creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:24011-200000236605

신경망과 의사결정트리를 이용한 Stream Data 예측 시스템 설계 및 구현

A Design and Implementation of Stream Data Prediction System Using Neural Network and Decision Tree

2008年 8月 日

朝鮮大學校 大學院

電算統計學科

梁 豪 原

梁豪原의 博士學位論文을 認准함

委 員 長	朝鮮大學校 敎授	이 웅 기	印
委 員	朝鮮大學校 敎授	정 채 영	印
委 員	朝鮮大學校 敎授	권 용 만	印
委 員	朝鮮大學校 敎授	장 인 홍	印
委 員	朝鮮大學校 敎授	배 상 현	印

2008年 8月 日

朝鮮大學校 大學院

목 차

표목차 ····· iii
그림목차 ····· iv
ABSTRACT ····· vi

Ⅰ. 서 론 ····· 1
A. 연구배경 및 관련 연구 ····· 4
B. 연구 목적 및 내용 ····· 6

Ⅱ. 주가 예측 시스템의 개요 ····· 8
A. 신경망(Neural Network) ····· 8
1. 신경망의 구조 ····· 11
2. 신경망의 특징 ····· 14
3. 신경망의 구성요소 ····· 16
4. 역전파 학습 알고리즘 ····· 17
B. SVM(Support Vector Machine) ····· 19
1. SVM의 기본원리 ····· 22
a. 경험적 리스크 최소화 원칙(ERM) ····· 22
b. VC 차원 ····· 23
c. 구조적 리스크 최소화 원칙(SRM) ····· 24
2. 최대 마진 분류기 ····· 26
3. 비선형 SVM ····· 30
a. 다항식(Polynomial) 커널함수 ····· 30
b. RBF 커널함수 ····· 31
C. 의사 결정 트리(Decision Tree) ····· 33

Ⅲ. 시스템 구성 및 설계 ····· 36
A. 시스템 구성도 ····· 36
B. 표준화 및 변수 설정 ····· 38
C. 모형 설계 ····· 40
1. 신경망을 이용한 모형 설계 ····· 43
2. 의사결정트리를 이용한 모형 설계 ····· 46
3. 개별주식에 대한 모형 설계 ····· 48
a. 삼성전자 ····· 48
b. 포스코 ····· 51

Ⅳ. 실행결과 및 평가 ····· 54
A. KOSPI 200 지수를 이용한 경우 ····· 54
1. 신경망을 이용한 경우 ····· 54
B. 개별 종목 주가를 이용한 경우 ····· 59
1. 삼성전자 ····· 59
2. 포스코(POSCO) ····· 62
3. KOSPI 200 지수의 모의실험 ····· 65

***V.* 결론** ····· 70

참고문헌 ····· 72

표 목 차

[표 3.1] 모형 구축에 이용된 변수 선정 ····· 39
[표 3.2] 자료의 분할 ····· 41
[표 3.3] 신경망 모형 설계 ····· 44
[표 4.1] 신경망의 실험 결과(KOSPI200지수) ····· 55
[표 4.2] 의사결정트리의 실험 결과 ····· 55
[표 4.3] KOSPI 200 지수를 이용한 각 기법들의 예측 성능비교 ····· 56
[표 4.4] 삼성전자 지수를 이용한 각 기법들의 예측 성능비교 ····· 59
[표 4.5] 포스코 주가를 이용한 각 기법들의 예측 성능비교 ····· 62
[표 4.6] 모의실험을 위해 사용된 기간 ····· 65
[표 4.7] 각 기법들의 예측 성능비교 ····· 65
[표 4.8] KOSPI 200지수와 미국 지수간의 상관관계 ····· 66
[표 5.1] 각 기법간 예측 성능 비교 ····· 70
[표 5.2] 알고리즘의 예측 성능비교 ····· 71

그림 목차

[그림 2.1] PE의 기본 구조와 신경망 ····· 10
[그림 2.2] 인공신경망의 일반적인 구조 ····· 11
[그림 2.3] 인공신경망의 활성함수의 종류 ····· 12
[그림 2.4] 4개의 입력노드를 가진 단층 퍼셉트론 ····· 13
[그림 2.5] 역전파 학습 알고리즘 ····· 17
[그림 2.6] SVM을 이용한 이진 분류 ····· 20
[그림 2.7] 커널 함수의 이용 ····· 21
[그림 2.8] 리스크의 경계범위 ····· 24
[그림 2.9] 최적 경계면의 개념도 ····· 26
[그림 2.10] 최대 마진 분류기 ····· 27
[그림 2.11] 다항식 커널함수 ····· 31
[그림 2.12] RBF 커널함수를 이용한 SVM ····· 32
[그림 2.13] 의사 결정 트리의 구조 ····· 33
[그림 3.1] 시스템 구성도 ····· 37
[그림 3.2] 모형 설계에 사용된 데이터 ····· 40
[그림 3.3] 신경망과 의사결정트리 알고리즘의 도식도 ····· 42
[그림 3.4] 신경망 모형의 최적화 은닉층 노드수 결정 ····· 43
[그림 3.5] 신경망 모형의 추정된 연결강도 ····· 45
[그림 3.6] 의사결정트리의 분석흐름도 ····· 46
[그림 3.7] 의사결정트리 모형의 분리 기준 ····· 47
[그림 3.8] 의사결정트리 모형 설계 결과(KOSPI200지수) ····· 47
[그림 3.9] 알고리즘의 모형도(삼성전자) ····· 48
[그림 3.10] 신경망 모형의 추정된 연결강도(삼성전자) ····· 49

[그림 3.11] 의사결정트리 모형 설계 결과(삼성전자) ····· 50
[그림 3.12] 알고리즘의 모형도(포스코) ····· 51
[그림 3.13] 신경망 모형의 추정된 연결강도(포스코) ····· 52
[그림 3.14] 의사결정트리 모형 설계 결과(포스코) ····· 53
[그림 4.1] Lift Chart(KOSPI200지수) ····· 57
[그림 4.2] ROC Curve(KOSPI200지수) ····· 58
[그림 4.3] Lift Chart(삼성전자) ····· 60
[그림 4.4] ROC Curve(삼성전자) ····· 61
[그림 4.5] Lift Chart(포스코) ····· 63
[그림 4.6] ROC Curve(포스코) ····· 64
[그림 4.7] KOSPI200지수와 DOW지수의 변동 그래프 ····· 67
[그림 4.8] KOSPI200 지수와 NASDAQ 지수의 변동 그래프 ····· 67
[그림 4.9] Test 2 기간 내 KOSPI200 지수와 DOW지수의 변동 그래프 ····· 68
[그림 4.10] Test 2 기간 내 KOSPI200지수와 NASDAQ지수의 변동 그래프 ····· 69

Abstract

A Design and Implementation of Stream Data Prediction System Using Neural Network and Decision Tree

Yang, Ho-Won
Advisor : Prof. Bae, Sang-Hyun, Ph.D.
Department of Computer Science and Statistics,
Graduate School of Chosun University

At stock market, about stock price index prediction the research was attempted to use many techniques from various field. Generally the stock price index prediction is using the fundamental analysis which uses the immanence value of technical analysis and stock market. But, stock market of fact about reduction prediction or estimate of situation the description below is difficult with complexity and uncertainty of stock price index. In this research, In order to raise effectiveness of stock price prediction system uses the neural network and a decision-making tree algorithm. because Information which are useful is latent below-mentioned mining. In order accomplishes a real-time data mining effectively from prepares the data which is appropriate probably has the necessity to which what kind of data becomes mining. It processes a data with the form which is suitable at mining and will use at mining the technique which and selects the result which has become mining and the chain which applies in prediction interprets process in necessity. Also data mining

relationship of the patterns which are useful is immanent in data or variables between uses the analytical model which is elaborate and seeks and accomplishes the role which reconfirms the experience knowledge of existing with the work and does not recognize simultaneously until now, provides new information and is a system which predicts. Consequently, this research uses the neural network and a decision-making tree and predicts the hereafter (tomorrow) rise and a depreciation of KOSPI 200 quotients the dual classification (binary classification) model which the prediction efficiency of each classification models and it plans it compares. The different meaning goal uses the prediction model which finally is constructed and to embody the prediction system of the individual stocks from stocks spot market, the contents which is concrete with afterwords is same.

1. Embodies the dual classification system which predicts the rise and a depreciation of parting stock price index from stocks spot market.
2, Uses a decision tree algorithm about the data which has a time series quality and reveals the prediction system which is efficient embodies the union or a relation of data and between.
3. After constructing the neural network and a decision-making tree model, divides TEST1 and TEST2 duration and simulation leads and the system which raises the accuracy of prediction about future point of view embodies.
4. Standardization process and above removal leads and application embodies the system which is possible from the actual money market.

Ⅰ. 서론

주식시장에서 주가 지수 예측에 대한 연구는 여러 분야에서 많은 기법들을 이용하여 시도되었다. 일반적으로 주가 지수 예측은 기술적 분석과 주식시장의 내재 가치를 이용한 기본적 분석을 이용하고 있으나 실제의 주식시장에서는 주가 지수의 복잡성과 불확실성으로 인하여 예측이나 상황 판단을 하기는 어렵다. 설령 예측을 하였다 하더라도 빗나가는 수가 많다. 이렇게 주가 지수는 사회 전반적으로 발생되는 문제에 대한 결과라고 할 수 있을 정도로 다양한 요인들에 의해 계속적으로 변동하기 때문에 정확한 결과를 이끌어 내기가 힘들다. 주가지수 예측에 관한 과거의 기존 연구들은 통계적 기법을 기반으로 예측하였지만 엄격한 가정에 의해 사용의 제한이 있어 좋은 연구 성과를 보여주지 못하였다[1]. 데이터 마이닝은 방대한 데이터에서 의미 있는 정보를 추출해 내는 데에 유용하게 쓰인다. 많은 데이터 마이닝 방법들이 이미 마련되어 변하지 않는 데이터들에서 정보를 뽑아내는 것에 기초를 두었다. 이러한 방법들은 데이터 마이닝을 시작하기 전에 필요한 데이터가 모두 갖춰져 있어야만 하고 마이닝이 진행되는 동안 시간이 얼마나 지나는지 여부와 상관없이 이 데이터가 변하지 않는다는 특징을 가지고 있다. 그러나 데이터가 시간이 지남에 따라 계속 추가 된다면 이렇게 계속 추가되는 데이터에 맞게 마이닝하는 방법이 더욱 유용할 것이다. 더군다나 데이터가 시간의 관계없이 늘기만 하는 것이 아니라 시간에 영향을 받는 다면 시간의 순서에 따른 data의 변화에 초점을 더 두어야 할 것이다.

주식시장에서 주식 가격의 변화에 미치는 요인, 주식 가격의 예측, 시간의 흐름에 따른 소비자들의 구매 변화 등등도 이러한 범주에서 분석될 수 있을 것이다. 변하지 않는 데이터를 대상으로 데이터 마이닝을 할 때에는 그 데이터에 가장 잘 맞는

방법을 선택하고 이 방법이 분석과정 내내 변할 필요가 없겠지만, 데이터가 시간의 흐름에 따라 변한다면 그것을 분석하는 방법 또한 시간의 흐름에 따라 변화되어야 할 것이다. 또한 이렇게 시간이 지남에 따라 계속 증가하는 데이터는 데이터가 늘어남에 따라 물리적인 저장 공간이 계속 늘어나야 할 것이다. 그러나 방대한 양의 데이터가 계속 늘어나는 경우 거기에 맞추어 계속 저장 공간을 늘리는 것은 물리적인 제한이 따르거나 혹은 엄청난 비용을 수반할 것이다. 따라서 자원 사용의 문제에 있어서도 시간의 흐름에 따라 변하는 데이터의 경우에는 고정되고 변하지 않는 데이터 마이닝 할 때와는 다른 방법이 선택 되어야 할 것이다. Stream data에 알맞은 마이닝 방법이 선택되면 이로부터 자료간의 조합이나 관련성을 밝히고 또한 효과적으로 새로운 데이터를 예측할 수 있게 되는 것이다. 이런 분류 및 예측의 문제들을 보완하기 위해 최근에 데이터 마이닝의 다양한 패턴분류(pattern classification) 모형들이 금융지수 분야에 적용되고 있다[2][3]. 그 중 신경망에 의한 예측 모형은 주가 변동을 나타내는 통계 모형의 설정이 필요하지 않는다는 특징을 가지고 있어 현실적으로 존재하는 거래제도나 규제 등 시장의 불완전성으로 인한 특이 패턴을 추정하기 위해 효율적으로 사용되어 왔다. 최근에 분류기법에서 집중적으로 연구되고 있는 SVM(Support Vector Machine : SVM) 모형은 1995년 Vapnik에 의해 이원(binary) 패턴분류 문제를 해결하기 위해 제안된 학습 방법이다[4]. 기존 신경망에서는 학습 알고리즘이 학습 집단을 이용하여 학습오류를 최소화하는 경험적으로 위험을 최소화하는 원칙(Empirical Risk Minimization : ERM)을 구현한다[5]. 따라서 신경망은 모형을 일반화하기 어렵고 모형이 과대적합(overfitting) 되는 단점을 갖는 반면에 SVM은 구조적 위험을 최소화하는 방법(Stucture Risk Minimization : SRM)에 기반을 두고 있어 ERM보다 우수한 일반화 성능을 얻을 수 있다. 최근에는 SVM을 응용한 연구영역이 문자 인식, 영상 인식, 홍채 인식, 그리고 문서분류 등의 다양한 분야에서 매우 활발히 진행되고 있다.

본 논문의 목적은 인공지능 기법인 신경망을 이용하며, 스트림 데이터마이닝 기법 중 의사결정 트리를 이용하여 KOSPI 200지수의 향후(내일) 상승 및 하락을 예측하는 이원분류(binary classification)모형을 설계하고 각 분류 모형들의 예측 성능을 비교하고, 최종적으로 구축된 예측모형을 이용하여 주식현물시장에서 개별주식의 예측 시스템을 구현하는 것이다. 본 연구의 구성은 총 5장으로 구성되어 있다. 제 1장에서는 연구의 배경과 목적 및 연구 방법에 관해서 설명하였고, 제 2장 본 시스템에서 알고리즘으로 사용한 신경망과 의사결정 트리의 이론을 연구하였고, 제 3장에서는 시스템 구성 및 설계에 대해서 연구하였고, 제 4장에서는 KOSPI 200지수와 개별주식 지수에 대한 실험 결과들을 정리하고, 제 5장에서는 시스템 구현 및 결론에 대해 기술하고 향후 연구 과제를 제시하였다.

A. 연구의 배경 및 관련 연구

주가지수 예측에 관한 과거의 기존 연구들은 통계적 기법을 기반으로 예측하였지만 엄격한 가정에 의해 사용의 제한이 있어 좋은 연구 성과를 보여주지 못하였다. 이런 분류 및 예측의 문제들을 보완하기 위해 최근에 데이터마이닝의 다양한 패턴 분류(pattern classification) 모형들이 금융지수 분야에 적용되고 있다. 국외의 경우를 살펴보면 Lapedes 와 Farber(1987)의 연구는 10주의 과거 자료를 토대로 다음 주의 주가 지수를 예측하는 방법으로 신경망이 단순 이동평균법보다 나은 예측력을 갖고 있음을 최초로 보였다[6]. Kimoto et al(1990)은 기술적 지표 및 경제지표와 주식을 매수, 매도하는 시점(timing)간의 관계를 인공신경망을 통하여 도쿄 주식시장 1부 종목의 모든 주식으로 구성된 TOPIX를 대상으로 추론하는 실험을 하여 단순한 매입 후 보유전략(buy-and-hold)보다 인공신경망이 우수함을 보였다[7]. George Tsibouris 와 Matthew Zeidenberg(1995)은 신경망 모형을 이용하여 미래의 주가의 방향을 예측하고, 예측의 정확성을 이용하여 효율적 시장 가설을 검증하고자 하였다. 인공신경망의 학습알고리즘 중에서 4가지의 학습규칙을 각각 적용하여 성과를 비교하였다[8][9]. 국내의 경우에는 노종래와 이종호(1991)가 최초로 인공신경망을 적용하여 종합주가지수의 등락을 예측한 결과 월별 예측의 경우에는 비교적 우수했지만 주별 예측의 경우에는 좋지 않은 결과를 보여주었고, 이재유(1993)는 주간 신용잔고, 고객 예탁금과 같은 자료를 입력변수로 하여 종합주가지수와 개별주식들의 주가를 예측하는 역전파 인공신경망을 구축하여 실험한 결과 통계적 모형보다 인공신경망 모형의 예측이 훨씬 좋은 결과를 가져왔다[10][11]. 김관문과 조광재(1994)는 주가 자료 및 거시경제지표 등을 입력변수로 하여, 인공신경망과 통계적기법(다변량 판별분석)으로 종합주가지수를 예측해 모의투자한 결과, 거래수수료를 감안하고도 두 모형 다 높은 수익을 거두었고, 인공신경망이 다변량 판별

분석보다 다소 우월하나 그 차이가 크지 않았다. 박종엽과 한인구(1995)는 종합주가지수, 거래량, 예탁금, 사채수익률, 원달러, 원엔, 다우존스지수를 중요 입력변수로 하여 신경망을 이용한 KOSPI 예측 시스템을 개발한 후 신경망 모형의 성과를 평가하기 위해 모의 투자한 결과 판별 분석보다 높은 수익률을 나타냈다[12]. 현창호(1996)는 인공신경망을 이용하여 일별 종합주가지수의 예측을 시도하였다. 입력변수로는 KOSPI, 일일거래량, 일일거래대금, 사채수익률, 원달러, 주가지수 수익률을 사용하여 인공신경망 모형과 VAR(Vector Auto Regression)모형의 성과를 비교한 결과 인공신경망이 VAR모형보다 적은 예측오차를 보여 인공신경망의 성과가 더 좋음을 보였다[13]. SVM을 이용한 연구들은 거의 문자 인식, 영상 인식, 홍채 인식 그리고 문서분류 등의 분야에서 우수한 분류 성능을 보이고 있지만, 아직까지는 국내에서는 주식 시장에서 주가 지수의 상승 및 하락을 예측하는 분류 성능 연구는 거의 없는 실정이다[14].

B. 연구 목적 및 내용

기존의 데이터베이스 관리 시스템에 저장되는 데이터의 양이 증가함에 따라서 저장된 자료로부터 원하는 정보를 효과적으로 추출하거나 데이터에 내재된 규칙들을 찾는 일이 대단히 어렵게 되고 있다. 본 논문에서는 신경망과 의사결정 트리 알고리즘을 이용하여 많은 잠재되어있는 유용한 정보들을 마이닝 함으로써 주가예측 시스템의 효율성을 높이고자 한다. 실시간 데이터 마이닝을 효과적으로 수행하기 위해서는 많은 사전 및 사후 작업이 필요하다. 즉 어떠한 데이터가 마이닝 될 필요가 있는지 적절한 데이터를 준비하고 마이닝에 적합한 형태로 데이터를 가공하고 마이닝에 사용할 기법을 선택하고 마이닝된 결과를 해석하고 예측에 활용하는 일련의 과정을 필요로 한다. 데이터마이닝은 데이터 베이스에 저장된 데이터에서 유용한 정보를 추출하는 작업이다. 또한 데이터마이닝은 데이터에 내재되어 있는 유용한 패턴이나 변수들 간의 관계를 정교한 분석 모형을 사용하여 찾아내는 작업으로 기존의 경험적 지식을 재확인하는 역할을 수행함과 동시에 지금까지 인식하지 못했던 새로운 정보를 제공하여 예측하는 시스템이다. 이러한 데이터 마이닝 기법에는 의미 있는 몇 개의 군집으로 나누는 클러스터링 기법과 다른 클래스에 대한 차별적 특성을 도출하는 분류 기법, 관련성을 구하는 연관 규칙(Association Rule) 기법 , 일반화된 대표적 표현으로 축약시키는 요약(Summarization) 기법 등 다양한 기법들이 있다. 이에 본 논문에서는 신경망과 의사결정 트리를 이용하여 KOSPI 200지수의 향후(내일) 상승 및 하락을 예측하는 이원분류(binary classification)모형을 설계하고 각 분류 모형들의 예측 성능을 비교하고, 최종적으로 구축된 예측모형을 이용하여 주식현물시장에서 개별주식의 예측 시스템을 구현하는 것을 목적으로 하며, 그 구체적인 내용은 다음과 같다.

1. 주식 현물 시장에서의 일별 주가지수의 상승 및 하락을 예측하는 이원분류 시스템을 구현한다.

2, 시계열 특성을 갖는 데이터에 대해 의사결정트리 알고리즘을 이용하여 자료간의 조합이나 관련성을 밝히고 효율적인 예측 시스템을 구현한다.

3. 신경망과 의사결정트리 모형을 구축한 후 TEST1, TEST2 기간을 나누어 모의실험을 통해 미래 시점에 대한 예측의 정확도를 높이는 시스템을 구현한다.

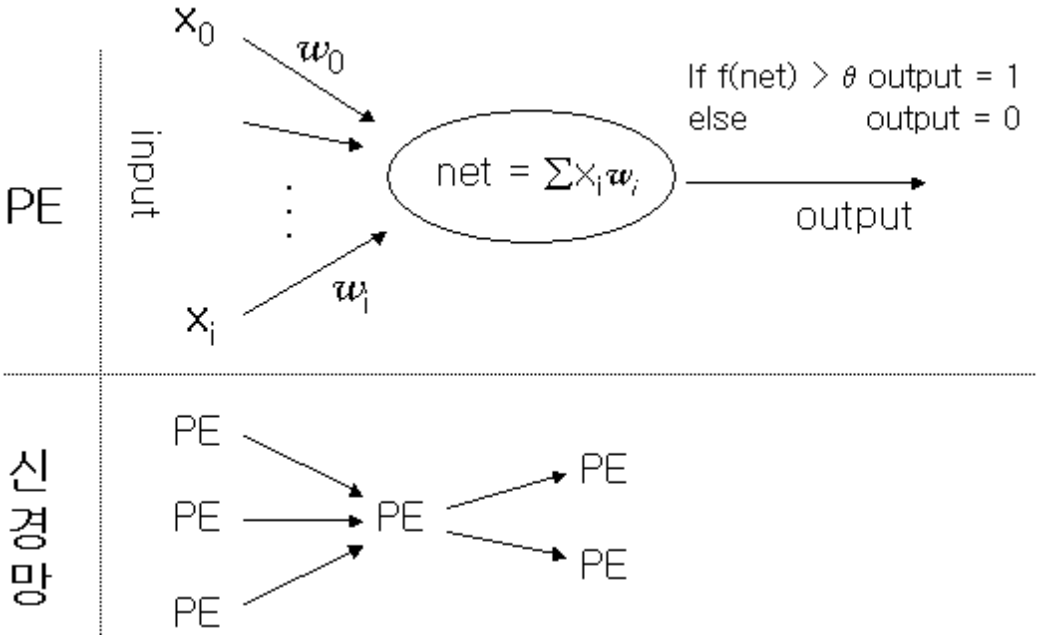
4. 표준화 과정 및 이상치 제거를 통하여 실제 금융시장에서 적용 가능한 시스템을 구현한다.

Ⅱ. 주가 예측 시스템의 개요

A. 신경망(Neural Network)

인공 신경망 모델에 관한 연구는 1943년 워렌 맥컬럭(Warren McCulloch)과 월터 피트(Walter Pitss)로부터 시작되었다. 맥컬럭과 피츠의 모델은 네트워크 내의 단순한 요소들의 연결을 통하여 무한한 컴퓨팅 능력을 가진다는 점에서 많은 관심을 불러 일으켰다. 이들의 모델은 인간 두뇌에 관한 최초의 논리적 모델링이라는 점에서 그 중요성이 매우 크며, 따라서 이 맥컬럭과 피츠의 모델이 인공 신경망 이론의 효시로 여겨지고 있다[15]. 그 이후 1949년에는 캐나다의 심리학자 도널드헵(Donald Hebb)이 두 뉴런 사이의 연결강도를 조정할 수 있는 학습규칙을 발표하였으며, 프랭크 로젠블럿(Frank Rosenblatt)이 1957년에 '퍼셉트론'이라는 최초의 인공 신경망 모델을 발표하였다. 이 모델은 처음 소개되었을 때 상당한 관심을 불 러 일으켰으며 여러 분야에 걸쳐 큰 영향을 끼쳤다. 이 퍼셉트론 모델은 당시에 매우 가능성이 큰 연구로 여겨졌으나 그 후 여러 가지 문제점들이 제시되었다. 민스키와 파퍼트가 '퍼셉트론즈'라는 저서를 통해 퍼셉트론 모델을 수학적으로 철저히 분석하여 퍼셉트론의 단점들을 밝혀내고 난 후에는 인공 신경망에 관련된 연구는 약 20년간 침체의 길을 걷게 된다. 그 이후 인공지능에 관한 연구가 진행되던 중 퍼셉트론의 문제점을 극복하여 네트워크 구성 시에 한 개의 계층을 더 추가해 구성한 다층 퍼셉트론의 개념이 등장하면서 인공 신경망 연구에 관한 새로운 계기를 마련하였다. 즉 은닉층(Hidden Layer)과 역전파(backpropagation) 학습 알고리즘을 이용함으로써 선형분리 문제뿐만 아니라 여러 가지 문제점 들을 해결할 수 있게 된 것이다[16]. 신경망은 인간의 뇌 구조를 이용하여 모델링 된 알고리즘으로 패턴 인

식 등 그 활용 범위가 대단히 광범위하다. 물질의 성질을 잃지 않는 최소 단위가 분자인 것처럼, 인간의 뇌를 포함한 신경들을 구성하는 최소 단위는 뉴론(neuron)이다. 뉴론과 뉴론의 연결은 시냅스(synapse)라는 물질이 관여를 한다. 아래 [그림 2.1]은 신경망의 최소 단위가 되는 PE(Processing Element)이다. Neural Network(신경망)은 인간 두뇌의 신경세포를 모델링 하여 지능을 구현하고자 하는 기법이다[17]. 신경망은 병행적으로 상호작용하는 여러 개의 계산요소들로 이루어져 있으며, 각 계산요소는 가중치 합(weighted sum)과 같은 단순한 계산만을 수행한다. 신경망의 계산능력은 다량의 계산요소들이 병행적으로 계산을 수행함으로써 얻어진다. 신경망은 학습데이터를 통하여 가중치를 학습하고, 그 데이터에 대한 특성을 일반화시킴으로서 학습 기능을 수행한다. 이처럼 신경망을 데이터마이닝 알고리즘에 적용할 경우 비선형적인 패턴분할을 수행할 수 있는 신경망의 장점 때문에 매우 정확한 지식을 생성할 수 있다[18][19]. 또한 신경망의 학습기능과 병렬분산처리 능력, 결함허용성 등도 신경망을 이용할 경우에 얻어지는 장점들이다. 그러나 신경망을 사용할 경우 생성되는 지식의 정확도는 매우 높지만, 얻어진 지식에 대한 설명을 하기 어려운 문제점을 가지고 있다. 이를 신경망 모델의 불투명성(opaqueness)이라고 한다.

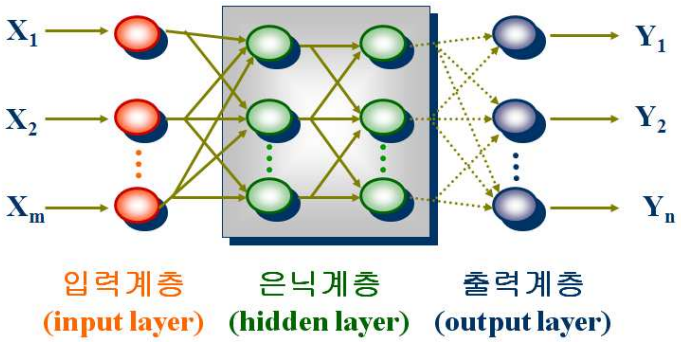


[그림 2.1]. PE의 기본 구조와 신경망

입력의 각 성분에 가중치(weight)를 곱한 값이 net 값이 되며, 출력은 net 값과 함수 f에 따라 달라진다.

1. 신경망의 구조

인공 신경망은 은닉마디라고 불리는 독특한 구성요소에 의해서 일반적인 통계 모형과 구별된다. 은닉마디는 인간의 신경세포를 모형화한 것으로 각 은닉마디는 입력변수들의 결합을 수신하여 활성함수라는 비선형 함수를 통해 목표변수에 전달한다. 이 때 결합에 사용되는 계수(coefficient)를 연결강도(synaptic weights)라 하며 활성함수는 입력값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하는 역할을 한다. 또한 인공신경망은 학습규칙이라는 반복적 알고리즘을 통하여 주어진 문제를 해결해 나간다[20].

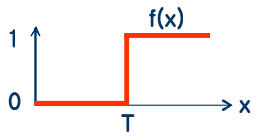


[그림 2.2] 인공신경망의 일반적인 구조

인공신경망 모형은 적용되는 신경망의 종류, 학습규칙 그리고 적용되는 활성함수의 종류에 따라 다양하게 분류될 수 있다. 따라서 인공신경망 알고리즘을 적용하기 위해서는 특정 문제의 해결에 적합한 알고리즘과 학습규칙, 활성함수의 선택이 필요하다[21].

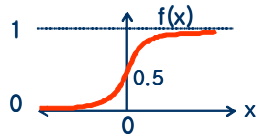
step function (계단함수)

$$F(x) = \begin{cases} 1 & \text{if } x > T \\ 0 & \text{otherwise} \end{cases}$$



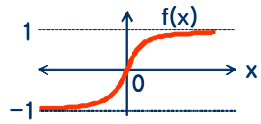
sigmoid function
(단극성 시그모이드 함수)

$$F(x) = \frac{1}{1+e^{-x}}$$



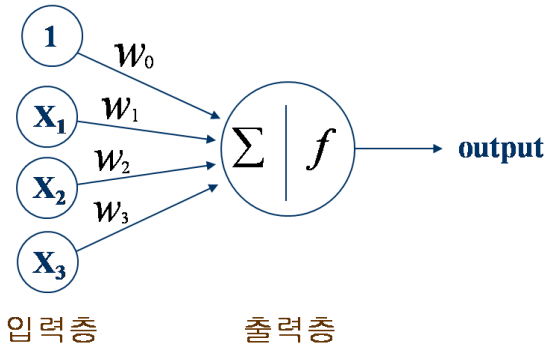
hyperbolic tangent function
(양극성 시그모이드 함수)

$$F(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$



[그림 2.3] 인공신경망의 활성함수의 종류

인공신경망 알고리즘은 해결하고자 하는 문제의 답 즉, 목표변수가 존재할 경우에 사용하는 지도학습법(supervised learning)과 목표변수의 값이 존재하지 않는 자율학습으로 구분할 수 있다. 지도학습에 속하는 인공신경망 알고리즘은 퍼셉트론, 다층퍼셉트론, RBF 신경망이 있으며 비지도학습(unsupervised learning)에 속하는 인공신경망 알고리즘으로는 코호넨의 자기조직화 네트워크가 있다[22]. 본 연구에서는 다층 퍼셉트론(Multi Layer Perceptron : MLP)의 역전파(backpropagation) 알고리즘을 사용하였다.



[그림 2.4] 4개의 입력노드를 가진 단층 퍼셉트론

2. 신경망의 특징

신경망은 단순한 기능을 가진 무수히 많은 신경세포 또는 처리소자(processing element)들이 병렬 연결된 연산구조로 되어 있다. 그 특징은 다음과 같다.

(1) 각 신경세포는 다른 신경세포들과 완전히 독립된 기능을 갖는다. 즉, 각 신경세포의 출력은 자신과 연결된 연결강도를 통하여 직접 전달되는 정보에만 의존할 뿐, 다른 정보들과는 무관하다. 이와 같은 특징으로 인하여 병렬처리가 가능하므로 연산속도가 매우 빠르다.

(2) 신경망은 무수히 많은 연결강도를 가지고 있다. 따라서, 정보의 분산표현 및 처리가 가능하다. 또, 중복성 (redundancy)이 크므로 fault - tolerant 할뿐만 아니라 일부의 정보로부터 전체를 얻을 수 있는 연산기억 (associative memory) 특성을 갖는다.

(3) 학습이나 훈련을 통해 연결강도를 조정함으로써 새로운 정보를 추가하거나 변경할 수 있는 적응 특성을 가지고 있다.

신경망은 알고리즘을 프로그램 명령어에 의해 순차적으로 처리하는 von Neumann 방식의 컴퓨터로는 해결하기가 곤란한 다음과 같은 문제들에 적합하다고 알려져 있다[23].

(1) 패턴분류 (pattern classification) : 입력패턴들을 그 특성에 따라 분류한다. 이때, 분류된 집단 (class)이나 군 (cluster)은 그것을 대표하는 고유한 패턴으로 식별될 수 있다.

(2) 패턴완성 (pattern completion) : 불완전한 패턴으로부터 완전한 패턴을 생성한다. 잡음제거 문제도 여기에 속한다.

(3) 함수근사 (function approximation) : 입력패턴에 대응하는 출력패턴을 생성한

다. 이 때, 입력과 출력 사이의 대응관계는 임의의 함수 또는 Mapping이라도 가능하다. 패턴정합 문제도 넓은 의미에서 여기에 속한다고 볼 수 있다.

이러한 문제들의 공통적인 특징은 다음과 같다. 첫째, 주어진 문제를 수학적 표현으로 정의하기가 곤란하기 때문에 알고리즘에 따라 해를 구하기는 어렵지만, 사람은 그 해법에 관한 노하우를 가지고 있다는 것이다. 둘째, 주어진 문제의 입력패턴과 출력패턴의 예를 충분히 많이 생성해 낼 수 있어서 이들로부터 입력과 출력 사이의 일반적인 관계를 추출하는 것이 가능하다. 이를 위해서는 처리해야 할 정보의 양이 매우 많다. 따라서 정보를 순차적으로 처리하는 컴퓨터로는 해를 구하는 과정이 매우 비효율적이고 시간도 많이 걸린다. 그러므로 병렬구조를 가지고 많은 양의 데이터를 사용하여 학습하는 신경망이 이와 같은 문제들을 해결하는데 적합할 것이다. 신경망 또는 신경회로 컴퓨터는 실제로 패턴인식, 음성인식, 음성합성, 자동번역, 최적화, 적응제어, 수치 예측 등 많은 분야에 응용되고 있다.

3. 신경망의 구성요소

생물학적 신경망이 단순한 신경세포들의 대단위 병렬연결로 이루어져 있듯이, 신경망도 단순한 기능을 수행할 수 있는 신경세포들의 수많은 병렬연결로 이루어져 있다. 기능면에 있어서도 생물학적 신경망과 마찬가지로 병렬분산 처리를 할 수 있을 뿐만 아니라, 학습이나 훈련을 통해서 연결강도를 조정하여 정보를 추가하거나 변경할 수 있는 적응특성을 가지고 있다. 이와 같이 거시적 차원에서 볼 때, 신경망은 생물학적 신경망의 구조와 기능을 닮았다고 할 수 있다. 그러나 신경세포들의 구체적인 기능, 연결선을 통한 신호의 전달체계, 연결강도의 조정 메카니즘 등과 같은 미시적 차원의 구조와 기능은 생물학적 신경망과 많은 차이가 있거나, 전혀 다를 수도 있다[24]. 신경망을 구성하는 기본적인 요소로는 다음과 같은 세 가지를 꼽을 수 있다.

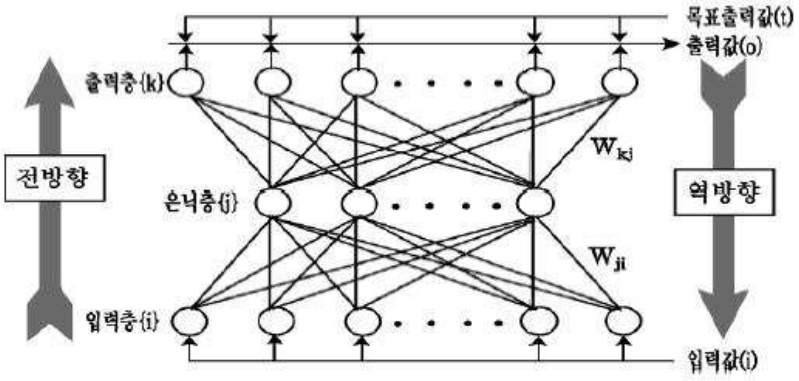
(1) 신경세포 (neuron) 또는 처리소자 (processing element)

(2) 연결강도 (connection weight)

(3) 학습규칙 (learning rule) 또는 훈련 알고리즘 (training algorithm)

4. 역전파 학습 알고리즘

역전파 알고리즘은 다층 전방향 퍼셉트론(Multilayer feed-forward perceptron)에서 사용되는 학습 알고리즘이며, 학습의 방법은 지도 학습(supervised learning)이다. 즉, 학습을 하기 위해서는 입력 데이터와 원하는 출력 (o) 데이터가 있어야 한다. 간단히 학습의 개념을 살펴보면 먼저, 입력이 신경망의 가중치(weights)와 곱하고 더하는 과정을 몇 번 반복하면 입력의 결과값인 출력 (y)이 나온다. 이 때 출력 (y)은 학습 데이터에서 주어진 원하는 출력 (o)와 다르다. 따라서 인공신경망의 학습에서 (y-o)만큼의 오차 (e)가 발생하며, 오차에 비례하여 출력층의 가중치를 갱신하고, 은닉층의 가중치를 갱신한다. 가중치를 갱신하는 방향이 신경망의 처리 방향과는 반대 방향이다. 즉 신경망의 처리는 입력층 → 은닉층 → 출력층의 방향으로 진행되며, 가중치 갱신의 학습방향은 출력층 → 은닉층으로 진행된다. 이런 이유로 역전파 알고리즘이라고 한다[25].



[그림 2.5] 역전파 학습 알고리즘

역전파 알고리즘의 단계를 간단히 정리하면 아래와 같다.

1. 신경망에 입력 데이터를 입력 노드에 적용하고, 입력에 따른 출력을 계산한다.

2. 입력에 따른 출력과 원하는 출력간의 오차를 계산한다.

3. 오차를 줄이기 위해 가중치의 증감 여부를 결정한다.

4. 각각의 가중치를 얼마나 변화시킬 것인가를 결정한다.

5. 4 단계에서 결정된 값으로 가중치를 갱신(변화)한다.

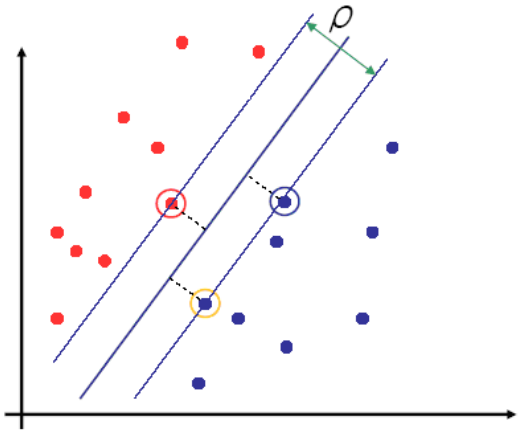
6. 모든 학습 데이터에 대해 오차가 적정 수준으로 감소하기까지 1단계에서 5 단계를 반복한다.

하지만, 역전파 알고리즘은 다음과 같은 몇 가지 문제점을 지니고 있다.

첫째, 역전파 알고리즘이 지역적 최소점(local minimum)에 빠질 우려가 있다. 역전파 알고리즘의 기본적인 원리는 최급하강법(gradient descent method)이다. 즉 최급하강법에 의한 학습은 전역적 최소점(global minimum)을 목표로 해서 가중치를 조정하는 것이 아니라 현재의 위치에서 경사면을 따라 내려가는 것이기 때문에 오차가 0이 아닌 지역적 최소점에서 더 이상 학습결과를 개선하지 못하고 정체될 수 있다. 두 번째 문제점은 전방향 신경망과 관련된 문제로서, 이는 이 방법이 기본적으로 신호가 입력에서 출력까지 한 방향으로만 진행되는 전방향 방식이라는 것이다. 이러한 전방향 신경망은 고차원의 시스템을 구성하는데는 한계가 있는 것으로 알려져 있다. 이외에도 역전파 알고리즘은 학습이 완료되기까지 많은 횟수의 반복학습과 응용분야에 따라 상이하게 학습모수를 조정하는 것이 필요하고, 추가 학습시 전체적인 재학습이 요구된다. 또한 학습의 완료시점을 예측할 수 없다는 문제점을 가지고 있다.

B. SVM(Support Vector Machine)

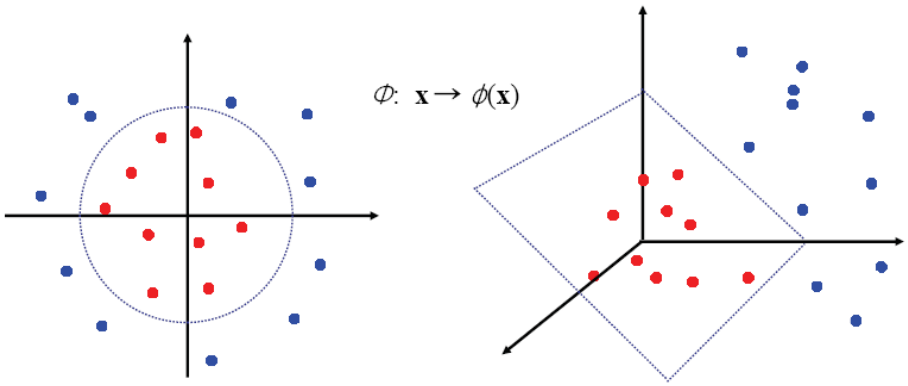
Support vector machine (SVM) 은 분류 (classification) 와 회귀 (regression) 에 응용할 수 있는 지도학습 (supervised learning) 의 일종으로서 기본적인 분류를 위한 SVM은 입력 공간에 maximum-margin hyperplane을 만든다. 학습데이터와 범주 정보의 학습진단을 대상으로 학습과정에서 얻어진 확률분포를 이용하여 의사결정함수를 추정한 후 이 함수에 따라 새로운 데이터를 이원 분류하는 것으로 VC(Vapnik-Chervonenkis) 이론이라고도 한다. 특히 SVM 은 분류 문제에 있어서 일반화 기능이 높기 때문에 많은 분야에서 응용되고 있다. 기존의 학습 알고리즘은 학습 집단을 이용하여 학습오류 (empirical error)를 최소화하는 경험적 위험 최소화 원칙 (Empirical Risk Minimization: EMR)을 구현하는 것인데 비해 구조적 위험 최소화 원칙 (Structural Risk Minimization: SRM) 은 전체집단을 하위 집단으로 세분화한 뒤 이 집단에 대한 경험적 위험도를 최소화하는 의사결정함수를 선택하는 것이다[26]. 어떤 학습데이터는 패턴분류 문제에 있어서 다른 데이터들보다 더 결정적인 영향력을 가질 수도 있으므로 우선적으로 올바르게 분류 되어야 하며 잡음 (noise) 또는 이상치 (outlier) 들은 기계학습 과정에서 영향력이 작아지길 원한다. SVM 알고리즘을 이용하는 이유는 첫째, 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내고, 셋째, 적은 학습 자료만으로 신속하게 분별학습을 수행할 수 있기 때문이다. SVM 은 원래 이진분류 (binary classification)를 위하여 개발되었으며 현재에는 생물 정보학(bioinformatics), 문자인식, 필기인식, 얼굴 및 물체 인식 등 다양한 분야에서 성공적으로 적용되고 있다[27][28].



(ρ = Margin)

[그림 2.6] SVM을 이용한 이진 분류

Support Vector Machine은 이차원 데이터 분류문제에서 가장 최적의 초평면(Hyperplane)을 구하여 이를 경계 결정면으로 선택한다. 최적의 초평면은 선형 분리가 가능한 두 집단에 대해 [그림 2.6]과 같이 마진을 최대로 하여 집단을 구분짓는다. 하지만 실제 문제의 경우 선형적으로 구성되지 않기 때문에 커널 함수를 이용하여 비선형적 특징공간을 선형적 특징공간으로 매핑(mapping)한 후에 선형 SVM으로 분류하게 된다[29][30][31]. 분리 가능의 경우 두 개의 범주로 구성된 N개의 객체가 p차원 공간에 위치하게 된다. 선형 SVM은 기존의 경험적 리스크 최소화 보다 뛰어난 성능을 가지고 있는 구조적 리스크 최소화 이론에 기반한다. [그림 2.7]은 선형적으로 구성되지 않은 input 데이터를 커널 함수를 이용해 다차원으로 매핑한 후 분류하는 하나의 예이다.



[그림 2.7] 커널 함수의 이용

1. SVM의 기본원리

a. 경험적 리스크 최소화 원칙(ERM)

일반적으로 예상 리스크 $R(\alpha)$를 추정하기 위해서 필요한 확률분포 $P(x,y)$는 미지의 분포이기 때문에 $R(\alpha)$를 계산하는 것은 어려운 일이다. 따라서, 경험적 리스크(empirical risk) $R_{emp}(\alpha)$은 학습 집단(training set)에서의 오차율(error rate)의 계산을 위해 사용된다.

$$R_{emp}(\alpha) = \int \frac{1}{2l} \, | y_i - f(x_i, \alpha) |$$

경험적 리스크인 $R_{emp}(\alpha)$은 확률분포 값인 $P(x,y)$가 포함되어 있지 않기 때문에 손실값을 $\frac{1}{2}|y_i - f(x_i,\alpha)|$ 을 계산할 수 있으며, 예상 리스크를 최소화시킬 수 있다. 경험적 리스크 최소화의 원칙(ERM)은 $R_{emp}(\alpha)$이 $R(\alpha)$로 수렴하면, $R_{emp}(\alpha)$의 최소값도 $R(\alpha)$의 최소값으로 수렴한다는 것이다.

$$\lim_{l \to \infty} R_{emp}(\alpha) = R(\alpha)$$

$$\lim_{l \to \infty} \min R_{emp}(\alpha) = \min R(\alpha)$$

경험적 리스크 최소화 원칙은 학습집단을 대상으로 경험적 리스크를 최소화시킴으로써 미지의 확률분포를 추정하기 때문에 학습집단의 크기와 정확도에 영향을 받

게 된다. 또한 수렴의 경계범위를 나타내기 위해서 VC차원(Vapnikchernonenkis dimension)을 도입하였다[32].

b. VC 차원

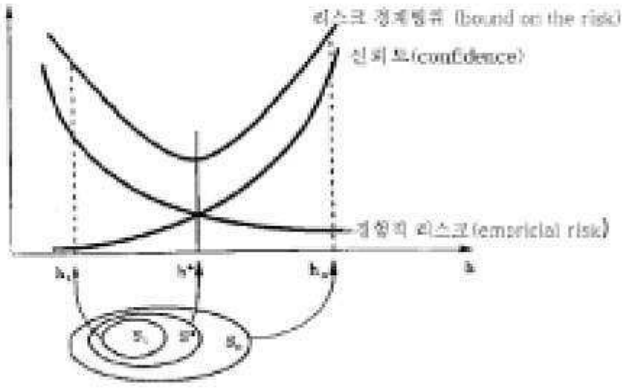
VC(Vapnik-Chernonenkis) 이론에서는 학습 집단을 적합한 용량(capacity)으로 학습할 수 있는 함수의 클래스에 대한 제한을 하는 것이 필수적이다. 이를 위해 경험적 리스크와 함수 클래스의 용량을 모두 고려한 검증 오류에 대한 경계범위(bounds)를 도입하였으며, 함수의 다양한 클래스를 측정하기 위해 VC차원을 이용하여 경계범위를 나타낸다[33].

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}}$$

여기서 h는 양수(non-negative)인 VC차원이고, l은 학습 데이터의 수, η는 신뢰도(confidence)를 의미하며 $0 \le \eta \le 1$의 범위를 갖는다. 위의 식에서 우변은 리스크의 경계범위(risk bound)를 나타내며, $\sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}}$ 는 VC 신뢰도를 나타낸다. 만약에 학습데이터의 주어진 양을 위해 자료가 복잡한 디자인을 하면, VC 신뢰도는 커지고 $R_{emp}(\alpha)$가 0까지 가서 적어진다 하더라도 검증집합(test set)에서의 오차(error)는 여전히 크다. 따라서 $R(\alpha)$을 최소화하기 위해서는 $R_{emp}(\alpha)$과 $\frac{h}{l}$ 를 동시에 줄임으로써, 우변의 항을 동시에 최저로 만드는 최적의 VC차원의 학습모형을 얻어야 한다[34][35].

c. 구조적 리스크 최소화 원칙*(SRM)*

기존의 학습 알고리즘은 학습 집단을 이용하여 경험적 오차(empirical error)를 최소화하는 경험적 리스크 최소화 원칙(ERM)을 구현하는 것을 목적으로 했다. 그러나 이러한 학습 알고리즘은 학습 집단의 크기와 정확도에 많은 영향을 받게 되어 수집할 수 있는 데이터의 범위가 큰 경우에는 적합하지 않다. Vapnik과 Chervonenkis에 의해 경험적 리스크 최소화의 원칙의 단점을 보완하면서 최적의 VC차원을 선택하기 위해 특정 데이터의 근사치에 대한 정확도(quality)와 경험적 리스크 함수의 복잡도(complexity) 사이의 타협점(trade-off)을 찾아주는 구조적 리스크 최소화의 원칙(Structural Risk Minimization : SRM)이 제안되었다. SRM은 전체 집단을 하위 집단으로 세분화 한 뒤 이 집단들에 대한 $R_{emp}(\alpha)$을 계산하여 이를 최소화하는 함수를 선택하는 것을 원칙으로 한다. 각 모형에 대한 경험적 리스크를 계산한 후 경계범위의 상한 값을 최소로 하는 모형을 선택하게 된다[36].



[그림 2.8] 리스크의 경계범위

[그림 2.8]은 리스크에 대한 경계범위를 보여 주고 있으며, 이때 경계범위는 경험적

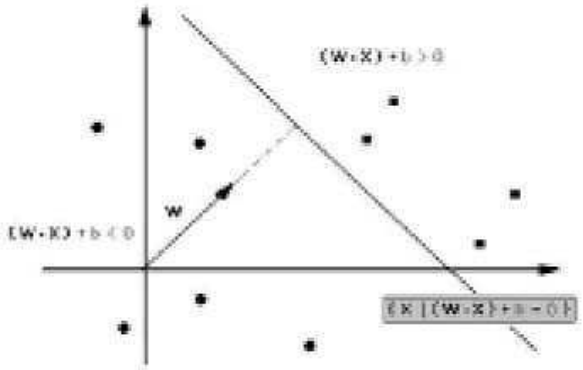
리스크와 신뢰도의 합으로 나타낸다. 즉, VC 차원이 증가함에 따라 경험적 리스크는 계속 감소하고 반대로 신뢰도는 증가하게 된다. 그러나 실제 오류는 최적의 VC 차원 지점까지 감소하다가 다시 증가하게 되는데, 이 최소 지점이 최적의 모형에 해당하는 VC 차원이 된다[37][38].

2. 최대 마진 분류기

[그림 2.9]에서 실선은 긍정예제와 부정예제를 분리하는 결정면을 의미하고, 실선과 평행인 점선들은 오류를 발생시키지 않으면서 결정면을 이동할 수 있는 여유한도를 나타내는 마진이다. 즉, SVM은 학습집단에서 마진을 최대화하는 결정면(optimal hyperplane)을 찾아내는 알고리즘이라 할 수 있다[39]. SVM은 Vapnik에 의해 처음 고안되었는데 [그림2.8]처럼 선형 분리가 가능한 두 클래스를 구분 지으며 마진(margin)을 최대로 하는 초평면 $w \cdot x + b = 0$을찾는 문제다. 여기서 w는 가중치 벡터이고, x는 입력벡터, b는 기준값이다. 학습 집합을 $D = \{(x_i, y_i)\}$라고 할 때, 입력데이터 x_i가 범주(class)에 속하면 y_i는 +1의 값을, 그렇지않으면 -1의 값을 갖는다. 결국 SVM은 다음을 만족하는 아래의 식에서 최적의 w와 b를 찾는 문제이다.

$x_i \cdot w + b \geq +1$($y_i = +1$인 경우)

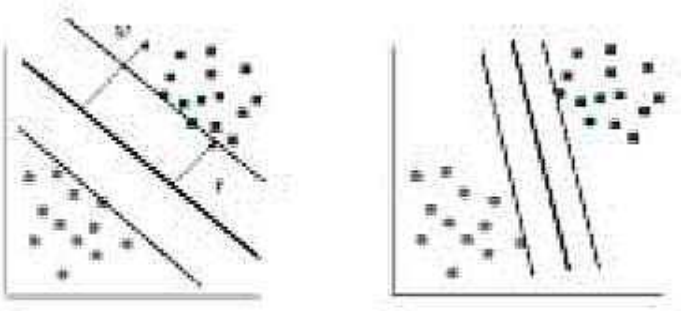
$x_i \cdot w + b \leq -1$($y_i = -1$인 경우)



[그림 2.9] 최적 경계면의 개념도

선형적으로 분리가 가능한 학습 집단에서는 최적의 초평면을 이용하여 최대 마진 분류기를 구축하며, 선형적으로 분리할 수 없는 경우에는 최적의 초평면 분류기 개념을 일반화시킨 소프트 마진 분류기를 구축하게 된다.

이원 분류를 위해서는 학습과정에서 $(x_1, y_1), \cdots, |(x_t, y_t) \in R^N \times \{\pm 1\}$ 을 만족시키는 함수 $f : R^N \rightarrow \{\pm 1\}$ 를 결정하여, 수많은 선형 분류기 중에서 일반화의 손실 없이 데이터를 분리해낼 수 있는 최대 마진을 갖는 최적의 선형 분류기를 선정해야만 한다. 이때, 일반화의 손실 없이 두 개의 범주를 분리해내는 선형 분류기라 할지라도 [그림 2.10]의 (b)와 같이 좁은 마진폭을 갖는 선형 분류기의 경우 예상 리스크가 높기 때문에 최대 마진 γ를 갖는 선형 분류기 (a)보다 낮은 일반화 성능을 보이게 된다.



(a) (b)

[그림 2.10] 최대 마진 분류기

그러므로 선형 SVM에서는 [그림 2.10]의 (a)와 같이 최대 마진을 찾기 위해서 $\|w\|$가 최소가 되는 최적의 값을 결정해야 하는데, SVM에서는 라그랑즈(Lagrange) 함수의 해를 구하는 문제로 최적화 문제를 해결하며, 이때 최적화 문제는 1차 영역(primal)의 해석과 2차 영역(dual)에서의 해석으로 나눌 수 있다. 1차 영역에서는 1차 변수 w와 b에 따라 최소화되어야 하며, 이를 라그랑즈 승수(Lagrange multipliers) α_i를 이용해서 표현하면 다음과 같다.

$$L(w, b, a) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} a_i\{[(x_i \cdot w) - b]y_i - 1\}, \ (a_i \geq 0, \forall_i)$$

이 식의 안장점(saddle point)을 찾기 위해서 다음과 같은 조건식을 구할 수 있다.

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = \sum_{i=1}^{l} \alpha_i^0 y_i = 0$$

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = w_0 - \sum_{i=1}^{l} y_i \alpha_i^0 x_i = 0$$

위의 결과를 이용해서 다음과 같은 식을 만들 수 있다.

$$W(a) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

최적 초평면을 설계하기 위해서는 이식을 최대화하는 α를 구하는데 이는 2차계획법(Quadratic Optimization)을 통해서 이루어진다. 이 과정을 통해서 최종적으로 얻

어진 해는 다음과 같다.

$$w_0 = \sum_{supportvectors} y_i \alpha^0_i x_i$$

이 해를 이용해서 얻어지는 분류법칙은 다음과 같다.

$$f(x) = sign(\sum_{supportvectors} y_i \alpha^0_i (x_i \cdot x) - b_0)$$

여기서 x_i를 지지벡터(support vector)라고 하는데 이것은 최적화 과정을 통해서 얻어진 α값 중에서 “0”이 아닌 값을 이용해서 결정되고 기준치 b_0는 다음 공식을 이용해서 구할 수 있다.

$$b_0 = \frac{1}{2}[(w-0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))]$$

여기서 $x^*(1)$는 긍정 범주의 지지벡터, $x^*(-1)$는 부정 범주의 지지벡터를 의미한다.

3. 비선형 SVM

입력데이터의 선형 분리가 불가능할 경우 입력공간을 분리하는 비선형 결정면(nonlinear decision surface)을 이용하게 된다. 함수 $K(\cdot,\cdot)$는 두 벡터의 내적을 나타내기 위한 커널 함수로 입력공간의 내적을 특징공간의 내적으로 변형시키는 역할을 한다. 이것은 특징공간으로 입력 벡터를 투영시킴으로써 내적에 대한 계산만을 필요로 하게 되어 고차원일 경우 계산상의 어려움을 덜 수 있다[40].

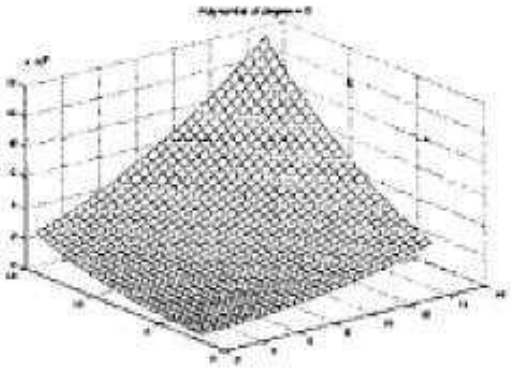
$$K(x, y) = (\Phi(x), \Phi(y))$$

즉, 벡터 x는 $\phi(\cdot)$를 통하여 특징 공간으로 변형시켜 두 벡터의 내적으로만 계산이 이루어지게 된다. 일반적으로 사용되는 커널함수는 다항식(polynomial) 커널함수, RBF(Radial Basis Function) 커널 함수, 다층 퍼셉트론(multi-layer perceptron) 커널함수 등이며, 본 논문에서는 아래의 커널 함수를 이용하여 실험을 수행하였다.

a. 다항식(Polynomial) 커널함수

다항식 커널함수는 스칼라 곱(Scalar product)으로 처리하기 때문에 두 벡터간의 방향에 의존적이며, 결국 동일한 방향을 갖는 벡터들은 다항식 커널 함수를 이용했을 때 높은 결과 값을 갖게 된다.

$K(x, x_i) = ((x \cdot x_i) + 1)^d$, d=다항식의차수



[그림 2.11] 다항식 커널함수

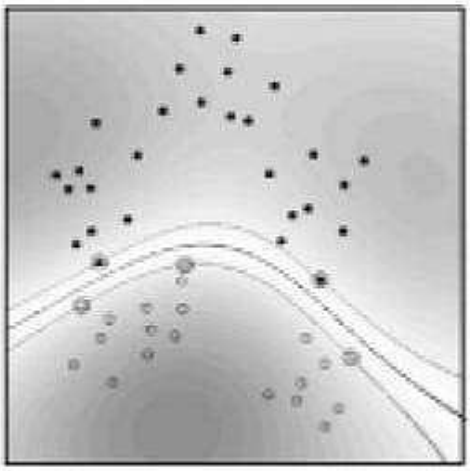
b. RBF 커널함수

$$f(x) = sign(\sum_{i=1}^{N} a_i K_r(|x - x_i|) - b)$$

여기서 $K_r(|x - x_i|)$은 두 벡터 사이의 거리인 $|x - x_i|$에 영향을 받게 되고, 가장 일반적인 함수 모형은 다음과 같다.

$$K(x, x_i) = \exp(\frac{\|x - x_i\|^2}{2\sigma^2})$$

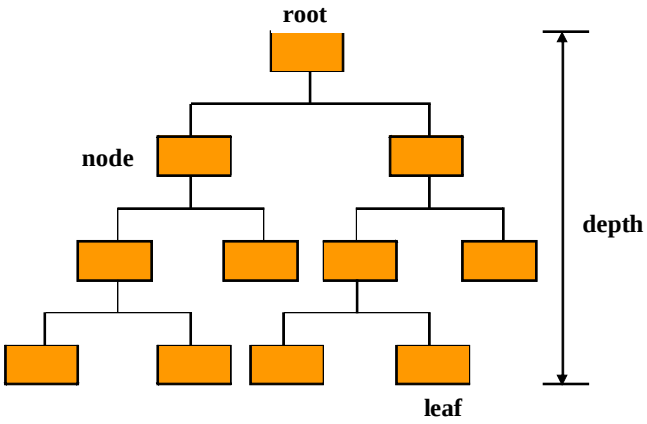
일반적으로 RBF 커널함수는 지지벡터와 검증 데이터간의 유클리드 거리와 연관되며, 여기서 지지벡터는 RBF의 중심점이 된다. σ값이 커질수록 완만한(smoother) 결정면을 갖게 된다.



[그림 2.12] RBF 커널함수를 이용한 SVM

C. 의사 결정 트리 (Decision Tree)

의사 결정 트리는 수집된 데이터의 레코드들을 분석하여 이들 집합의 분류를 위한 규칙을 생성해내어 트리 모양의 구조로 의사 결정의 집합을 나타낸다. 분류 작업에 주로 사용되는 기법으로, 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류 모형을 나타내는 것이다. 그리고 이렇게 만들어진 분류 모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용된다[41]. 이 기법은 분류나 예측의 근거를 알려주기 때문에 이해하기가 쉽고 변수간 상관관계와 영향을 알 수가 있어 데이터 선정이 용이하며 구조가 단순하여 모형 구축에 소요되는 시간이 짧다는 장점이 있다. 의사 결정 트리는 흔히 [그림 2.13] 과 같은 형태로 그 결과가 출력된다



[그림 2.13] 의사 결정 트리의 구조

[그림 2.13] 에서 사각형으로 표시된 것은 노드(Node)를 나타내는 것으로, 관측치들의 집합을 의미한다. 최상위에 위치하는 마디를 루트 노드(Root Node)라 하며

모든 관측치를 의미한다. 루트 마디를 특정 법칙에 의해 나누어 생성되는 마디들은 자식 노드(Child Node)라고 한다. 의사 결정 트리는 순서도와 같은 트리 구조이다. 안쪽 노드는 속성에 대한 검사표시이고 가지는 검사의 결과를 나타내며 리프 노드는 클래스 레이블이나 클래스 분포를 나타낸다. 의사 결정 트리는 두 가지 과정을 통해 생성 된다. 하나는 트리 생성으로 처음에 학습 데이터는 루트에 있고 선택된 속성을 기준으로 재귀적으로 분할한다. 다른 하나는 가지치기 과정으로 반사잡음이나 이상치의 가지를 식별하고 잘라낸다. 구축된 의사 결정 트리를 사용해서 트리에 대응하는 표본의 속성 값을 테스트함으로써 알려지지 않은 표본을 분류한다. 가장 보편적으로 사용 되는 것으로는 CART, CHAID, C4.5 등의 의사 결정 트리 알고리즘이 있으며, 본 논문에서는 C4.5을 이용하였다. C4.5 알고리즘은 J. Ross Quinlan에 의해 오랫동안 정립된 의사 결정 나무 알고리즘의 유용한 단편이다. C4.5 알고리즘은 기계학습알고리즘분야의 효력 있는 ID3알고리즘을 기반으로 한다. C4.5 알고리즘과 ID3알고리즘은 엔트로피(Entropy)를 최소화 하는 방향으로 가지치기를 수행하는 방법이다[42]. 다음 식은 엔트로피를 계산하는 방법이다.

$$Entropy(D) \equiv \sum_{i=1}^{c} - p_i \log p_i$$

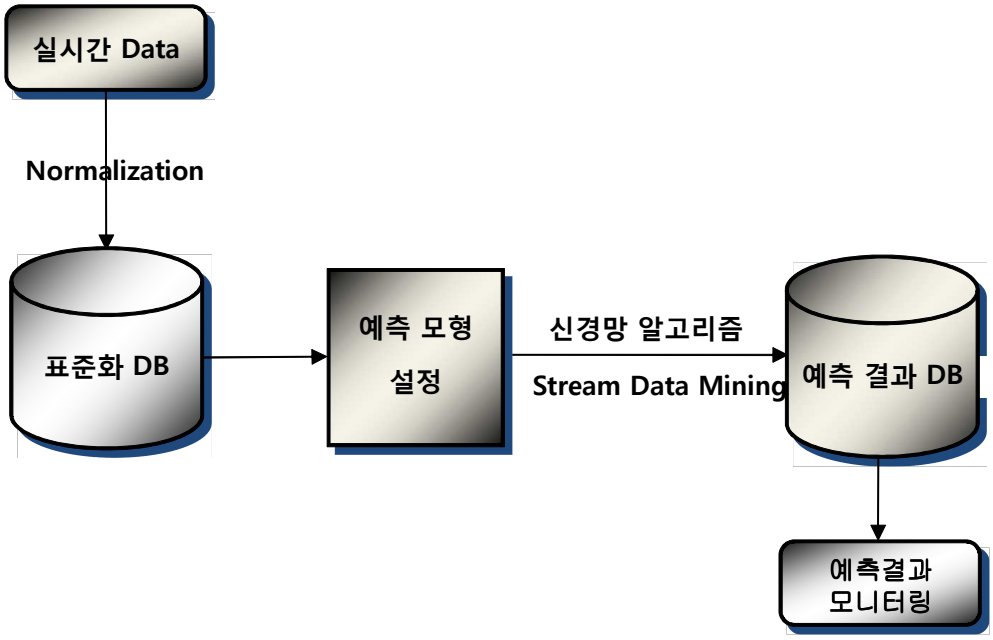
C4.5는 이진분리는 하는 CART와 달리 가지의 수를 다양화 할 수 있다. 또한 연속변수에 대해서는CART와 비슷한 방법을 사용하지만 범 주형 에서는 좀 다른 방법을 사용한다. 만약 "색깔"이 분리변수로 선택되면 트리의 다른 레벨은 각 색깔별로 노드를 형성한다. 의사 결정 트리의 장점은 이해하기 쉬운 규칙을 형성하고 많은 컴퓨팅 작업 없이 분류과정을 형성하며 연속변수와 범주형 변수에 모두 사용 가능 하다는 것이다. 또한 예측과 분류부분에서 가장 효과적인 방법이라는 것이다.

하지만 몇몇 의사 결정 트리 알고리즘이 이진분리를 하기 때문에 분리 가지의 수가 너무 많고 error-prone이 발생하는 단점이 있다. 또 트리 형성 시 컴퓨팅 비용이 많이 든다는 것이다. 가지치기 시, 분리기준 설정 시 각각의 경우를 다 고려해야 되므로 각 조합의 경우를 모두 고려할 경우 그리고 가지치기를 할 경우 컴퓨팅이 제공되어야 한다는 단점이 있다[43][44].

Ⅲ.시스템 구성 및 설계

A. 시스템 구성도

본 시스템의 전체 구성도는 [그림 3.1]과 같다. 본 논문에서는 KOSPI 200 지수와 개별종목 주가지수의 향후(내일) 상승 및 하락을 예측하기 위한 최적의 이원분류 모형을 구축하는 것이 주된 목적이다. 이를 위해서 입력변수로서 우리나라 주식시장의 흐름을 대표할 수 있는 기본 변수들을 선정하였고, 분석의 성능을 향상시키기 위해 신경망 알고리즘과 의사결정 트리 알고리즘을 이용하였다. 우선 기본 변수(24개)로는 주식지표인 KOSPI 지수, KOSPI 200지수, 미국지수인 DOW-JONES지수와 NASDAQ지수의 시가, 고가, 저가, 종가 지수와 각 지수들의 거래량과 거래대금, 경제지표인 회사채(3년 만기)와 원달러 환율을 선정하였고, 개별종목에 대한 주가지수를 예측 분류할 때에는 각 개별종목에 대한 주식지표(14개)가 더해져서 총 38개의 변수가 이용되었다. 개별종목 선정에 있어서는 KOSPI 200의 상위 20종목에서 베타계수가 큰 삼성전자, 포스코 2개 기업을 선정하였다.


실시간 Data
Normalization
표준화 DB
예측 모형
설정
신경망 알고리즘
Stream Data Mining
예측 결과 DB
예측결과
모니터링

[그림 3.1] 시스템 구성도

B. 표준화 및 변수 설정

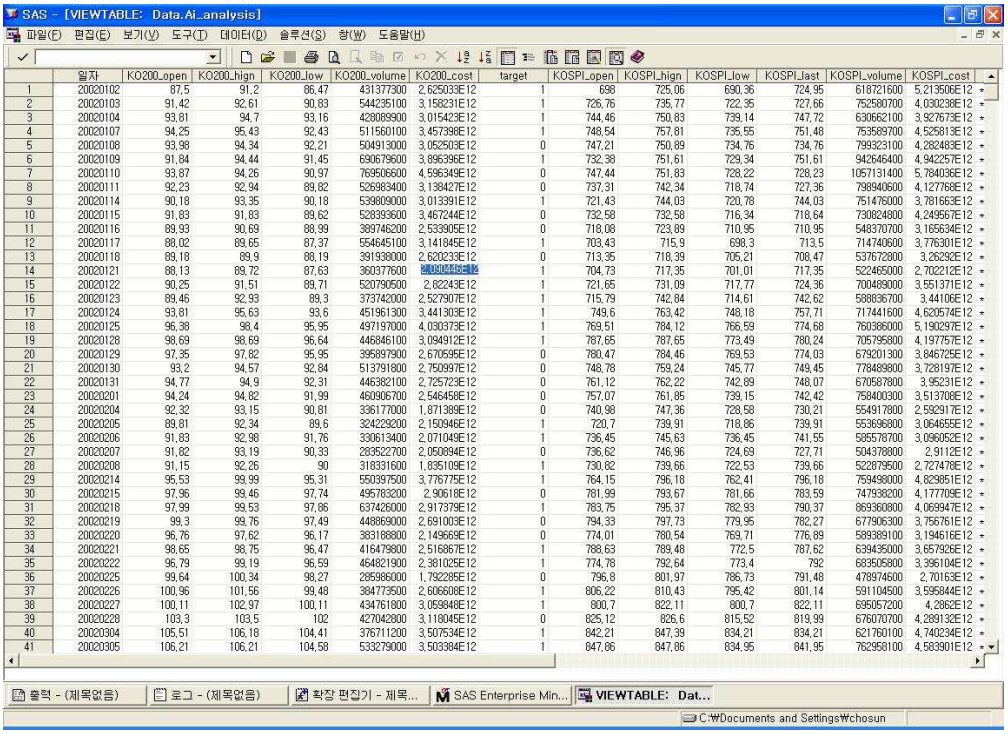
데이터 분석에서 Raw Data를 사용하게 되면 불규칙적으로 변동 폭이 심한 관찰치가 분석에 영향을 미치어 잘못된 결과를 초래할 수 있기 때문에 자료들을 그대로 사용하지 않고 전처리를 하는 것이 보통이다. 본 논문에서는 자료의 전처리를 위해 2단계 표준화 방법을 사용하였다. 여기서 사용된 2단계 표준화 방법은 우선 원자료를 표준화하여 표준화된 자료의 표준화 값이 4이상이면 3.5로, 3이상이면 3으로 코딩하여 다시 재 표준화하는 방법으로 원자료의 이상치(outlier)를 조정하였다. 분석에 들어가기에 앞서 위의 방법으로 사전 처리된 자료를 이용하여 변수를 선정하였다. 입력변수의 수는 분석모형의 예측정확도에 크게 영향을 끼치는 요소이다. 분석모형에서 변수의 수가 많아졌을 때 모형이 복잡해지고 모형구축에 많은 시간이 소요되며 과대적합(overfitting)으로 실제 예측하고자 하는 자료에 대한 예측율이 낮아지는 현상이 있으므로 학습속도와 과대적합에 많은 영향을 미치는 변수의 수를 조절하기 위해 변수 선정과정이 필수적이라 할 수 있다. 기본 변수(24개)로는 KOSPI 지수, KOSPI 200지수, 미국지수인 DOW-JONES지수와 NASDAQ지수의 시가, 고가, 저가, 종가 지수와 각 지수들의 거래량과 거래대금, 경제지표인 회사채(3년 만기)와 원달러 환율을 선정하였고, 개별종목에 대한 주가지수를 예측분류할 때에는 각 개별종목에 대한 주식지표(7개)가 더해져서 총 31개의 변수가 이용되었다. [표 3.1]은 모형 구축에 이용된 변수이다.

대분류	중분류	소분류
목표변수	Tarket변수	1 - 상승 0 - 하락
기본변수(24개)	KOSPI 지수	시가,고가,저가,종가,거래량, 거래대금
	KOSPI 200지수	시가,고가,저가,종가,거래량, 거래대금
	DOW 지수	시가,고가,저가,종가,거래량
	NASDAQ 지수	시가,고가,저가,종가,거래량
	경제지표	회사채(3년만기), 원달러환율
개별주식변수(7개)	개별주식지표	시가,고가,저가,종가,거래량,거래대금,시가총액

[표 3.1] 모형 구축에 이용된 변수 선정

C. 모형 설계

모형을 설계하기 위하여 데이터를 표준화하여 원자료의 이상치(outlier)를 조정하였다. 분석에 들어가기에 앞서 위의 방법으로 사전 처리된 자료를 이용하여 변수를 선정하였다. [그림 3.2]는 모형 설계에 사용된 데이터 자료이다.



SAS - [VIEWTABLE: Data.Ai_analysis]

	일자	KO200_open	KO200_hign	KO200_low	KO200_volume	KO200_cost	target	KOSPI_open	KOSPI_hign	KOSPI_low	KOSPI_last	KOSPI_volume	KOSPI_cost
1	20020102	87,5	91,2	86,47	431377300	2,625033E12	1	698	725,06	690,36	724,95	618721600	5,213506E12
2	20020103	91,42	92,61	90,83	544235100	3,158231E12	1	726,76	735,77	722,35	727,66	752580700	4,030238E12
3	20020104	93,81	94,7	93,16	428089900	3,015423E12	1	744,46	750,83	739,14	747,72	630662100	3,927673E12
4	20020107	94,25	95,43	92,43	511560100	3,457398E12	1	748,54	757,81	735,55	751,48	753589700	4,525813E12
5	20020108	93,98	94,34	92,21	504913000	3,052503E12	0	747,21	750,89	734,76	734,76	799323100	4,282483E12
6	20020109	91,84	94,44	91,45	690679600	3,896396E12	1	732,38	751,61	729,34	751,61	942646400	4,942257E12
7	20020110	93,87	94,26	90,97	769506600	4,596349E12	0	747,44	751,83	728,22	728,23	1057131400	5,784036E12
8	20020111	92,23	92,94	89,82	526983400	3,138427E12	0	737,31	742,34	718,74	727,36	798940600	4,127768E12
9	20020114	90,18	93,35	90,18	539809000	3,013391E12	1	721,43	744,03	720,78	744,03	751476000	3,781663E12
10	20020115	91,83	91,83	89,62	528393600	3,467244E12	0	732,58	732,58	716,34	718,64	730824800	4,249567E12
11	20020116	89,93	90,69	88,99	389746200	2,533905E12	0	718,08	723,89	710,95	710,95	548370700	3,165634E12
12	20020117	88,02	89,65	87,37	554645100	3,141845E12	1	703,43	715,9	698,3	713,5	714740600	3,776301E12
13	20020118	89,18	89,9	88,19	391938000	2,620233E12	0	713,35	718,39	705,21	708,47	537672800	3,26292E12
14	20020121	88,13	89,72	87,63	360377600	3,030444E12	1	704,73	717,35	701,01	717,35	522465000	2,702212E12
15	20020122	90,25	91,51	89,71	520790500	2,82243E12	1	721,65	731,09	717,77	724,36	700489000	3,551371E12
16	20020123	89,46	92,93	89,3	373742000	2,527907E12	1	715,79	742,84	714,61	742,62	588836700	3,44106E12
17	20020124	93,81	95,63	93,6	451961300	3,441303E12	1	749,6	763,42	748,18	757,71	717441600	4,620574E12
18	20020125	96,38	98,4	95,95	497197000	4,030373E12	1	769,51	784,12	766,59	774,68	760386000	5,190297E12
19	20020128	98,69	98,69	96,64	446846100	3,094912E12	1	787,65	787,65	773,49	780,24	705795800	4,197757E12
20	20020129	97,35	97,82	95,95	395897900	2,670595E12	0	780,47	784,46	769,53	774,03	679201300	3,846725E12
21	20020130	93,2	94,57	92,84	513791800	2,750997E12	0	748,78	759,24	745,77	749,45	778489800	3,728197E12
22	20020131	94,77	94,9	92,31	446382100	2,725723E12	0	761,12	762,22	742,89	748,07	670587800	3,95231E12
23	20020201	94,24	94,82	91,99	460906700	2,546458E12	0	757,07	761,85	739,15	742,42	758400300	3,513708E12
24	20020204	92,32	93,15	90,81	336177000	1,871389E12	0	740,98	747,36	728,58	730,21	554917800	2,592917E12
25	20020205	89,81	92,34	89,6	324229200	2,150946E12	1	720,7	739,91	718,86	739,91	553696800	3,064655E12
26	20020206	91,83	92,98	91,76	330613400	2,071049E12	1	736,45	745,63	736,45	741,55	585578700	3,096052E12
27	20020207	91,82	93,19	90,33	283522700	2,050894E12	0	736,62	746,96	724,69	727,71	504378800	2,9112E12
28	20020208	91,15	92,26	90	318331600	1,835109E12	1	730,82	739,66	722,53	739,66	522879500	2,727478E12
29	20020214	95,53	99,99	95,31	550397500	3,776775E12	1	764,15	796,18	762,41	796,18	759498000	4,829851E12
30	20020215	97,96	99,46	97,74	495783200	2,90618E12	0	781,99	793,67	781,66	783,59	747938200	4,177709E12
31	20020218	97,99	99,53	97,86	637426000	2,917379E12	1	783,75	795,37	782,93	790,37	869360800	4,069947E12
32	20020219	99,3	99,76	97,49	448869000	2,691003E12	0	794,33	797,73	779,95	782,27	677906300	3,756761E12
33	20020220	96,76	97,62	96,17	383188800	2,149669E12	0	774,01	780,54	769,71	776,89	589389100	3,194616E12
34	20020221	98,65	98,75	96,47	416479800	2,516867E12	1	788,63	789,48	772,5	787,62	639435000	3,657926E12
35	20020222	96,79	99,19	96,59	464821900	2,381025E12	1	774,78	792,64	773,4	792	683505800	3,396104E12
36	20020225	99,64	100,34	98,27	285986000	1,792285E12	0	796,8	801,97	786,73	791,48	478974600	2,70163E12
37	20020226	100,96	101,56	99,48	384773500	2,606608E12	1	806,22	810,43	795,42	801,14	591104500	3,595844E12
38	20020227	100,11	102,97	100,11	434761800	3,059848E12	1	800,7	822,11	800,7	822,11	695057200	4,2862E12
39	20020228	103,3	103,5	102	427042800	3,118045E12	0	825,12	826,6	815,52	819,99	676070700	4,289132E12
40	20020304	105,51	106,18	104,41	376711200	3,507534E12	1	842,21	847,39	834,21	834,21	621760100	4,740234E12
41	20020305	106,21	106,21	104,58	533279000	3,503384E12	1	847,86	847,86	834,95	841,95	762958100	4,583901E12

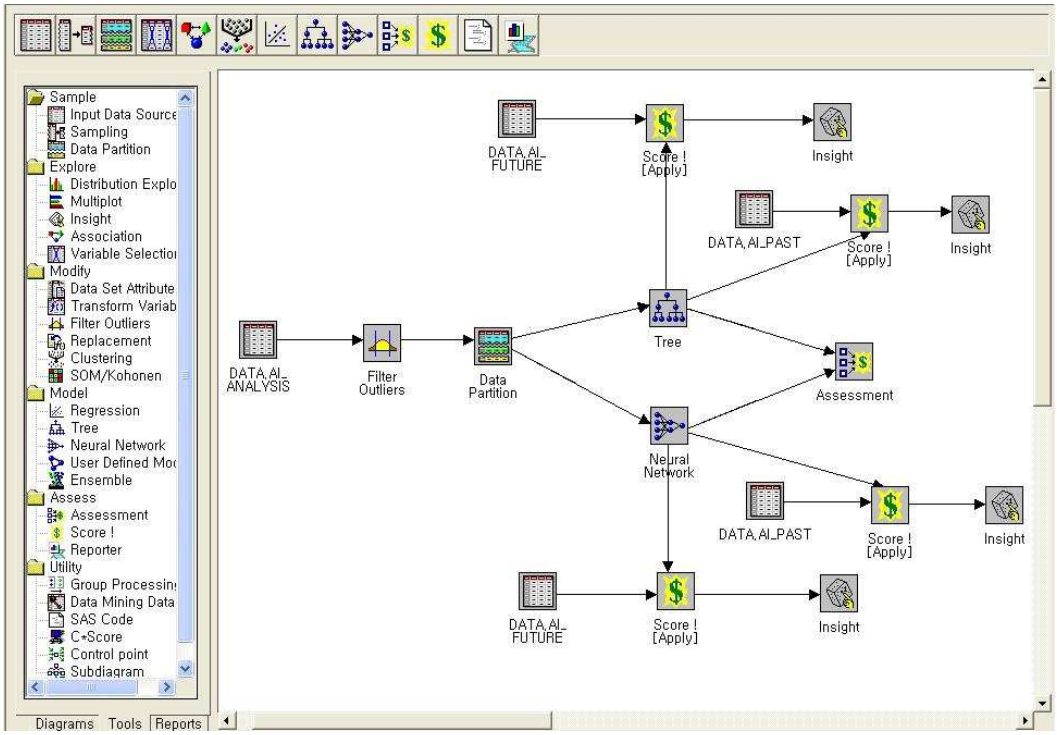
[그림 3.2] 모형 설계에 사용된 데이터

모형을 구축하는 과정에서 자료를 분할할 때에는 분석용, 평가용, 검증용 3단계로 분할하였다. [표 3.2]은 자료의 분할을 나타낸다.

[표 3.2] 자료의 분할

Method		분석용	평가용	검증용
신경망	(3 partition)	40%	30%	30%
의사결정트리	(3 Partition)	40%	30%	30%

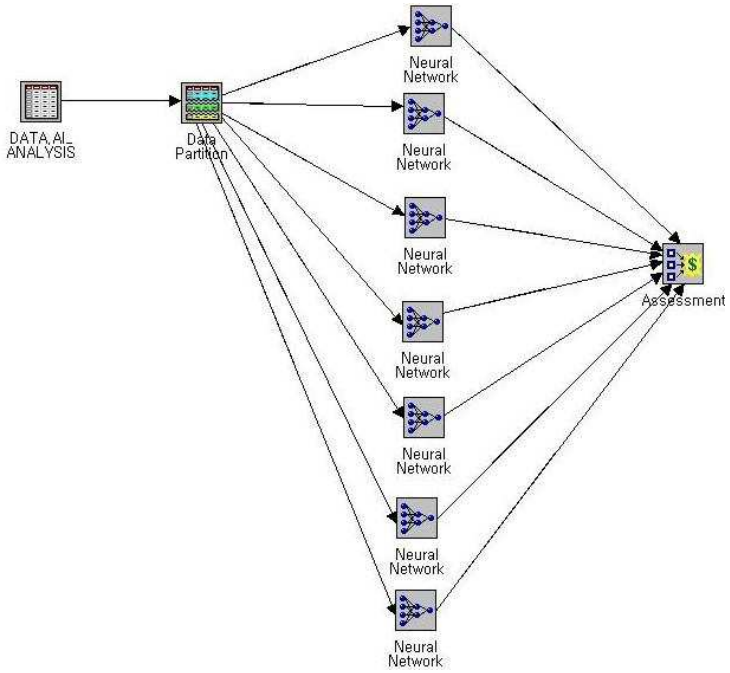
분할된 자료를 통한 두 알고리즘의 분석 과정은 다음과 같다. [그림 3.3]는 두 알고리즘의 도식도를 나타낸다.



[그림 3.3] 신경망과 의사결정트리 알고리즘의 도식도

1. 신경망을 이용한 모형 설계

분석에 사용된 신경망 모형은 지도학습을 이용한 예측 모형에서 가장 보편적으로 사용되고 있는 MLP와 역전파 알고리즘을 이용하였다. 하나의 은닉층이 있는 3 layer 모형으로 은닉층의 노드(node)수는 1개에서 7개까지 변화시켜가면서 학습하였고 오차가 가장 적은 최적 은닉노드 수를 발견하는 방법을 사용하였다. 은닉층의 노드 수는 입력변수의 수에 의해 결정된다. 신경망 모형의 구조를 결정할 때 가장 중요한 것은 바로 은닉층의 노드 수를 결정하는 것이라 할 수 있다[45]. 은닉층의 노드 수에 따라 신경망 모형의 예측력에 차이가 생기기 때문이다.



[그림 3.4] 신경망 모형의 최적화 은닉층 노드수 결정

은닉층의 노드 수를 결정하기 위한 방법은 아직까지 정형화되지 않았는데, 일반적으로 은닉 노드의 수가 증가하면 신경망 모형의 예측력이 향상되는 것으로 알려지고 있다. [그림 3.4]는 신경망의 최적화된 은닉층의 노드수를 결정하기 위한 과정이다. 입력받은 모든 변수는 선형 함수(linear function)에 의해 결합되고 이렇게 선형 결합된 값을 비선형으로 변환하는 과정을 거치게 되는데 이때 활성함수로 로지스틱(logistic)함수와 쌍곡 탄젠트(hyperbolic tangent)함수를 사용하였고, 모형 선택의 기준으로 Misclassification Rate와 Average Error를 사용하여 신경망 모형의 성능을 향상시키려 하였다. [표 3.3]는 신경망 모형의 설계를 나타낸다.

[표 3.3] 신경망 모형 설계

Model Selection Criteria	Combination Fn.	Activation Fn.	No. Hidden Node
Misclassification Rate Average Error	Linear	Hyperbolic Tangent Logistic	1 ~ 7

이 모형에서는 최적의 은닉층과 은닉노드의 수가 1개로 결정되었고 데이터에 반복된 학습을 통해 연결강도를 추정한 결과 [그림 3.5]와 같은 결과를 얻었다.

Results - Neural Network

Model | Tables Weights | Plot | Code | Log | Output | Notes |

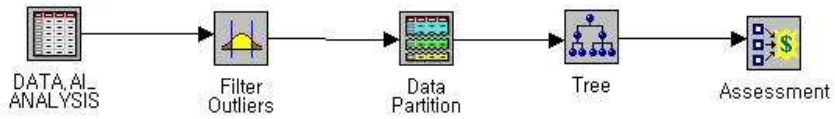
	From	To	Weight
1	DOW_HIGN	H11	1,6676967238
2	DOW_LAST	H11	-0,085513339
3	DOW_LOW	H11	0,6690176727
4	DOW_OPEN	H11	-2,182502776
5	DOW_VOLUME	H11	0,0210854514
6	KO200_COST	H11	-0,051569378
7	KO200_HIGN	H11	1,6806764337
8	KO200_LOW	H11	0,9730183347
9	KO200_OPEN	H11	-2,543112546
10	KO200_VOLUME	H11	0,0865114959
11	KOSPI_COST	H11	0,1267478802
12	KOSPI_HIGN	H11	-6,353140007
13	KOSPI_LAST	H11	13,97644871
14	KOSPI_LOW	H11	-1,564011505
15	KOSPI_OPEN	H11	-6,226278263
16	KOSPI_VOLUME	H11	-0,109744681
17	NAS_HIGN	H11	-0,190799264
18	NAS_LAST	H11	2,3421454055
19	NAS_LOW	H11	0,4894004631
20	NAS_OPEN	H11	-2,70835624
21	NAS_VOLUME	H11	-0,031402257
22	BIAS	H11	0,1481572534
23	H11	TARGET1	11,692439445
24	BIAS	TARGET1	-1,443724751

[그림 3.5] 신경망 모형의 추정된 연결강도

KOSPI200지수의 신경망 모형의 추정된 연결강도를 보면 NASDAQ 종가가 2.34로 가장 높게 나타났고 KOSPI200의 고가, DOW지수의 고가 순으로 높게 나타남으로써 모형에 가장 많은 영향을 미치는 변수는 KOSPI200의 고가, DOW지수의 고가인 것으로 나타났다.

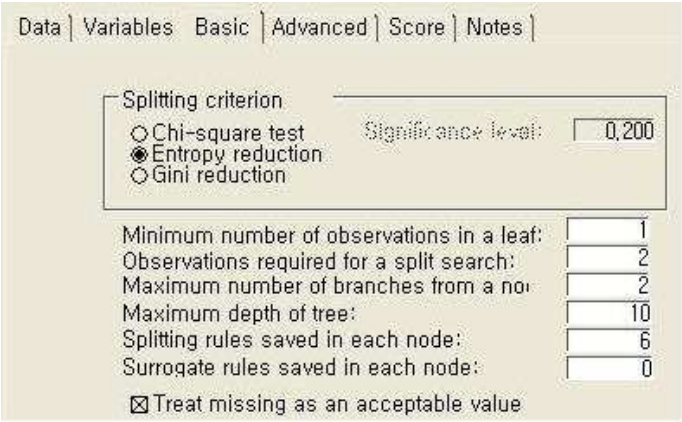
2. 의사결정트리를 이용한 모형 설계

의사결정트리 모형을 설계하기 위해서 먼저 분석흐름도를 작성하고 Model Role을 설정하였다. KOSPI200지수 패턴을 분석하고자 목표 변수를 KOSPI 200지수의 시가로 하였다. [그림 3.6]은 의사결정트리 모형의 분석흐름도이다.



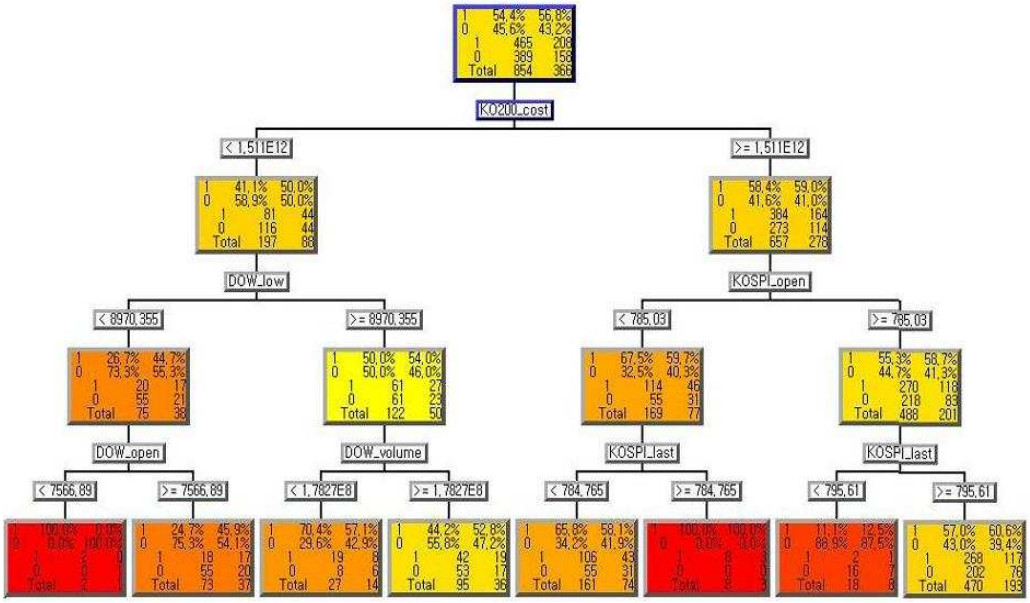
[그림 3.6] 의사결정트리의 분석흐름도

변수들의 관측치 중에 이상치를 제거하기 위해 분석흐름도에 Fillter Outlier를 추가시켜서 분석하였다. Tree 노드에서는 부모마디로부터 자식마디를 분리하기 위해서 마디에서 불순도를 재는 측도인 엔트로피 지수E(Entropy index)를 기준으로 부모마디에서 자식마디로 분리하는 Entropy reduction 방법을 이용하였다[46][47]. [그림 3.7]은 의사결정트리 모형의 분리 기준인 Entropy reduction 방법의 실행화면이다.



[그림 3.7] 의사결정트리 모형의 분리 기준

[그림 3.8]은 KOSPI200지수를 목표변수로 설정하였을 때 의사결정트리 기법을 이용한 모형 설계의 결과이다. 모형 설계를 한 후 목표 변수에 가장 영향을 미치는 속성은 시가총액, 시가, 종가 순임을 알 수 있다.

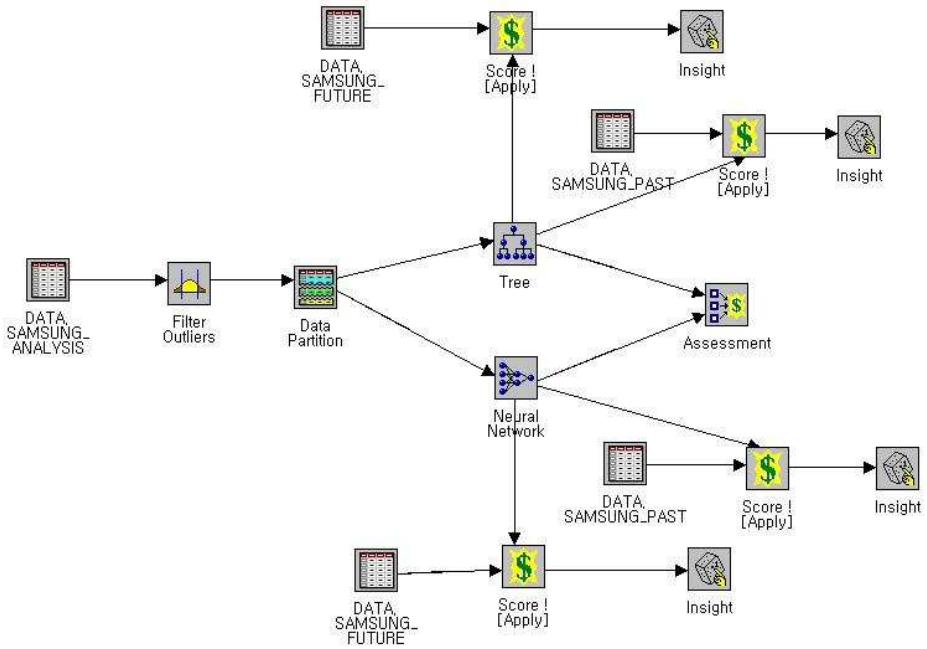


[그림 3.8] 의사결정트리 모형 설계 결과(KOSPI200지수)

3. 개별주식에 대한 모형 설계

a. 삼성전자

개별 주식에 대한 모형을 구축하기 위해서 먼저 분석흐름도를 작성하고 Model Role을 설정하였다. 삼성전자 주가의 패턴을 분석하고자 목표 변수를 삼성전자 주가의 시가로 하였다. 삼성전자 주가의 예측을 위해 기본변수(24개)에 개별주가지표(7개) 시가,고가,저가,종가,거래량,거래대금,시가총액을 이용하여 분석하였다. 사용된 알고리즘의 분석 과정은 다음과 같다. [그림 3.9]는 알고리즘의 모형도를 나타낸다.



[그림 3.9] 알고리즘의 모형도(삼성전자)

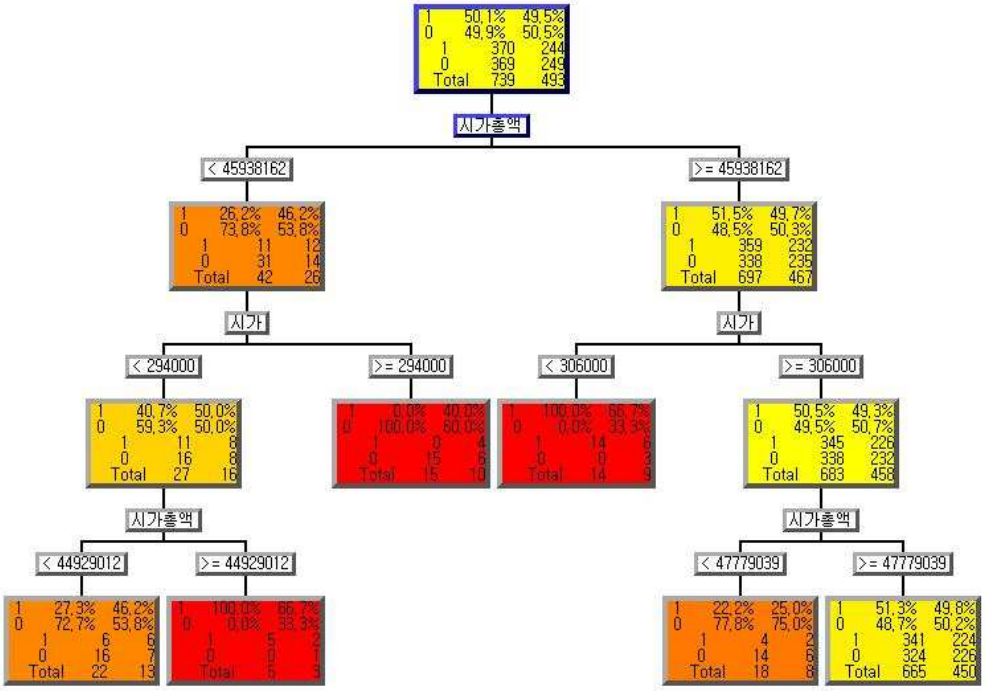
신경망 모형에서는 최적의 은닉층과 은닉노드의 수가 1개로 결정되었고 데이터에 반복된 학습을 통해 연결강도를 추정한 결과 [그림 3.10]과 같은 결과를 얻었다.

	From	To	Weight
1	_COL1	H11	-5.510651409
2	_COL2	H11	0.6312081496
3	_COL3	H11	2.4992321694
4	_COL6	H11	2.3560958732
5	_COL7	H11	-0.034932428
6	_COL8	H11	0.0393280855
7	BIAS	H11	-0.371652416
8	H11	_COL51	9.0227163696
9	BIAS	_COL51	3.0852246964

[그림 3.10] 신경망 모형의 추정된 연결강도(삼성전자)

삼성전자 주가의 신경망 모형의 추정된 연결강도를 보면 전날의 저가가 2.50으로 가장 높게 나타났고 전날의 고가가 2.35로 거의 차이가 없이 가장 영향을 미치는 것으로 나타났다.

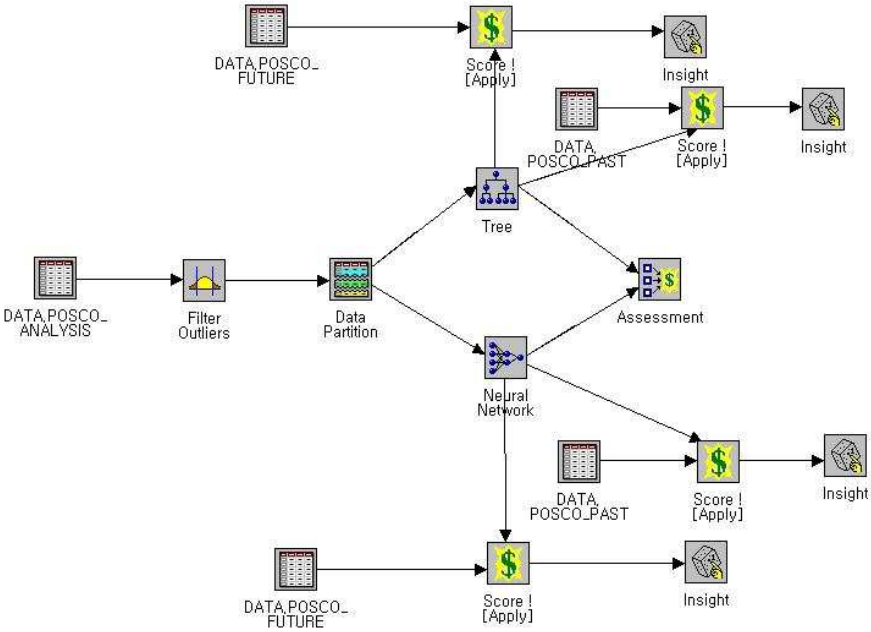
[그림 3.11]은 삼성전자 주식지표를 목표변수로 설정하였을 때 의사결정트리 기법을 이용한 모형 설계의 결과이다. 모형 설계를 한 후 목표 변수에 가장 영향을 미치는 속성은 시가총액, 시가 순임을 알 수 있다. 또한 마디의 구분값을 나타낸다.



[그림 3.11] 의사결정트리 모형 설계 결과(삼성전자)

b. 포스코

포스코에 대한 모형을 구축하기 위해서 먼저 분석흐름도를 작성하고 Model Role을 설정하였다. 포스코 주가의 패턴을 분석하고자 목표 변수를 포스코 주가의 시가로 하였다. 포스코 주가의 예측을 위해 삼성전자와 마찬가지로 기본변수(24개)에 개별주가지표(7개) 시가,고가,저가,종가,거래량,거래대금,시가총액을 이용하여 분석하였다. 사용된 알고리즘의 분석 과정은 다음과 같다. [그림 3.12]는 알고리즘의 모형도를 나타낸다.



[그림 3.12] 알고리즘의 모형도(포스코)

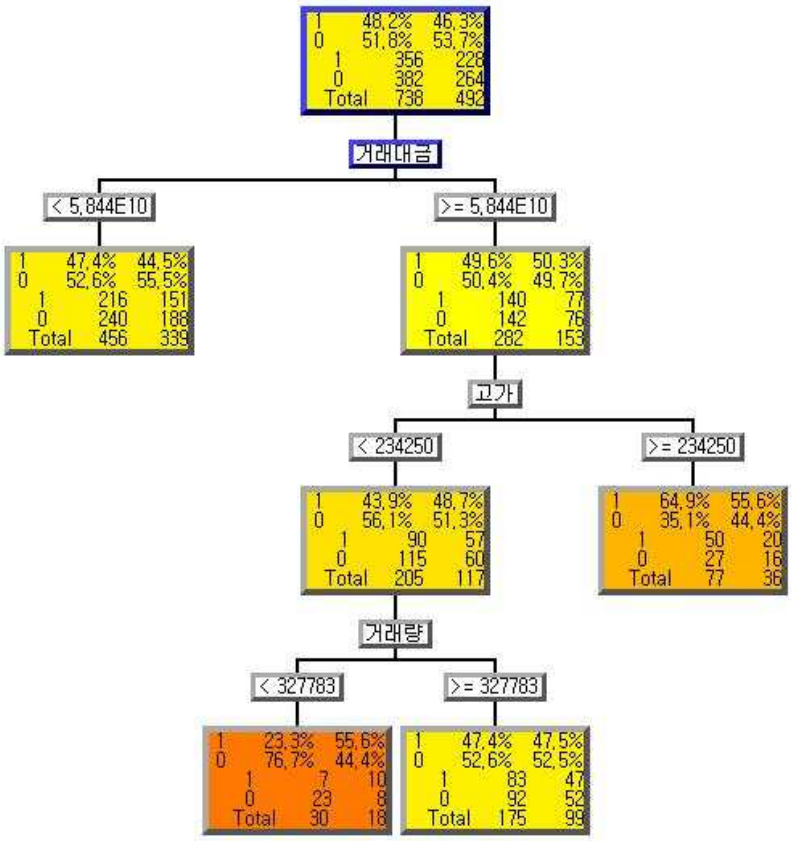
신경망 모형에서는 최적의 은닉층과 은닉노드의 수가 1개로 결정되었고 데이터에 반복된 학습을 통해 연결강도를 추정한 결과 [그림 3.13]과 같은 결과를 얻었다.

	From	To	Weight
1	_COL1	H11	11.181801641
2	_COL2	H11	-2.619539291
3	_COL3	H11	-5.776688174
4	_COL6	H11	-2.735017142
5	_COL7	H11	0.115748425
6	_COL8	H11	-0.176942235
7	BIAS	H11	-0.177080962
8	H11	_COL51	-3.63859082
9	BIAS	_COL51	-0.680425811

[그림 3.13] 신경망 모형의 추정된 연결강도(포스코)

포스코 주가의 신경망 모형의 추정된 연결강도를 보면 전날의 시가가 11.18로 가장 높게 나타남으로써 가장 많은 영향을 미치는 것으로 나타났다.

[그림 3.14]는 포스코 주식지표를 목표변수로 설정하였을 때 의사결정트리 기법을 이용한 모형 설계의 결과이다. 모형 설계를 한 후 목표 변수에 가장 영향을 미치는 속성은 거래대금, 고가, 거래량 순임을 알 수 있다. 또한 마디의 구분값을 나타낸다.



[그림 3.14] 의사결정트리 모형 설계 결과(포스코)

Ⅳ. 실행결과 및 평가

A. KOSPI 200 지수를 이용한 경우

1. 신경망을 이용한 경우

표준화된 데이터를 이용해서 신경망 기법의 실험을 수행하였다. 분석에 사용된 신경망 모형은 지도학습을 이용한 예측 모형에서 가장 보편적으로 사용되고 있는 MLP와 역전파 알고리즘을 이용하였다. 하나의 은닉층이 있는 3 layer 모형으로 은닉층의 노드(node)수는 1개에서 7개까지 변화시켜가면서 학습하였고 오차가 가장 적은 최적 은닉노드 수를 발견하는 방법을 사용하였다. 신경망 실험 예측 결과를 보면, 활성함수에 따른 차이, 모형설정기준에 따른 차이는 거의 없다 할 수 있겠고 가장 좋은 예측력을 보인 신경망 모형은 은닉노드수가 1개인 경우로 분석용 집단 73.2%, 평가용 집단 71.1%, 검증용 집단 72.8%라는 최적의 모형을 찾을 수 있었다.

[표 4.1] 신경망의 실험 결과(KOSPI200지수)

분석용			평가용			검증용		
	실제 하락	실제 상승		실제 하락	실제 상승		실제 하락	실제 상승
하락(예측)	136	52	하락(예측)	99	48	하락(예측)	100	33
상승(예측)	55	158	상승(예측)	46	135	상승(예측)	48	118
Total (%)	73.2		Total (%)	71.1		Total (%)	72.8	

[표 4.1] 신경망 모형의 결과를 보면 집단별로 실제하락을 예측한 경우보다는 실제 상승을 예측하는 경우가 높아 이 모형을 실제 주식시장에 적용하여 예측한다면 하락을 예측하는 경우에는 위험성이 커져 문제가 될 수 있을 것이다. 의사결정트리를 이용한 예측 시스템을 위해서 총 1028개의 학습예제를 가지고 의사결정 트리를 구성하였으며 생성된 의사결정트리를 규칙기반(Rule Base)으로 만들어서 실험하였다. 실험 결과는 분석용 집단 81.6%, 평가용 집단 83.2%, 검증용 집단 82.9%라는 최적의 모형을 찾을 수 있었다. [표 4.2]는 의사결정트리의 실험 결과이다.

[표 4.2] 의사결정트리의 실험 결과

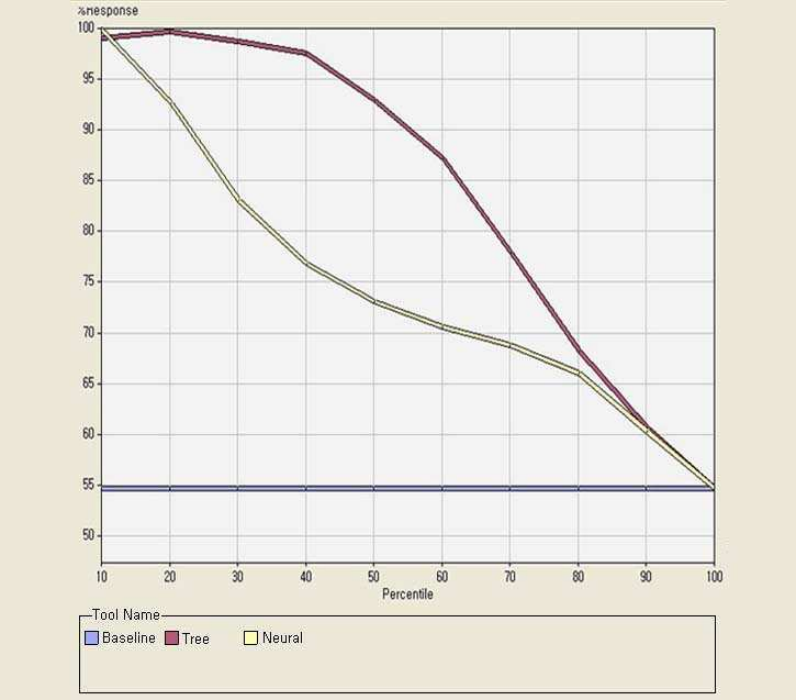
분석용			평가용			검증용		
	실제 하락	실제 상승		실제 하락	실제 상승		실제 하락	실제 상승
하락(예측)	149	32	하락(예측)	116	26	하락(예측)	118	21
상승(예측)	42	178	상승(예측)	29	157	상승(예측)	30	130
Total (%)	81.6		Total (%)	83.2		Total (%)	82.9	

KOSPI200 지수의 예측 결과를 종합해 보면, 신경망 기법은 71.1%의 예측력을 지닌 것으로 나타났다. 하지만 의사결정트리를 이용해서 데이터마이닝을 한 결과 예측력이 83.2%로 약 10%이상의 높은 예측력을 보여줌을 알 수 있다. 이러한 결과가 도출될 수 있었던 것은 기존의 연구 방법과는 다르게 실시간으로 들어오는 데이터에 대한 분석을 통해서 예측력을 높일 수 있음을 알 수 있다.

[표 4.3] KOSPI 200지수를 이용한 각 기법들의 예측 성능비교 (단위:%)

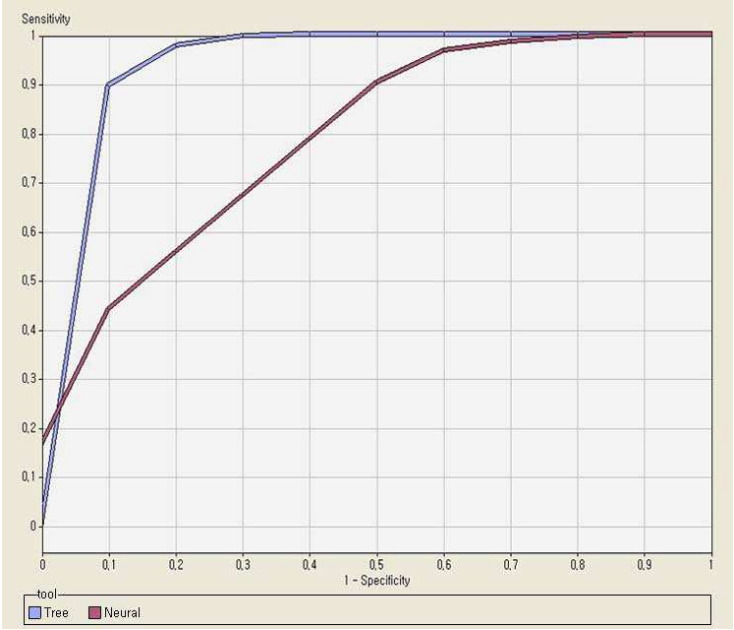
	분석용	평가용	검증용
신경망	73.2	71.1	72.8
의사결정트리	81.6	83.2	82.9

모형화를 통하여 얻어진 결과의 성능을 평가하기 위하여 Lift Chart와 ROC Curve를 이용하여 성능을 평가하고 신경망과 의사결정트리중에 좋은 성능을 보이는 모델을 산출하였다. Lift Chart는 %Response이익도표로서 목표 변수의 특정 범주의 빈도를 해당집단에서 전체빈도로 나눈값을 도표로 나타낸 것이다. 즉, %Response = (해당집단에서 목표변수의 특정범주의 빈도/해당 집단에서의 전체빈도) × 100을 이용하여 모델의 성능을 평가하는데 %Response값이 높을수록 모형의 성능이 좋다고 평가한다[48]. [그림 4.1]은 신경망 모형과 의사결정트리 모형의 Lift Chart를 나타낸다.



[그림 4.1] Lift Chart(KOSPI200지수)

[그림 4.1]을 보면 베이스 라인을 기준으로 의사결정트리 모형이 신경망 모형보다 예측력이 우수하다는 것을 알 수 있다. ROC Curve는 이진형 목표 변수를 가지는 모형들의 성능을 비교, 평가 하는데 매우 유용한 도표로서 민감도(Sensitivity) 와 특이도(Specificity)를 이용하여 X축에는 1-특이도를 Y축에는 민감도를 각 분류 기준값에 대해 나타낸 도표이다[49]. 그래프가 도표의 왼쪽 상단에 가까울수록 성능면에서 우수한 모형으로 판단된다. [그림 4.2]는 KOSPI200지수에 대한 ROC Curve 이다.



[그림 4.2] ROC Curve(KOSPI200지수)

위의 ROC Curve를 통해 모형을 비교하면 의사결정트리 모형이 신경망 모형보다 우수함을 알 수 있다.

B. 개별 종목 주가를 이용한 경우

개별종목 주가의 상승 또는 하락을 예측하기 위해서 KOSPI 200지수에서 구축한 모형과 동일한 모형을 설계하고 위에서 언급한 방법으로 변수를 선정하여 신경망, 의사결정트리 알고리즘을 이용하여 분석을 하였다.

1. 삼성전자

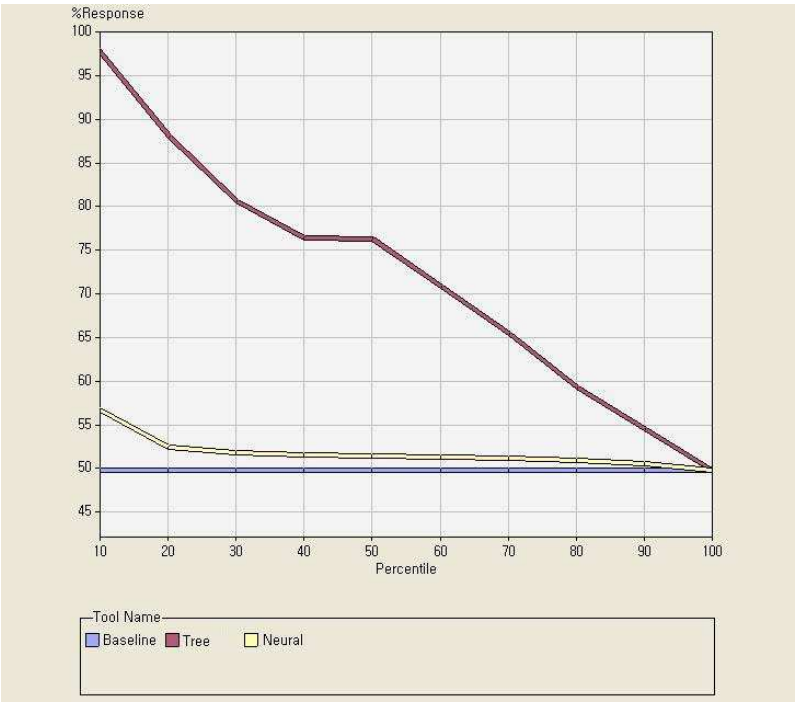
신경망이나 의사결정트리를 이용한 삼성전자 주식의 예측 결과는 KOSPI 200지수의 예측 결과와 마찬가지로 의사결정트리를 이용한 예측 결과(75.2%)가 신경망 모형의 예측결과(67.4%)보다 높았다. 신경망 모형에서 가장 좋은 예측력을 보인 모형은 은닉노드수가 1개인 경우로 분석용 집단 70.3%, 평가용 집단 64.1%, 검증용 집단 67.4%라는 최적의 모형을 찾을 수 있었다. 의사결정트리를 이용한 실험 결과 분석용 집단 71.3%, 평가용 집단 68.5%, 검증용 집단 75.2% 이었다. [표 4.4]는 삼성전자 지수를 이용한 각 기법들의 예측 성능비교를 나타낸다.

[표 4.4] 삼성전자 지수를 이용한 각 기법들의 예측 성능비교 (단위:%)

	분석용	평가용	검증용
신경망	70.3	64.1	67.4
의사결정트리	71.3	68.5	75.2

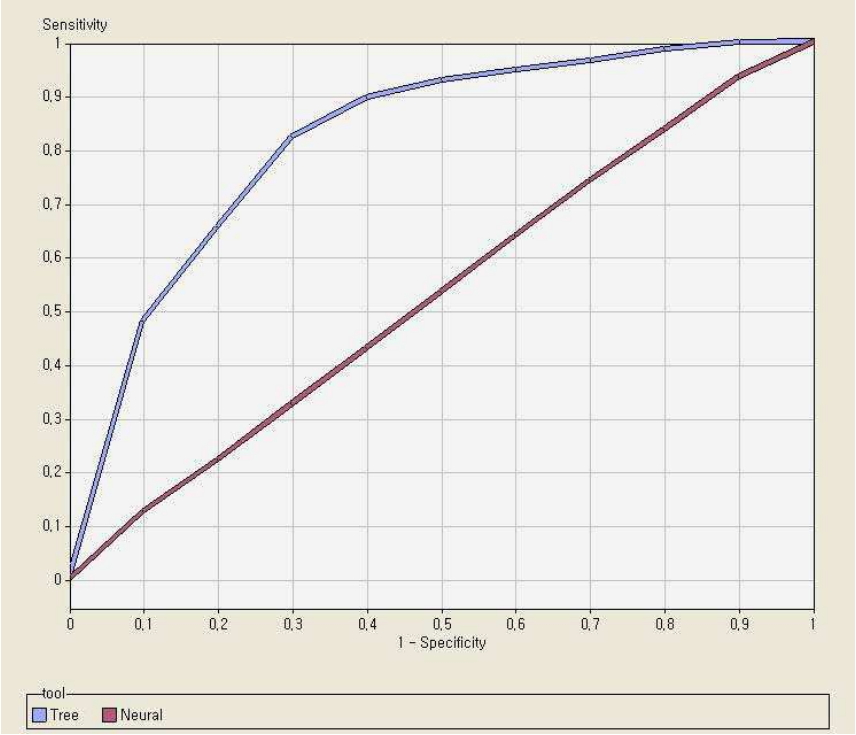
삼성전자 지수의 예측 결과를 종합해 보면, 의사결정트리의 75.2%로 가장 높게 나

타났고, 신경망 모형은 64.1%의 예측력을 보임으로써 의사결정트리모형이 보다 높은 예측력을 나타냄을 알 수 있다. 또한 KOSPI200지수와 마찬가지로 성능을 평가하기 위해 Lift Chart 와 ROC Curve를 산출하였다. [그림 4.3]은 신경망 모형과 의사결정트리 모형의 Lift Chart를 나타낸다.



[그림 4.3] Lift Chart(삼성전자)

[그림 4.3]을 보면 베이스 라인을 기준으로 의사결정트리 모형이 신경망 모형보다 훨씬 예측력이 우수하다는 것을 알 수 있다. [그림 4.4]는 삼성전자 주가에 대한 ROC Curve이다.



[그림 4.4] ROC Curve(삼성전자)

위의 ROC Curve를 통해 모형을 비교하면 의사결정트리 모형이 신경망 모형보다 우수함을 알 수 있다.

2. 포스코(POSCO)

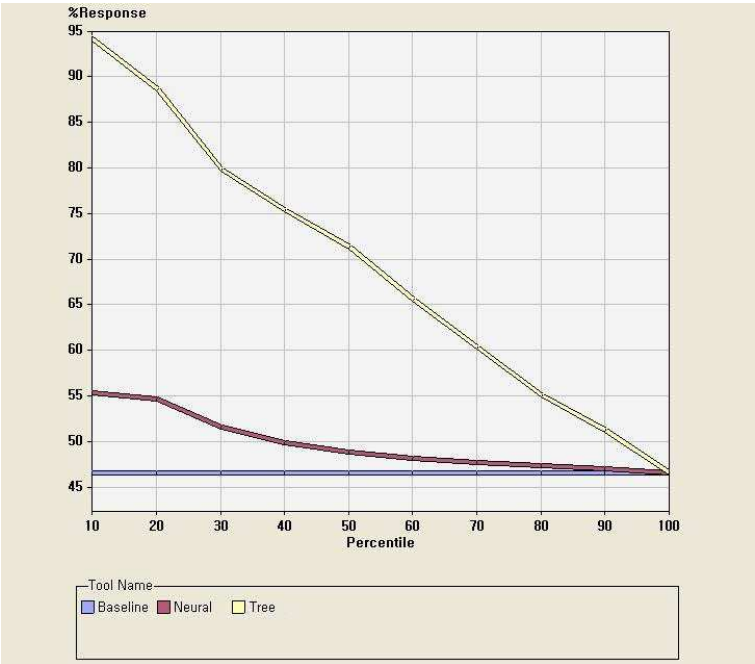
신경망이나 의사결정트리를 이용한 포스코 주가의 예측 결과는 KOSPI 200지수의 예측 결과와 마찬가지로 의사결정트리를 이용한 예측 결과(81.5%)가 신경망 모형의 예측결과(73.9%)보다 높았다. 신경망 모형에서 가장 좋은 예측력을 보인 모형은 은닉노드수가 1개인 경우로 분석용 집단 76.6%, 평가용 집단 72.5%, 검증용 집단 73.9%라는 최적의 모형을 찾을 수 있었다. 의사결정트리를 이용한 실험 결과 분석용 집단 73.1%, 평가용 집단 79.9%, 검증용 집단 81.5% 이었다. [표 4.5]는 포스코 주가 지수를 이용한 각 기법들의 예측 성능비교를 나타낸다.

[표 4.5] 포스코 주가를 이용한 각 기법들의 예측 성능비교 (단위:%)

	분석용	평가용	검증용
신경망	76.6	72.5	73.9
의사결정트리	73.1	79.9	81.5

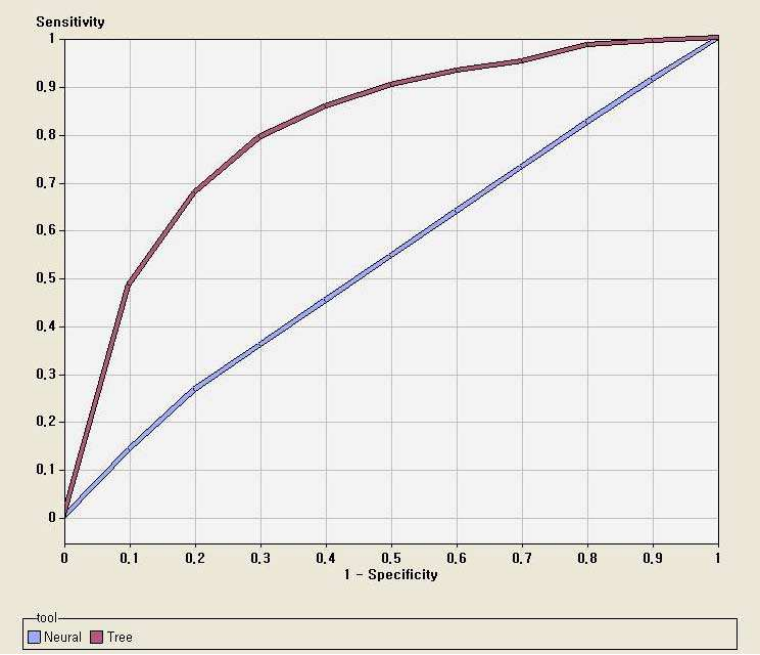
포스코 주가의 예측 결과를 종합해 보면, 의사결정트리를 이용한 예측 결과가 81.5%로 우수한 예측력을 보여주었고, 신경망 모형은 73.9%로 약간의 차이를 보였고 삼성전자 지수와 마찬가지로 포스코 주가 또한 의사결정트리 모형이 보다 나은 예측력을 보였다.

성능을 평가하기 위해 Lift Chart 와 ROC Curve를 산출한 결과 [그림 4.5]는 신경망 모형과 의사결정트리 모형의 Lift Chart를 나타낸다.



[그림 4.5] Lift Chart(포스코)

[그림 4.5]를 보면 베이스 라인을 기준으로 의사결정트리 모형이 신경망 모형보다 훨씬 예측력이 우수하다는 것을 알 수 있다. [그림 4.6]은 포스코 주가에 대한 ROC Curve이다.



[그림 4.6] ROC Curve(포스코)

위의 ROC Curve를 통해 모형을 비교하면 의사결정트리 모형이 신경망 모형보다 우수함을 알 수 있다.

3. KOSPI 200 지수의 모의실험

KOSPI 200지수를 이용하여 구축된 모형의 성능을 평가하기 위해 모형구축에 사용되지 않은 기간을 검증용 집단(Test1, Test2)으로 넣어 신경망, 의사절정트리 모형의 성능을 알아보았다.

[표 4.6] 모의실험을 위해 사용된 기간

모형 구축(Test)	Test 1	Test 2
2003.01 ~ 2007.12	2001.02 ~ 2002.12	2008.01 ~ 2008.05

[표 4.7] 각 기법들의 예측 성능비교 (단위:%)

	모형 구축(Test)	Test 1	Test 2
신경망	72.8	63.8	67.2
의사결정트리	82.9	78.7	84.6

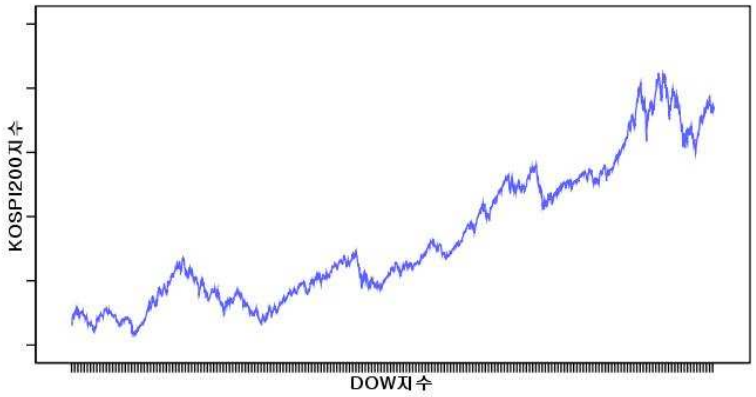
KOSPI 200지수로 구축된 모형에서 검증용 집단을 달리하여 모의실험을 실시한 결과, Test1 집단에서는 의사결정트리의 예측 결과가 78.7%로 높았으나 모형에서의 예측력 82.9%보다는 떨어지는 결과가 나타났고, Test 2 집단에서는 의사결정트리를 이용했을 때 예측결과가 84.6%로 매우 우수한 성능을 보여주었다. 이러한 결과

가 도출될 수 있었던 이유를 살펴보면, 최근 우리나라 증권시장의 흐름은 미국시장의 DOW-JONES 지수나 NASDAQ 지수와 거의 연동되어 나타나기 때문이라 할 수 있고, 이러한 변수들을 사용하여 우리나라 주가지수를 예측한다면 위의 결과에서 보이는 것과 같이 우수한 예측 성능을 보일 수가 있을 것이다. 이에 우리나라 지수와 미국지수와의 변동의 흐름을 자세히 살펴보기 위해, 본 연구에서 사용된 기간(2001.01～2008.05.)내의 KOSPI 200지수와 미국지수(DOW지수, NASDAQ지수)와의 변동 그래프를 그려보고, 각각의 상관관계를 구해보면 [표 4.8]과 같다.

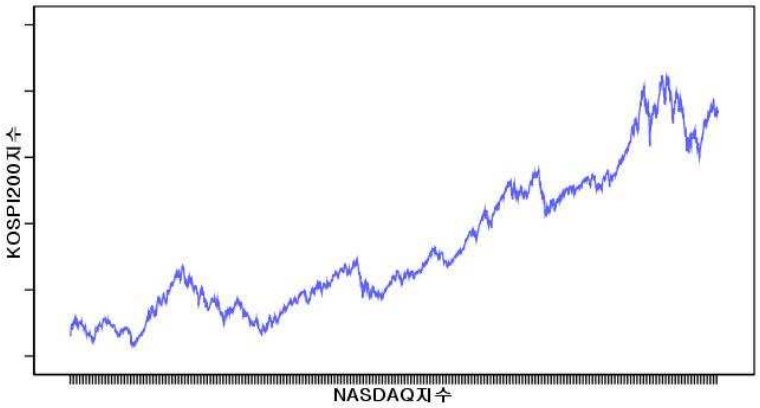
[표 4.8] KOPSI 200지수와 미국지수간의 상관관계

상관 계수	모형구축(Test)	Test 1	Test 2
KOSPI200/DOW종가	0.880	-0.903	0.854
KOSPI200/NASDAQ종가	0.777	-0.872	0.871

먼저 모형구축에 사용된 기간에서의 KOSPI 200지수와 미국지수(DOW지수, NASDAQ지수)의 변동관계를 알아보면, [그림 4.7]과 [그림 4.8]에서 보면 알 수 있듯이 몇 몇 구간에서 KOSPI 200지수와 미국지수간의 변동 흐름이 유사함을 알 수 있고, 상관계수 값을 구해보면 0.880(KOSPI 200지수-Dow지수), 0.777(KOSPI200지수-NASDAQ지수)로 변동 흐름이 유사함을 알 수 있겠다.



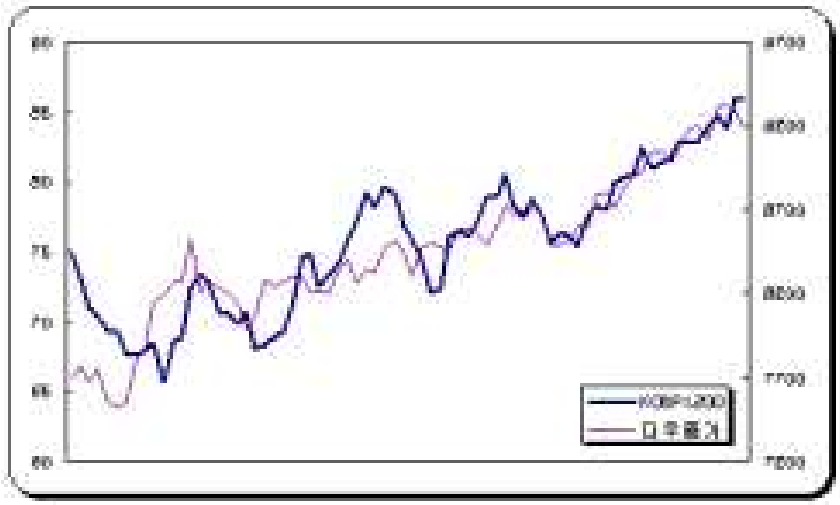
[그림 4.7] KOSPI 200지수와 DOW지수의 변동 그래프



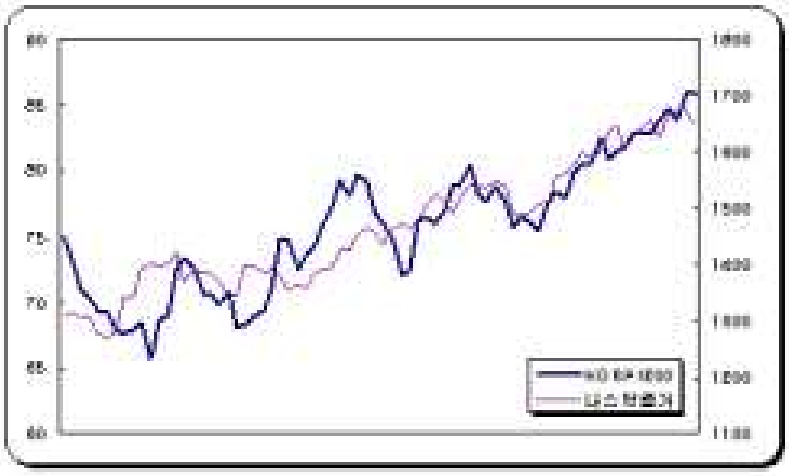
[그림 4.8] KOSPI 200지수와 NASDAQ지수의 변동 그래프

Test1 기간내의 KOSPI 200지수와 미국지수간의 변동 흐름을 살펴보면, [그림 4.7], [그림 4.8]에서 보면 알 수 있듯이 KOSPI 200지수는 계속 내려가고 미국지수는 계속 올라가는 현상으로, 거의 반대의 흐름을 보이고 있고 상관계수의 값을 구해보면 -0.903 (KOSPI 200지수/Dow지수), -0.872 (KOSPI 200지수/NASDAQ지수)로 Test1기간 내의 KOSPI 200지수와 미국지수간의 변동 흐름은 반대의 현상을 보여 모형구축에 사용된 기간 내의 예측결과보다는 떨어지는 결과를 나타내고 있다.

우수한 예측 결과를 보여준 Test 2기간 내에서의 KOSPI 200지수와 미국지수간의 변동 흐름을 살펴보기 위해 [그림 4.9], [그림 4.10]을 보면, Test 2기간 대부분의 구간에서 KOSPI 200지수가 미국지수와 거의 연동되어서 변동하고 있음을 보이고 있고, 상관계수 값을 구해보아도, KOSPI 200지수와 Dow지수는 0.854, KOSPI 200 지수와 NASDAQ지수는 0.871로 매우 높은 상관관계를 보여, 본 논문에서 구축된 모형의 성능이 우수함을 입증할 수 있었다.



[그림 4.9] Test 2 기간 내 KOSPI 200지수와 DOW지수의 변동 그래프



[그림 4.10] Test 2 기간 내 KOSPI 200지수와 NASDAQ지수의 변동 그래프

V. 결론

본 논문에서는 신경망과 의사결정트리의 예측모형을 KOSPI 200지수와 개별주식(삼성전자, 포스코)에 적용하여 각 예측모형들의 성능을 비교하였다. 모형의 예측성능을 높이기 위해 변수를 선정한 후 표준화를 하였으며 성능을 평가하기 위해 Lift Chart 와 ROC Curve를 이용하였다. 각 기법간의 예측 성능 결과는 다음과 같다.

[표 5.1] 각 기법간 예측 성능 비교 (단위:%)

	신경망	의사결정트리
KOSPI 200지수	72.8	82.9
삼성전자	67.4	75.2
포스코	73.9	81.5

가장 좋은 예측력을 보인 모형은 신경망의 경우 은닉노드수가 1개인 경우로 포스코 주가에서 73.9%의 예측력을 보였고 KOSPI 200지수에서 72.8% 삼성전자에서 67.4% 순으로 나타났다. 의사결정트리의 경우 KOSPI 200지수에서 82.9%로 가장 높은 예측력을 보였고 포스코, 삼성전자 순으로 나타났다. 또한 성능을 평가하기 위해 Lift Chart 와 ROC Curve를 산출한 결과 의사결정트리의 성능이 신경망 모형의 성능보다 높게 나타났으며 각 지수를 이용한 실험 결과 또한 신경망 모형의 성능보다는 의사결정트리를 이용한 예측 성능이 우수하게 나타났다. 또한 KOSPI 200지수의 모의실험 결과도 신경망 모형보다 의사결정트리의 예측 성능이 우수하

다. 특히 최근 기간(Test2)동안의 예측 성능이 매우 우수하게 나타났는데 이유는 최근 우리나라 증권시장의 흐름이 미국시장의 DOW지수나 NASDAQ지수와 매우 흡사하게 연동하고 있으며 본 논문에서는 이러한 변수들이 모형 구축에 사용되어 모형의 예측 성능이 우수하게 나타난 것으로 보인다.

[표 5.2] 알고리즘의 예측 성능비교 (단위:%)

	모형 구축(Test)	Test 1	Test 2
신경망	72.8	63.8	67.2
의사결정트리	82.9	78.7	84.6

첫째, 주식시장에서 향후(내일)의 주가지수를 예측하는 기존의 연구에서는 일별 주가지수의 예측성능이 낮았지만, 최근의 주식시장에서는 우리나라 주가지수가 미국지수와 연동하여 매우 비슷한 변동성을 보임으로 인해 본 논문에서는 일별 주가지수의 예측성능을 84.6%까지 향상시킬 수 있었다.

둘째, Stream data는 시계열 특성을 갖게 되는데 데이터 특성에 맞는 의사결정트리 알고리즘을 이용함으로써 자료간의 조합이나 관련성을 밝히고 또한 효과적인 예측 시스템을 구현할 수 있다.

향후 연구과제로는 본 논문에서 구축된 모형을 이용하여 실제 주식시장에서 수익률을 검증할 필요가 있으며, 표준화 과정 및 이상치 제거를 통하여 실제 금융시장에서 데이터마이닝의 예측기능의 향상을 기대 한다.

참고문헌

[1] 정용관, 윤용섭, "인공신경망모형을 이용한 주가의 예측가능성에 관한 연구", 1998.

[2] Vladimir N. Vapnik, "Statistical Learning Theory", WILEY-INTERSCIENCE Pub, 1998.

[3] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory", 1999.

[4] 권기택, "신경망을 이용한 주가예측시스템의 개발", 동양대학교 논문집 제 6집, 15-23, 2000.

[5] 임혜영, "SVM 분류기를 이용한 문서범주화 연구", 2000.

[6] 강상규, "인공신경망을 이용한 우리나라 증권시장에서의 주가의 비선형적 규칙성에 관한 연구", 1997.

[7] Michael Negnevitsky, "Artificial Intelligence - A Guide to Intelligent Systems", ADDISON WESLEY, 2000.

[8] Kryzanowski L. Galler, M., Wright, D. W., "Using Artificial Neural Networks to Pick Stocks", In Rober R. Trippi, Efraim Turban(Ed.), "Neural Networks in Finance and Investing", IRWIN, p525-541, 1996.

[9] Kimoto, A. K, Yoda, M. and Takeoka, M, “Stock market prediction system with modular neural network”, Proceedings of the I. Joint Conference on Neural Networks, San Diego, California, 1990.

[10] 김경재, 한인구, "Simultaneous Optimization Method of Feature Transformation and Weighting for Artificial Neural Networks Using Genetic Algorithm : Application to Korean Stock Market", 한국지능정보시스템학회, 1999.

[11] 김관문,조광재, "인공신경망을 이용한 주가지수의 방향예측", 대우경제연구소,

조사월보 11월호, p28-36, 1994.

[12] 정일찬, "인공신경망을 이용한 주식선별에 관한 연구", 2000.

[13] 오경석, "인공신경망을 이용한 우리나라 주식시장의 예측에 관한 연구", 1999.

[14] 류태욱, "부스팅을 이용한 서포트벡터 머신의 분류", 2001.

[15] 이재유, "인공신경망을 이용한 주가 예측", 1993.

[16] 이정길, "인공신경망 모형의 환율예측능력에 관한 실증연구", 1997.

[17] 현창호, "Artificial Neural Network를 이용한 일별 종합주가지수 예측", 1996.

[18] Richard O. Duda, Peter E. Hart, David G. Stork, "Pattern Classification", WILEY-INTERSCIENCE Pub, 2002.

[19] Vladimir N. Vapnik, Chervonenkis, "Theory of pattern Recognition", Moscow, Nauka, 1974.

[20] George Tsibouris, and Matthew Zeidenberg, "Testing the Efficient Markets Hypothesis with Gradient Descent Algorithms", Neural Networks in the Capital Markets Edited by Apostolos-Paul References, John Wiley & Sons Inc, 1995.

[21] Lapedes, A. and Faver, R., "Nonlinear Signal Processing Using Neural Networks", Prediction and System Modeling, Las Almos National Laboratory LA-UR-87-2662, 1987.

[22] Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh., "*Speaker Recognition Using Neural Networks and Conventional Classifiers.*", IEEE Transactions of Speech and Audio Processing, Vol.2, No.1, Jan, 1994.

[23] 윤여창,허문열, "신경망이론에 의한 시계열자료의 분석", 한국 통계 학회 Vol.4,pp.91-99, 1997

[24] 김대수, "신경망 이론과 응용", 하이테크 정보,pp.169-188, 1992

[25] Simon Haykin, "Netral Network(Acomprehensive foundation)", Prentice Hall International, pp.161-172, 1999

[26] 임은영, "패턴인식을 위한 가중치 Support Vector Machine 기법", 한양대학교, 2000.

[27] 이수용,이일병, "FUZZY이론과 SVM을 이용한 KOSPI200지수 패턴분류기", 한국데이터마이닝학회 추계학술대회 PROCEEDINGS, p331-348, 2002.

[28] Chin, K. K., "Support Vector Machines applied to speech pattern classification", PH.D. diss, Darwin College, 1998.

[29] C. F. Lin, "Fuzzy Support Vector Machines, IEEE Transactions on Neural Networks", Vol.13, No.2, 2002.

[30] Cortes and Vapnik, "Support Vector Network", Machine Learning 20, p273-297, 1995.

[31] Cristianini Nello, and John Shawe-Taylor, "An introduction to Support Vector.

[32] D. Roobaert, Direct SVM -, "A Simple Vector Machine Perceptron", J of VLSI Signal Processing 32, p147-156, 2002.

[33] D. Yang, "Probably Fast Training Algorithms for Support Vector Machines", www.kernel-machines.org.

[34] F. Alan, "Stock Selection using Support Vector Machines", www.kernel-machines.org.

[35] F. E. H. Tay, "Application of Support Vector Machines in Financial time series forecasting", Omega 29, p309-317, 2002.

[36] I. Takuya, "Fuzzy Support Vector Machines for Pattern Classification", www.kernel-machines. org.

[37] J. T. Jeng, “Support Vector Machines for the Fuzzy Neural Networks”, www.kernel-machines. org.

[38] M. W. Chang, “Analysis of nonstaionary time series using support vector machines”, www.kernel-machines. org.

[39] Steve Gunn, "*Support Vector Machines for Classification and Regression.*", ISIS Technical Report, 1998.

[40] Corinna Cortes, and Vladimir Vapnik, "*Support Vector Networks",* Machine Learning Vol.20, p273-297, 1995.

[41] Biggs, D., B. de Ville and E. Ville, A method of choosing multiway partitions for classification and decision tree, Journal of Applied Statistics, Vol. 18, pp.46-62, 1991

[42] Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone., Classification and regression trees, Wadsworth, Belmont, 1984

[43] 강현철, 한상태, 최종후, 김은석, 김미경, "데이터마이닝 방법론 및 활용", 자유아카데미, 2001

[44] Cezary Z. Janikow, "Fuzzy Decision Trees . Issues and Methods", IEEE Transactions on Systems, Man, and Cybernetics-Part B : Cybernetics, Vol. 28, No. 1, Feb., 1998

[45] 조용준, 허준, 최인규, "데이터마이닝 신경망분석" SPSS아카데미, 2000.

[46] 강현철, 한상태, “SAS Enterprise Miner를 이용한 데이터마이닝 방법론 및 활용”, 자유 아카데미, 1999.

[47] 최종후, 한상태 외, "SAS Enterprise Miner를 이용한 데이터마이닝 - 기능과 사용법", 자유아카데미, 1999.

[48] 오동근, "고객분석을 위한 데이터 마이닝 기법 비교", 2001.

[49] 이병국, "기술적지표의 조합을 통한 시스템 트레이딩의 유용성 평가 - KOSPI200 주가지수선물 중심으로", 2002.