대한산업공학회/한국경영과학회 2000 춘계공동학술대회 논문집. 2000년4월21-22일. 경남대학교
Session 0B5.5

신경망을 사용한 매도/매수 주식 종목 선정

임도형*, 이일병

연세대학교 인지과학

dhrim@csai.yonsei.ac.kr, yblee@csai.yonsei.ac.kr

주가는 시계열 데이터의 일종으로 많은 변수들이 주가의 변동에 영향을 미친다. 그러나 몇 개의 어떠한 변수가, 어떻게 영향을 미치는 지 정확히 알려져 있지 않다. 그렇기 때문에 주가를 예측하는 것은 쉽지 않으며 단지 등락을 예측하는 것 조차도 쉽지 않다. 본 논문에서는 주가를 신호와 잡음이 혼합된 것으로 가정하고 그 특성을 고려하여, 전 종목에 대한 등락을 예측하지 않고, 예측율이 높은 종목을 선정하는 것을 목표로 하였다. MLP를 BP로 학습시켰으면 입력으로는 28개의 주가분석 지표값이 사용되었다. 여러 예측 기간으로 실험하였으며, 예측기간이 60일일 때 77.1%의 예측율을 보였고 선정된 종목의 등락 예측율은 88%였다.

서론

주가의 특성

주가는 매수 세력과 매도 세력에 의해서 변동폭이 결정되는 시계열 데이타이다. 매수 세력과 매도 세력은 과거의 주가 뿐 아니라 환율이나 금리, 혹은 각 종목에 대한 특정 뉴스에 의하여 영향을 받는다. 심지어 각 투자자의 개인적인 관심이나 심리에 의해서 영향을 받는다. 이렇게 수많은 변수들이 존재하고 변수들이 무엇인지 파악하기 어렵다.

가능한 모든 변수 중에 우리가 모르거나 측정하기 불가능한 변수가 있다고 가정한다. 그러한 변수에 의한 예측 불가능한 변동을 잡음으로 볼 수 있다. 그리고 알려진 변수에 의한 예측 가능한 변동을 신호로 볼 수 있다. 이렇게 가정한다면 주가변동은 신호와 잡음이 혼합된 시계열 데이타이다. 특정 투자가의 매수매도로 인해 주가가 최대 변동폭으로도 변할 수 있는, 신호 대 잡음비가 무척 큰 시계열 데이타이다. 예측된 변동 크기보다 잡음이 클 경우 예측 결과는 틀릴 수 있다.

이러한 주가의 특성으로 인해 다른 시계열 데이타와 달리 예측이 어렵다. 하지만 잡음의 크기보다 예측된 변동의 크기가 크도록 예측을 조절한다면 보다 정확하게 예측할 수 있을 것이다. 주가의 잡음은 그 크기가 크지만 다른 잡음과 마찬가지로 그 평균은 0으로 볼 수 있다.

주가의 경우 하루의 최대 변동 폭은 15%이기(한국 증권시장의 경우) 때문에 잡음의 최대 크기를 15%로 볼 수 있다. 이러한 사항을 고려한다면 1일 후의 주가는 신호의 변동 폭이 잡음의 변동폭에 비해 크지않기 때문에 신호 자체 보다는 잡음에 의해 변동이 좌우될 수 있다. 그렇기 때문에 그 예측이 불가능하다고 할 수 있다. 하지만 짧은 기간이 아닌 충분한 기간 후의 주가는 잡음의 특성으로 인해 그 영향이 상쇄되고, 신호를 예측할 수 있다. 그러나 과거와 현재의 변수에 의한 영향은 시간에 따라 감쇄하기에 그 기간이 커질 경우 현재 데이터에 의한 예측의 정확도가 감소하게 된다. 결국 잡음의 영향과 현재 신호의 감쇄를 고려한 적당한 예측 간격을 선택하여야 한다.

종목 선정

과거와 현재의 변수에 의한 신호는 단기적인 기간

후에는 잡음에 의해 예측 불가능하고, 장기적인 기간 후에는 그 신호의 영향이 감쇄되어 역시 예측이 불가능해 진다. 위의 두 가지의 경우를 고려하여 극히 짧거나 길지 않은 기간을 선택한다면 예측이 가능할 것이다. 그러나 적당한 예측 간격의 경우에도 신호의 변동 폭이 작을 경우 잡음의 영향으로 인해 역시 예측이 어렵고 큰 경우에만 예측이 가능할 것이다. 그렇기 때문에 모든 종목에 대한 예측은 불가능하다고 보고, 본 논문에서는 기존의 방법에서와 같은 전체 종목에 대한 등락 예측을 하지 않고, 예측이 정확한 일부 종목을 선정하는데 중점을 두었다.

신경망의 보편적 근사능력과, 입력의 누락

신경망은 보편적인 근사능력을 갖는다[3]. 이는 임의의 함수를 근사화 할 수 있다는 것이다. 그러나 입력의 일부분이 누락될 경우, 누락된 부분으로 인한 영향은 근사화 되지 않을 것이고, 단지 입력에 의한 부분만이 근사화 되거나 근사화 되는 정확도가 감소할 것이다. 이 경우에 입력된 부분에 의한 출력을 신호라 하고, 누락된 부분에 의한 출력을 잡음이라 할 수 있다. 지금 근사화 할 함수가 1 또는 0의 출력값 만을 가지고, [0,1]외의 출력값은 불가능하다면, 입력 부분의 누락된 비중이 커질 수록 잡음에 의해 0과 1사이의 출력값이 발생할 것이다. 1과 0에 해당하는 집합의 출력은 각각의 [0.1] 사이에서 어떤 출력 분포를 갖을 것이다. 잡음이 증가할 수록 두개의 집합의 각각의 평균은 서로 가까워 질 것이며 0의 집합에 속하는 데이터의 출력이 1에 가까운 경우도 가능할 것이다. 그러나 어떤 거리 안에서 1에 가까운 출력값을 보이는 데이터가 1의 집합에 속하는 확률은 0의 집합에 속하는 확률이 클 것이다. 이 때 어떤 거리보다 작은 데이터의 집합 판별의 정확도는 상대적으로 높을 것이다.

신경망을 사용한 주가 등락 예측

과거와 현재의 주가 데이터를 사용한 특징값을 입력으로 하고 등락을 출력으로 한 함수는 위에서 언급한 입력부분이 누락된 함수로 가정할 수 있다. 과거 주가 데이터를 입력으로 하고 등락을 1과0 값의 출력으로 하여 학습된 신경망은, 어떠한 출력 분포를 갖을 것이고, 이 출력 분포는 잡음으로 인해 확산된 상승집합과 하락집합의 분포의 합으로 볼 수 있다. 1과 0에 가까운 종목은 각각 상승과 하락의 집합에 속할 확률이 높다. 이러한 방법으로 등락 예측율이 높은 종목을 선정한다.

실험

신경망 구조

주가 등락 함수를 근사화 하기 위해 MLP[4,6]를 사용하였으며 BP를 사용하여 학습하였다. MLP는 5개의 은닉노드가 있는 한 개의 은닉층을 갖고 있다. 과거와 현재의 주가를 분석한 29개의 값을 입력으로 하는 29개의 입력노드가 있고, 등락을 출력하는 한 개의 출력노드가 있다. 1의 경우 상승을 0의 경우 하락을 의미한다. 모멘텀을 사용했으며, 학습율, 모멘텀 학습율, 바이어스 학습율은 각각 0.1, 0.1, 0.1이었다.

사용 데이터

한국 증시에 상장되어 있는 900여 개의 종목 중, 99년 이전에 상장되었고 정상적인 거래가 진행되는 848개의 종목 중 임의의 144개 종목을 대상으로 하였다. 1998년 8월 14일부터 1999년 12월 6일까지의 종가를 사용하였다.

입력값

실험에 사용된 입력값으로 <표1>과 같은 28개의 지표[7]에 의해 구하여진 매매신호과 과거 10일 대비 당일의 종가의 변동 비율을 사용하였다. 매매신호가 매수일 경우 1, 매도일 경우 -1, 중립일 경우 0을 입력하였다. 과거 대비 당일의 변동 비율은 [0,1]로 정규화 하여 입력하였다.

1	Aroon Cross
2	Linear Regression
3	MACD(MA Convergence Divergence)
4	ADX
5	Bollinger Band Composite
6	Moving Average Composite
7	DM (Directional Movement)
8	Average True Range (ATR)
9	Bollinger Band
10	Moving Average(MA)
11	Standard Deviation
12	Envelope
13	Linear Regression Signal
14	Psychological Line
15	CCI (Commodity Channel Index)
16	Momentum Indicator
17	Range Indicator
18	RSI (Relative Strength Index)
19	Japanese Candlestick Signal

20	Stochastic
21	A/D (Accumulation/Distribution)
22	Ease of Movement
23	MFI(Money Flow Index)
24	Moving Average (Volume)
25	Negative Volume
26	Positive Volume Index
27	Volume Oscillator
28	Volume Rate-of-Change (VROC)

<표1> 신경망 입력에 사용된 28개 주가 지표

각 지표는 과거 최대 30일 간의 종가를 사용하여 매매 신호를 발생한다. 이러한 지표값은 과거와 현재의 주가 데이터를 분석한 값이며 추후의 주가의 움직임을 예측하기 위한 특징으로 사용될 수 있다. 매수 신호의 경우 1, 중립의 경우 0, 매도 신호의 경우 -1의 값으로 신경망에 입력하였다.

실험 결과

당일 종가 대비 특정 기간 후의 등락을 출력값으로 하여 실험하였다. 상승일 경우 1, 하락일 경우 0의 출력값으로 학습시켰다. 학습집학과 테스트집합은 전체 데이터에서 4:1의 비율로 랜덤하게 선택하여 구성하였다. 테스트 집합의 경우 학습을 고려하여 상승과 하락의 수를 갖게 하여 학습시켰다.
<표2>는 여러 예측 기간에 따른 전체 종목에 대한 학습집합과 테스트집합의 등락 예측율이다.

예측 기간	등락예측율(%)	
	학습집합	테스트 집합
1일	60.6	53.3
2일	59.8	54.1
3일	61.2	53.8
5일	62.4	54.8
10일	70.5	61.0
20일	73.6	64.6
30일	74.1	67.6
40일	76.8	71.7
50일	77.0	73.8
60일	83.0	77.1
80일	83.1	76.6
100일	82.5	75.4
120일	82.6	77.5

<표2> 예측 기간에 따른 예측율 변화

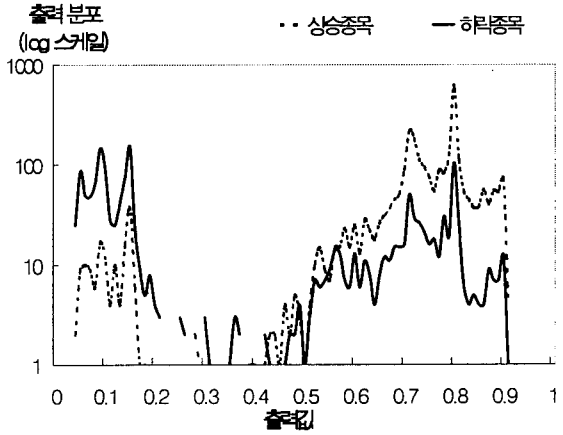
예측기간이 10일 미만이 경우 60% 미만의 등락 예측율을 보였다.

학습된 신경망에 의해 각 종목의 출력값을 구하여 0.8보다 큰 종목을 상승 예측 종목으로 0.2보다 작은 종목을 하락 예측종목으로 선정하였다. <표3>은 선정된 종목의 비율과 예측율이다.

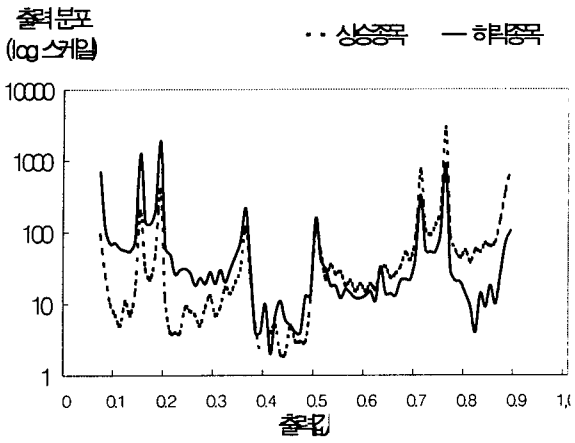
예측 기간	학습집합(37510개)		테스트집합(9378개)	
	선정갯수 (비율%)	예측율(%)	선정갯수 (비율%)	예측율(%)
1일	4876(13)	90.6	1031(11)	57.2
2일	5003(13)	92.2	938(10)	58.3
3일	5480(15)	91.1	1324(14)	57.2
5일	6752(18)	91.0	1500(16)	59.9
10일	11253(30)	91.3	2344(25)	68.8
20일	11628(31)	93.8	2813(30)	72.1
30일	13129(35)	92.7	3000(32)	75.3
40일	14629(39)	93.0	3469(37)	77.0
50일	12754(34)	95.1	3000(32)	82.0
60일	13879(37)	95.7	3188(34)	88.7
80일	14254(38)	94.0	3376(36)	87.5
100일	13504(36)	96.6	3000(32)	88.3
120일	12378(33)	94.5	3000(32)	88.1

<표3> 예측 기간에 따른 선정 종목의 비율과 선정된 종목의 예측율

테스트 집합의 수는 9378개 였다. 종목에 대한 결과를 보면 60일에 대하여 예측율이 가장 좋음을 알 수 있다. 60일의 경우 선정된 종목의 수는 전체 종목의 34%에 해당하며 선정 종목에 대한 예측율 역시 60일이 88%로 가장 좋음을 알 수 있다.
[그림1]과 [그림2]은 학습집합과 테스트 집합에 대한 출력값의 분포이다. 0.5를 기준으로 상승과 하락이 각각 우세함을 볼 수 있다.



[그림1] 실험1의 학습집합에 대한 출력값 분포



[그림2] 실험1의 테스트집합에 대한 출력값 분포

[그림 3]은 출력값에 대한 예측 정확도를 나타낸다. 0.5의 출력부근에서 최저의 정확도를 보이고 0.0과 1.0쪽으로 갈수록 예측 정확도가 증가하는 것을 볼 수 있다. 0.2와 0.8에서 각각 68.9%와 81.4%의 정확도를 보이고 있다.

[그림 6] 실험2의 테스트 집합에 대한 출력에 따른 예측정확도

결론

주가는 다른 시계열 데이터와 달리 신호대잡음비가 큰 데이타이다. 또한 다른 시계열 데이타 예측과 달리 전체의 예측의 정확도 보다는 적은 수의 대상이라도 그 예측 정확도가 더 중요하게 적용될 수 있다. 이러한 주가의 특성을 고려할 때 전체 데이터의 등락을 예측하는 것보다 예측율이 높은 데이터를 선정하는 것이 더 유용할 수 있다.

실험에 사용된 28개의 각각의 지표는 특정 기간후의 등락을 예측한다. 이때의 특정기간은 6-30일 정도이다. 그러나 각각의 지표는 전체 주가 흐름의 추세에 따라 예측이 좋은 것과 그렇지 않은 것이 있다. 그렇기 때문에 복수의 지표를 사용한다. 실험 결과 28개의 지표가 같이 사용된 경우 예측기간이 60일인 경우 그 예측 결과가 가장 좋았다. 이러한 실험결과는 복수개의 지표가 같이 사용되어 주가를 예측할 때 그 적정 예측기간이 60일이라는 점을 시사한다. 그리고 이렇게 특정 기간에 예측율이 최대점을 보이는 사실은 주가가 신호와 잡음이 혼합된 데이타라는 가정을 뒷받침해 주며, 특히 6일 미만의 단기 예측의 경우의 실험결과는, 잡음으로 인한 신호의 왜곡으로 인해 단기간의 예측이 불가능하다는 가정을 강하게 뒷받침해준다.

참고문헌

[1] Kamijo, K., Tanigawa, T., "Stock Price Pattern Recognition - A Recurrent Neural Network Approach-", In Proceedings of the IEEE International Joint Conference on Neural Networks, I 215-221, 1990

[2] Kwon, Y. S., "The Prediction of Industry stock Using Artificial Neural Networks: Cases of Construction Industry and Banking," M.S thesis, KAIST, 1996

[3] Kurt Hornik, "Multilayer Feedforward Networks are Universal Approximators", Neural Networks, Vol.2 pp.359-366, 1989

[4] Simon Haykin, "Neural Networks, A Comprehensive Foundation 2nd Ed", Prentice Hall International, Inc., 1999

[5] Robert Schalkoff, "Pattern Recognition : statistical, structural and neural approaches", John Wiley & Sons, Inc., 1992

[6] James A. Freeman, David M. Shapura, "Neural networks : algorithms, applications, and programming techniques", Addison-Wesley Publishing Company, Inc., 1992

[7] http://www.vipstock.com