Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

데이터마이닝 기법을 이용한 주가자료 분석

손인석[1] · 황창하[2] · 조길호[3] · 김태윤[4]

<요약>

본 연구의 주된 목표는 1997년 주가자료를 데이터마이닝 기법인 로지스틱모형, 의사결정트리, 신경망, SVM(support vector machine), 뉴로퍼지모형을 사용하여 분석한 후 우리나라 경제상황을 진단하고 예측하는데 가장 적합한 모형을 찾고 그 모형을 해석하는데 있다. 1997년 주가자료를 훈련자료로 간주하여 그 당시 경제 상황에 따라 적절한 구간으로 나누고 훈련시킨 결과 중요한 변수로는 주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비로 나타났으며 적절한 기법으로는 의사결정트리, 신경망, SVM임을 알 수 있다. 1997년 이외의 주가자료를 데이터마이닝 기법(신경망, 의사결정트리, SVM)에 적용한 결과, 우리나라 경제상황을 고려해 볼 때 신경망이 가장 정확도가 좋은 기법으로 보여진다.

주제어 : 경제상황, 경제위기, 주가자료, 데이터마이닝.

1. 서론

소규모 개방경제인 우리나라의 경제상황을 가장 잘 나타내는 변수로서 여러 가지가 있을 수 있으나 그 중 대표적인 것이 환율, 금리, 주가 등이다. 경제위기 이전까지의 이들 세 변수에 대한 분석은 김명기 문소상(1998) 등에 의해 이루어졌는데 그들의 분석에 의하면 주가, 환율, 금리 중 주가의 변동성(volatility)이 가장 높은 것으로 관찰되었으며 이는 경제여건의 변화에 주가가 가장 민감하게 반응할 수 있다는 사실을 암시하고 있다. 김태윤 황창하(2001)는 신경망을 사용하여 1997년 주가자료를 분석하고 경제상황의 지표를 개발한바 있다.

본 논문에서는 훈련자료인 1997년 주가자료와 검정자료인 1997년을 제외한 1994년부터 현재까지의 자료에 데이터마이닝 기법인 로지스틱모형, 의사결정트리, 신경망, SVM, 뉴로퍼지모형을 적용하여 분석한 후 우리나라 경제상황을 진단하고 예측하는데 가장 적합한 모형을 찾고 그 모형을 해석하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 데이터마이닝 모형에 대해 간단히 소개를 하고 3절에서는 1997년 경제 위기 연도의 주가자료를 데이터마이닝 기법에 적용한 결과를 집중분석하며 4절에서는 1997년 이외의 주가자료에 적용하여 그 정확도 및 타당성을 평가한다. 5절에서는 결론을 제시한다.

2. 데이터마이닝 모형

1) 대구광역시 북구 산격동 370번지 경북대학교 통계학과 석사과정
2) 경북 경산시 하양읍 금락1리 330번지 대구가톨릭대학교 정보통계학과 교수
3) 대구광역시 북구 산격동 360번지 경북대학교 통계학과 교수
4) 대구광역시 달서구 신당동 1000번지 계명대학교 통계학과 교수

2.1 신경망

신경망 기법은 1943년 McCulloch과 Pitts에 의해 이루어졌으며, 인간의 뇌의 학습기능을 모방하여 만들어진 자료분석 도구이다. 즉 신경단위(neuron)라 불리는 여러 개의 간단한 정보처리 단위들을 연결하여 이루어지며 학습과정을 통해 정보를 축적하여 활용하는 특성을 갖고 있다. 본 논문에서 사용하는 신경망은 역전파 알고리즘 기반의 분류를 위한 신경망이다.

역전파 알고리즘은 다층의 구조를 갖는 신경망에 사용되는 학습알고리즘으로서 최급하강법(steepest decent method)을 기본으로 하는 함수근사기법이다. 기본원리를 살펴보면, 입력 층의 각 노드에 입력패턴이 제시되면 중간층으로 변환된 신호를 전달하게 되고 최후에 출력 층에서 신호를 출력하게 된다. 이때 출력된 값과 목표 값을 비교하여 그 차이를 줄여나가는 방향으로 가중치를 조절하며 그 과정은 상위 층에서 하위 층으로 역전파하는 순서를 밟는다. 결국 이 알고리즘은 최급하강법을 이용하여 오차제곱의 합이 최소가 되도록 신경망의 모수 즉, 가중치를 반복적으로 조정하는 알고리즘이라 할 수 있다. 최급하강법은 반드시 가장 낮은 골짜기를 목표로 하는 방법이 아니라 지금 있는 점에서 보아 가장 급경사면을 따라 내려가는 것이므로 극소점(local minima point)에 머무를 가능성이 높다.

함수근사과정에서 신경망은 스스로가 주어진 자료를 갖고 적응적(adaptive)으로 문제를 풀어나가며, 따라서 근사과정에서 신경망에 의해 만들어지는 기저함수(basis function)가 다양하며 결국 다른 방법에 비해 적은 수의 기저함수로 미지의 비선형 다변량 함수를 잘 근사할 수 있다는 장점이 있다. 그리고 가장 중요한 점은 신경망이 사후확률을 추정하여 최대의 사후확률을 갖는 그룹으로 패턴을 분류하는 베이즈 분류규칙이라는 점이다. 한편 분류를 위한 신경망은 지역극소(local minimum)에 빠질 수 있고 훈련시간이 길며 이상치에 민감하다는 단점이 있다.

2.2 의사결정트리

의사결정트리(decision tree)는 의사결정규칙(decision rule)을 트리구조로 도표화하여 분류와 예측(prediction)을 수행하는 분석방법이며, 탐색(exploration)과 모형화(modeling)라는 두 가지 특성을 모두 가지고 있다고 할 수 있다. 즉, 의사결정트리는 판별분석(discrimination analysis) 또는 회귀분석(regression analysis) 등과 같은 모수적(parametric) 모형을 분석하기 위해서 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 교호효과를 찾아내기 위해서 사용될 수도 있고, 그 자체가 분류 또는 예측 모형으로 사용될 수도 있다. 의사결정트리는 하나의 트리구조를 이루고 있으며, 마디(node)라고 불리는 구성요소들로 이루어져 있다. 의사결정트리 분석을 위해서 CHAID(Kass, 1910), CART(Breiman et al., 1914), C4.5 (Quinlan, 1993) 등과 같은 다양한 알고리즘이 제안되어 있으며, 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 있다. 의사결정트리 분석의 장점은 이해하기 쉬운 규칙을 형성하고 규칙에 근거한 영역을 생성시켜주며 연속변수와 범주형 변수에 모두 사용가능하고 가장 좋은 변수를 명확히 알아내고, 이상치에 덜 민감하다는 것이다. 단점은 계층이 너무 많은 경우엔 오류가 나기 쉽고 비용이 많이 들며 비 사각영역에선 문제가 있다.

2.3 로지스틱모형

로지스틱모형은 반응변수가 이항형(binary) 또는 다항형일 때, 반응변수와 입력변수들간의 관계를 조사하기 위해 사용하며 사후확률 $p(k \mid \boldsymbol{x})$를 훈련표본으로부터 직접 구하는 분류기법이다. 그룹이 두

개인 경우의 로지스틱 분류기법을 생각해 보자. 그룹 2에 대한 로그 승산비(log odds ratio)는

$$\text{logit}\, p(2 \mid \boldsymbol{x}) = \log \frac{p(2 \mid \boldsymbol{x})}{p(1 \mid \boldsymbol{x})}$$

가 되며 로짓함수는 $\text{logit}\, p = \log p/(1-p)$이고, 역함수는 로지스틱함수 $\ell(x) = e^x/(1+e^x) = 1/(1+e^{-x})$이다. 이제 로그 승산비는 다음과 같이 입력벡터의 선형함수라고 가정할 수 있다. 즉

$$\text{logit}\, p(2 \mid \boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x},$$

$$p(2 \mid \boldsymbol{x}) = \ell(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}), \quad p(1 \mid \boldsymbol{x}) = 1 - p(2 \mid \boldsymbol{x})$$

이 되며 이것은 은닉층이 없고 한개의 출력노드를 갖는 신경망이 된다. 이러한 로지스틱 분류기법은 입력값들의 선형함수와 로지스틱 출력 두 단계로 나누어 생각할 수 있다.

따라서 패턴인식에 사용되는 로지스틱모형은 사후확률 $p(k \mid \boldsymbol{x})$를 추정하여 최대의 사후확률을 갖는 그룹으로 패턴을 분류하는 최적의 분류기법인 베이즈 규칙으로서 신경망의 특수한 경우이며 주어진 자료에 잘 부합하는 선형분류기법이고 해석이 가능하다는 장점을 가지고 있다. 그리고 Fisher의 선형판별분석법과 비교할 때 모집단분포가 정규분포가 아니거나 공분산행렬이 다를 때에도 성능이 비교적 좋은 선형분류기법이다. 그러나 단점은 교호작용의 결여이다.

2.4 SVM

SVM(support vector machines)은 Vapnik(1995)에 의해 고안되었다. 선형으로 분리 가능한 두 그룹을 구분 지으며 마진(margin)을 최대로 하는 초평면 $\boldsymbol{w}'\boldsymbol{x} + b = 0$을 찾는 분류기법이다. 통계적 학습이론 관점에서 보면 SVM은 모형복잡도의 측정수단으로 VC 차원(Vapnik Chervonenkis dimension) h를 도입하여 아래의 기대위험(expected risk) $R[f]$의 상계(upper bound)

$$R[f] \le R_{\text{emp}}[f] + \sqrt{\frac{h(\ln\frac{2l}{h}+1) - \ln\frac{\delta}{4}}{l}},$$

를 최소화하는 초평면을 찾는 방법이다. 여기서 l은 관측치의 개수이며 $R_{\text{emp}}[f]$는 경험적 위험(emprical risk)이다. 즉, 최적의 모형을 찾기 위해서 $R_{\text{emp}}[f]$ 만을 최소로 하는 모형을 찾는 것만으로는 부족하고 모형복잡도까지 같이 고려해야 한다는 것이며 결과적으로 예측을 잘하는 모형을 찾겠다는 것이다.

SVM은 VC 차원의 직접적인 제어가 가능하기 때문에 VC 차원의 상한 값(upper bound)과 학습에러를 동시에 최소로 줄일 수 있다는 장점이 있다. 그리고 SVM은 SV(support vector)만으로 이루어진 모형을 구성하고 특징공간(feature space)의 차원에 독립이며 통계적 이론에 충실하다. 간단히 그룹이 두 개일 경우를 예를 들어 생각하면, SVM의 가장 큰 장점은 $\text{sign}(\text{사후확률} - \frac{1}{2})$을 추정하는 일종의 베이즈 규칙이라는 점이다. 단점은 QP(Quadratic Programming) 문제가 있다는 것이다.

2.5 뉴로퍼지모형 NEFCLASS

신경망(neural network)과 퍼지시스템(fuzzy system)을 결합한 방법들이 공학분야에서 많이

활용되고 있다. 그러나 대부분의 방법들은 매우 다른 네트워크 구조(architecture)와 활성함수(activation function) 그리고 학습알고리즘(learning algorithms)을 사용하기 때문에 비교가 쉽지는 않다. Nauck은 다층퍼지신경망(multilayer fuzzy neural network)의 일반적 모형을 위해 퍼지퍼셉트론(fuzzy perceptron)을 소개하고 뉴로퍼지모형 NEFCLASS를 개발하였다. 퍼지퍼셉트론은 신경망과 퍼지시스템을 결합한 여러 방법들의 비교를 용이하게 하는 뉴로퍼지모형 구조의 공통 기반으로 사용될 수 있다. 뉴로퍼지모형 NEFCLASS의 목적은 서로 다른 그룹으로 분리될 수 있는 자료들의 집합으로부터 퍼지규칙을 유도하는 것이다. 이때 애매성(fuzziness)은 주로 입력패턴을 정확한 그룹으로 분류하는 것을 어렵게 만드는 입력변수들의 부정확하고 불완전한 측정에 기인한다. 자료를 설명하는 퍼지규칙은 다음과 같은 형태이다.

만약 x_1이 μ_1, x_2가 μ_2, … , x_n이 μ_n이면 패턴 x는 그룹 i에 속한다.

여기서 $x = (x_1, x_2, \cdots, x_n)'$이고 $\mu_1, \cdots, \mu_n$은 퍼지집합이다. 뉴로퍼지모형 NEFCLASS는 이런 퍼지규칙을 유도하고 소속함수(membership function)의 형태를 학습한다. 사전지식에 의해 초기화 될 수 있는 뉴로퍼지 시스템 모형은 퍼지의 **if-then** 규칙을 사용하여 학습을 한 후 해석이 가능하다는 강한 장점이 있다. 즉 일반적 신경망과 같은 블랙박스 방법은 아니고 해석 가능한 퍼지분류 방법이다. 뉴로퍼지 모형은 훈련자료에 대해 한번 학습한 후 퍼지규칙을 생성할 수 있다. 퍼지규칙을 생성한 후 뉴로퍼지 모형은 감독 학습알고리즘을 사용해 소속함수의 모수들을 적응 적으로 추정하여 최종적으로 퍼지규칙을 완성한다. 유도된 규칙은 다른 뉴로퍼지 방법처럼 가중평균을 사용하여 계산되지 않는다. 따라서 의미론적 문제(semantical problem)를 피하고 학습결과를 간단하게 만든다.

3. 1997년도 주가지수를 이용한 경제상황 예측 모형 개발

훈련자료를 경제위기가 발생된 연도인 1997년 주가자료를 사용했으며 세 구간으로 나누었다. 신호가 발생한 9월 30일을 전후로 하여 9월 19일부터 10월 21일까지를 위기직전의 불안정 구간(구간 2), 1월 3일부터 9월 18일까지를 정상구간(구간 1), 10월 22일부터 12월 27일까지를 위기구간(구간 3)으로 하였다. 구간 1은 정상 구간으로서 더 이상의 설명이 필요 없다고 생각되며, 구간 2는 위기상황을 앞두고 변동성의 급작스런 증가등 신호가 발생하고 경제가 혼란스러워 지기 시작하는 구간이며 구간 3은 경제위기가 이미 발생하여 시장이 그것을 인지하여 반응하고 있는 구간이다. 자료분석결과 세 구간이 나름대로 구조적인 차이점이 명백한 것으로 판단되었기 때문에 이들 세 구간을 경제상황 판단의 근거가 되는 그룹으로 간주하였다.

97년 자료를 훈련자료로 사용하여 데이터마이닝 기법에 훈련시키기 위해 입력변수는 5개(주가지수, 등락률, 등락률 10일 이동평균, 등락률 10일 이동분산, 10일 이동분산의 변동비)가 사용되었다. 표 3.1을 보면 SVM, 신경망, 의사결정트리는 세 구간을 정확하게 분류하고, 로지스틱모형과 뉴로퍼지모형은 정상구간, 위기구간을 어느 정도 정확하게 분류하나 불안정구간을 잘 분류하지 못함을 보여준다. 이것은 그룹별 자료의 수가 불균등하기 때문인 것으로 생각된다. 5개의 입력변수 중 주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비가 선택되었다. 그 변수에 대한 타당성을 고려하기 위해 자료를 살펴보면 등락률 10일 이동분산은 97년 9월경부터 증가함을 알 수 있는데 이 시점에서부터 다가오는 위기에 대해 시장이 사전적으로 어떤 반응(혹은 신호)을 보이고 있다고 여겨지고, 10일 이동분산의 변동비는 97년 9월 30일경 증가비가 4이상인 명백한 신호가 있었으며 그 이전에 변동비 2이상인 값들도 가끔 관찰되었던 것을 알 수 있다. 따라서 주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비가 중요한 변수라고 여겨진다.

[표 3.1] 훈련자료의 오분류 및 변수 선택

	변수 선택	오분류(%)
뉴로퍼지모형	주가지수, 10일 이동분산의 변동비	10.27
SVM	×	0
신경망	×	0.3224
의사결정트리	주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비	1.370
로지스틱모형	주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비	7.8767

4. 개발된 경제상황 예측 모형의 적용

본 절에서는 1997년 이외의 자료에 적용한 후, 실제 적용시의 정확성 및 문제점을 분석하고자 한다. 이를 위해 1994년 1월 4일부터 2001년 9월 8일까지의 주가지수 자료를 사용하였다. 오분류가 심각한 뉴로퍼지모형 NEFCLASS와 로지스틱모형을 제외한 3가지 데이터마이닝 기법(의사결정트리, SVM, 신경망)을 1994년부터 최근까지의 주가지수에 적용한 결과를 보면 97년 이전구간은 정상구간 1로 구분해 내고 있다. 98년에서 99년 3월경까지 구간(T_3)은 기법에 따라 약간씩 차이를 보이고 있다. 의사결정트리와 신경망은 위기구간 3으로 분류하고 있으며 SVM은 위기구간 3, 안정구간 1, 위기구간 3 순으로 분류하고 있다. 99년 3월 중순에서 4월초까지 구간(T_4)은 잠시 불안정 구간 2를 거친 후 99년 4월경부터 2000년 9월초까지 구간(T_5)은 안정구간 1로 분류하고 있으며 2000년 9월 중순부터 현재까지 구간(T_6)도 기법에 따라 약간씩 차이를 보이고 있다. SVM은 분류가 아주 불안정하고 신경망과 의사결정트리는 2001년 2월 초순까지는 위기구간 3으로 분류하고 그 이후는 불안정구간 2로 분류하였다. 구간의 분류가 사후확률의 값을 가장 크게 갖게 되는 구간을 찾는 방법으로 진행되기 때문에 가끔 분류된 경제상황이 각 구간 사이에서 며칠 간격으로 빠르게 변화 이동하는 것을 볼 수 있는데, 이러한 현상은 데이터마이닝 기법으로 구분하기가 까다로운 시점들 때문인 것으로 보이며 전체적인 분류의 흐름으로 보았을 때에는 큰 의미가 없는 것으로 무시해도 좋을 것으로 판단된다.

이와 같은 분류 결과를 보면 실제 우리나라 경제가 겪어온 상황과 거의 일치하는 것으로 판단된다. 즉 IMF직후 T_3기간에서는 많은 어려움을 겪은 위기상황이었으나 강력한 경제구조개혁 및 조정 그리고 미국경제의 지속적인 호황 등을 통해 T_5기간 중 정상상황을 회복할 수 있었다. 여기서 흥미로운 것은 T_3 와 T_5 기간 사이에서 이동시 짧은 불안정기간 T_4를 겪었다는 점이다. 이는 불안정 구간 2의 주된 특징인 변동성의 증가가 미래의 경제상황 호전시에도 발생할 수도 있음을 보여주고 있다. 우리나라 경제는 2000년 9월경부터 현대사태 및 해외경제 여건의 악화로 인한 어려움을 겪고 있는데 이 기간(T_6)을 의사결정트리와 신경망은 위기 및 불안정 구간으로 구분하고 SVM은 분류가 매우 불안정한 것으로 경제상황을 예측하고 있다. 이는 위기상황이 제대로 극복되지 못할 경우 경제 불안이 증대되기 시작한다는 점을 보여주고 있으며 2001년 3월경부터 또 다른 경제위기에 관한 논의가 사회 전반에 걸쳐 공개적으로 활발히 전개된 것과도 무관하지 않은 듯하다. 특히 현재까지 5개월간의 장기간 불안정 상황이 지속되고 있는데 이는 경제위기 혹은 불황이 다가올 가능성을 말해주고 있는 듯하다.

5. 결론

본 연구의 주된 목표는 1997년 주가자료를 데이터마이닝 기법인 로지스틱모형, 의사결정트리, 신경망, SVM(support vector machine), 뉴로퍼지모형을 사용하여 분석한 후 우리나라 경제상황을 진단하고 예측하는데 가장 적합한 모형을 찾고 그 모형을 해석하는데 있다. 1997년 주가자료를 훈련자료로 간주하여 그 당시 경제 상황에 따라 적절한 구간으로 나누고 훈련시킨 결과 중요한 변수로는 주가지수, 등락률 10일 이동분산, 10일 이동분산의 변동비로 나타났으며 적절한 기법으로는 의사결정트리, 신경망, SVM임을 알 수 있다. 1997년 이외의 주가자료를 데이터마이닝 기법(신경망, 의사결정트리, SVM)에 적용하여 분류한 결과를 우리나라 경제상황에 고려해 볼 때 의사결정트리와 신경망이 상당한 정확도를 갖고 있는 것으로 판단된다. 의사결정트리는 주로 해석력을 강조하는 자료분석에 사용되고 신경망은 주로 예측력을 강조하는 자료분석에 사용된다. 주가자료에 대해서는 신경망이 의사결정트리보다 더 높은 정확도를 가지고 있는 것으로 보여진다.

본 연구 결과는 다른 변수들과의 연계에 의해 (예를 들면 환율, 금리, 외환보유고등) 더욱 정확도가 높고 의미 있다고 생각된다.

참고문헌

[1] Bishop, C. M. (1995), *Neural Networks for Pattern Recognition,* Clarendon Press, Oxford.

[2] Chen, D. S. and Jain, R. C. (1994), *A Robust Back Propagation Learning Algorithm for Function Approximation,* IEEE Transactions on Neural Networks, **5**, 467-479.

[3] Cheng, B. and Titterington, D. M. (1994), *Neural Network: A Review from a Statistical Perspective.* Statistical Science Vol. **9**, No. 1, 2-54.

[4] Cherkassky, V., Friedman, J. H. and Wechsler, H. (1994), *From Statistics to Neural Networks Theory and Pattern Recognition Applications*, Springer-Verlag.

[5] Cherkassky, V. and Mulier, F. (1998). *Learning from data*, John Wiley & Sons, Inc.

[6] Hornik, K., Stinchcombe, M. and White, H. (1989), *Multilayer Feedforward Networks are Universal Approximators*, Neural Networks, **2**, 359-366.

[7] Nauck, D. (2000). Data Analysis with Neuro-Fuzzy Methods. Habilitation Thesis, University of Magdeburg.

[8] Ripley, B. D. (1994), *Neural Networks and Related Methods for Classification,* J. R. Statist. Soc. B **56**, No. 3, 409-456.

[9] Vapnik, V (1998), *Statistical Learning Theory,* John Wiley & Sons, Inc.

[10] 김명기 문소상 "환률, 금리, 주가변동의 상호 연관성" 『경제분석』 4권 2호 1998 93-113

[11] 김태윤 황창하 (2001) "신경망을 이용한 경제상황지표 개발 " 한국은행 보고서

[12] 박원암 최공필 (1998) "한국외환위기의 원인과 예측 가능성" 『한국경제의 분석』 4권 2호 1-73

[13] 이종규(2000A) 『IMF/IBRD 경제개혁 프로그램과 앞으로의 경제정책방향』 금융경제총서 2000-1, 한국은행 특별연구실

[14] 이종규 (2000B) 『경제위기: 원인과 발생과정』 금융경제총서 2000-2, 한국은행 특별연구실

[15] 한국은행 (2000) 『한국의 금융경제 연표』 한국은행 조사국

[16] 황선영 김은주(2000) "TAR-GARCH" 모형을 이용한 주가자료분석" 『응용통계연구』, 13권 2호 437-444