 검증패턴에 대한 과 적합 현상이 나타났으며, 이 경우 학습오차가 잡음 분산의 신뢰구간 안에 있는 것을 알 수 있다. 따라서 여기서 주어진 시계열의 학습에서 학습오차가 99% 신뢰구간 밖에 있는 커널의 개수 42가 과 적합이 일어나지 않는 보다 적절한 선택임을 알 수 있다.

4. 주가예측 실험 및 결론

제안된 방법의 실험을 위하여 KOSPI 200 주가 시계열(1003개, 2008.06.13.～2012.06.14.)을 사용하였고 시계열의 정규화를 위하여 시계열의 값을 200으로 나누어 전체 시계열의 범위가 대략 0.5에서 1.5사이의 값을 갖도록 하였다. 그리고 전체학습 자료 중에서 앞선 70%, 80%, 90%를 학습자료로 사용하였고 나머지를 검증 자료로 사용하였다. 먼저 학습 자료가 의미하는 시계열 발생 동역학을 파악하기 위하여 사상의 유연성 측도 (2)를 이용하여 동역학 예측이 용이한 삽입차원 E와 지연시간 τ를 구하여 식 (3)과 같은 예측모형을 구하였다. 그 후 시계열에 내재된 잡음의 분산을 식 (7)을 이용하여 구하고 99% 신뢰구간을 식 (9)와 같이 구하였다. 그리고 RBFN 예측모형을 [5]의 방법으로 커널의 개수를 증가시키면서 학습(incremental learning)하였고 학습오차가 99% 신뢰구간에 인접한 경우 학습을 멈추었다. 그리고 학습된 예측모형을 검증 자료를 이용하여 성능평가를 실시하였다. 여기서 주가 시계열예측의 성능평가를 위하여 RMSE(Root Mean Squre Error)와 회귀모형 분석을 위한 결정계수(Coefficient of Determination) R^2을 사용하였다. 여기서 RMSE와 결정계수 R^2은 다음과 같다 :

$$\text{RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(x_i - \hat{x_i})^2} \quad (11)$$

$$R^2 = 1 - \left(\sum_{i=1}^{l}(x_i - \hat{x_i})^2 / \sum_{i=1}^{l}(x_i - \bar{x})^2\right) \quad (12)$$

여기서 l은 검증 시계열의 자료 개수 그리고 $\bar{x}$는 시계열의 평균값을 나타낸다.

제안모형과 비교예측을 위한 SVM(Support Vector Machine)모형의 학습비율 70%, 80%, 90%에 대한 주가 시계열예측의 성능평가 실험결과는 아래의 표1과 같다.

학습 비율	커널개수		RMSE		R^2	
	제안모형	SVM	제안모형	SVM	제안모형	SVM
70%	37	3	0.035879	0.058423	87%	60%
80%	42	6	0.020517	0.039960	90%	64%
90%	63	6	0.016172	0.036820	88%	41%

표1 주가 예측 실험 결과

제안된 예측모형의 실험결과 학습비율이 높아지면 필요한 커널의 개수가 증가하는 것을 알 수 있었고, 이에 따라 RMSE도 감소함을 알 수 있었다. 그런데 학습된 회귀모형이 얼마나 주어진 시계열의 변동성을 잘 쫓아가는가의 측도로서 결정계수 R^2는 87%~90% 정도로 표시되었다. 일반적으로 회귀모형의 결정계수가 90%이상이면 매우 우수한 회귀모형이라 할 수 있는데, 제안된 모형은 이에 매우 근접함을 알 수 있었다.

결론적으로 본 논문에서 제안된 주가 시계열예측을 위한 비선형 회귀모형의 예측성능은 우수하다고 할 수 있고, 이러한 결과는 시계열의 동역학 파악을 위한 사상의 유연성을 이용한 시계열 위상공간의 분석, 그리고 시계열에 내재된 잡음의 분산을 이용한 RBFN 예측모형의 최적화에 기인한다고 할 수 있다. 그리고 제안된 방법은 일반적인 시계열의 예측에도 적용 가능함으로 앞으로 다양한 시계열의 예측에 제안된 방법이 매우 효과적으로 사용되리라고 기대된다.

참 고 문 헌

[1] 김상락 (2000) “주가 예측 : 허구인가? 사실인가?”, 물리학과첨단기술, 12-17

[2] 장경천, 김현석 (2004) “주식 수익률의 비선형 결정론적 특성에 관한 연구”, 재무관리연구 제21권 제1호, 149-181

[3] 박형중 (2006) “한국 주식시장의 비선형적 동태성과 카오스 현상에 대한 실증분석”, 성균관대학교

[4] Kil, R., Park, S., & Kim, S. (1999) “Time series analysis based on the smoothness measure of mapping in the phase space of attractors”, International Joint Conference on Neural Networks, Vol. 4, 2584-2589

[5] Kil, R. (1993) "Function Approximation Based on a Network with Kernel Functions of Bounds and Locality : an Approach of Non-Parametric Estimation", ETRI Journal, vol. 15, no. 2, 35-51

[6] Rice, J. (1984) “Bandwidth choice for nonparametric regression”, Annals of Statistics, 1215-1230