

# 유전자 알고리즘을 활용한 인공지능 예측모형간 결합 기법: 주식시장에의 응용

안 현 철<sup>a</sup>, 이 형 용<sup>b</sup>

<sup>a</sup> 성신여자대학교 사회과학대학 경영학과  
서울 성북구 동선동 3가 249-1, 136-742

Tel: +82-2-920-7640, Fax: +82-2-920-2043, E-mail: [hcahn@sungshin.ac.kr](mailto:hcahn@sungshin.ac.kr)

<sup>b</sup> 한성대학교 경영학부  
서울 성북구 삼선동 2가 389, 136-792

Tel: +82-2-760-5960, Fax: +82-2-760-4482, E-mail: [leemit@hansung.ac.kr](mailto:leemit@hansung.ac.kr)

## 초록

각종 인공지능 기법들을 활용하여, 주식시장의 흐름을 예측하려는 연구가 지금까지 많은 인공지능 및 금융공학의 연구자들에 의해 시도되어 왔으며, 그 결과 다양한 인공지능 기법들이 예측 방법론으로 제시되어 왔다. 이런 가운데 서로 다른 예측모형들이 산출하는 예측결과를 종합 - 보완하는 결합기법에 관련된 연구가 90년대 후반부터 오늘날까지 꾸준히 발표되고 있다. 본 연구 역시 유전자 알고리즘 기반의 새로운 인공지능 예측모형간 결합기법을 제시하고 있다. 다만, 기존의 연구모형들이 각 개별모형 결과의 상대적 가중치에 초점을 맞추고 있었다면, 본 연구의 제안모형은 등락을 판단하는데 활용되는 임계치까지 유전자 알고리즘을 이용해 동시에 최적화하도록 설계되어 있다는 점에서 차별화된다. 제안모형의 유용성을 검증하기 위해, 본 연구에서는 지난 1998년부터 2007년까지의 KOSPI 지수 등락 예측을 위해 구축된 로지스틱 회귀모형, 인공신경망, SVM모형의 결과들을 제안모형을 이용해 결합하였다. 그 결과, 예측력 향상에 본 연구의 제안모형이 기여할 수 있음을 확인할 수 있었다.

## Keywords:

모형간 결합, 임계치 최적화, 유전자 알고리즘, 주가 예측

## 1. 서론

주식시장(stock market)의 예측은 자료가 방대하고, 잡음이 포함되어 있는 경우가 많은 특징이 있다. 또한 주식시장은 불확실성이 높고, 눈에 보이지 않는 제 3의 영향요인이 너무나도 많이 존재하고 있다. 예를 들어, 정치적 사건, 일반적인 경제 상황, 거래자(trader)의 기대심리 등 다양한 요소들이 주식시장에 복잡하게 영향을 미칠 수 있다. 이런 연유로 주식시장을 예측하는 것은 가장 난제(難題)에 속하는 경영학 문제라고 할 수 있다 (Carpenter et al., 1992).

금융공학 분야에서는 주가의 움직임이 무작위적(random)인지, 혹은 그렇지 않은지에 대한 많은 논란이 있어왔다. 하지만, 많은 학술 연구들을 통해, 주가의 움직임은 무작위적이지 않으며, 설명하기는 어렵지만 고도의 비선형적이면서도, 유동적인 방식에 의해 결정되고 있음이 지속적으로 제시되고 있다 (McNelis, 2005).

이러한 비선형적이면서도 복잡한 원리에 의해 결정되는 주식시장을 예측하기 위해, 최근에는 ‘인공지능 기법’을 도입하는 연구가 활발하게 소개되고 있다. 특히 LOGIT(Logistic Regression, 로지스틱 회귀분석), ANN(인공신경망, Artificial Neural Networks), SVM(Support Vector Machines) 등의 기법들은 주식시장의 등락을 예측하는데 있어서,

가장 대표적으로 적용되어 온 인공지능 기법들이다. 하지만, 이러한 기법들은 지금까지 개별적으로 적용되거나, 혹은 소위 앙상블(ensemble) 방법을 이용해 단일 기법을 서로 다른 설정을 통해 다양한 모형으로 구축해 결합하는 방식으로만 확장되었을 뿐, 서로 다른 이종(異種)의 인공지능 기법들이 동시에 적용되어, 종합적으로 결합되는 모형에 관한 연구는 거의 소개되지 않았다 (Kim et al., 2003; Kim et al., 2006; Lunga and Marwala, 2006). 특히 이종의 기법들을 통해 산출된 예측결과를 종합할 때, 기존 연구들은 대부분 투표(voting) 혹은 단순 평균(averaging)만을 통해 결합하는 것이 일반적이었다 (Huang et al., 2008).

본 연구는 LOGIT, ANN, 그리고 SVM 등 3가지 기법들의 예측결과를 결합하여, 보다 정확성을 높일 수 있는 주식시장 예측모형을 제안한다. 본 연구의 제안모형은 각 기법별 예측결과에 대한 가중치 뿐만 아니라, 등락을 가능하는 임계치(cutoff-value)까지 동시에 최적화함으로써, 보다 정밀한 예측결과를 생성할 수 있도록 설계하였다. 이러한 서로 다른 설계요소들을 동시에 최적화하기 위한 도구로서, 본 연구는 유전자 알고리즘(GA, genetic algorithms)를 제안하고 있다.

본 논문의 뒷부분은 다음과 같이 구성된다. 우선 2절에서는 기존 문헌 연구를 간단히 살펴보고, 3절에서는 본 연구의 제안모형인 유전자 알고리즘 기반의 새로운 이종 인공지능 기법간 결합 모형을 소개한다. 4절에서는 앞서 제시한 모형의 유용성을 검증하기 위한 실험 데이터 및 설계 내용을 설명하고, 5절에서는 실험 결과를 종합적으로 정리해 제시하도록 한다. 끝으로 마지막 절에서는 결론과 함께 본 연구와 관련한 향후 연구방향이 함께 제시된다.

## 2. 문헌연구

본 연구에서는 각 개별 통계 및 인공지능 모형들을 산출결과를 도출하고, 이를 결합하는 모형을 제시하려고 한다. 이에, 기존 문헌을 검토하게 될 본 절에서는 우선 본 연구에서 적용하는 기계학습 기법인 LOGIT, ANN, SVM에 대해 각각 간략히 살펴보고, 이어 경영분야의 예측과 관련하여 모형간 결합을 시도한 기존 연구들에 대해 간략하게 살펴보도록 한다.

### 2.1. 로지스틱 회귀분석(LOGIT)

분류를 하는데 있어서, 가장 흔한 경우는 이분법(二分法)을 기준으로 분류하는 경우다. 예를 들어, 인터넷 쇼핑물의 경우에도, 특정 고객이 물건을 살 것인가, 말 것인가 혹은 특정 이벤트에 고객이 반응을 할 것인가, 말 것인가 하는 문제들과 같이 이분법을 필요로 하는 분류 문제들의 해결이 종종 요구되는 것이다. 그런데, 이처럼 이항 확률을 가진 종속변수를 통계적으로 설명하고자 할 때, 일반적인 회귀 분석은 적용하기에 어려움이 있다. 왜냐하면 일반적인 회귀분석에서는 종속변수를 연속변수의 형태를 가정하는데 반해, 이러한 분류의 문제에서는 종속변수가 이항변수이므로 오차의 분포가 정규분포를 한다는 회귀분석의 일반적인 가정이 위배되기 때문이다. 또한 일반적인 회귀분석을 사용할 경우에는 예상값이 0과 1 사이에만 국한되지 않는다는 문제점 역시 발생할 수 있다. 이러한 문제들을 해결하기 위해 사용되는 기법이 바로 로지스틱 회귀분석(Logistic Regression) 기법, 약자로 로짓(LOGIT) 분석이다.

LOGIT의 일반적인 식은 다음과 같이 표현될 수 있다. 만일 어떤 사건이 일어날 확률에 1값을, 일어나지 않을 확률에 0값을 준 종속변수를 사용한다고 할 때, 로지스틱 회귀분석은 다음과 같은 형태를 갖게 된다.

$$P(event) = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}} \quad (1)$$

이 때,  $Z$ 는 독립변수들의 조합으로서, 아래와 같이 계산되어 진다.

$$Z = B_0 + B_1X_1 + B_2X_2 + \dots + B_pX_p \quad (2)$$

이처럼, LOGIT 분석은  $Z$  값이 연속적인 값을 가지더라도, 로지스틱 함수의 결과값은 마치 정규분포의 누적확률분포와 유사하게 0과 1사이의 값을 갖도록 되어 있기 때문에, 앞서 설명한 이분법적인 분류 문제를 해결하는 모형으로 적합하다. 이러한 장점으로 인해 LOGIT 분석은 지금까지 기업부도예측, 주가지수 예측 등 다양한 경영학 분야에 적용되어 왔다 (Ohlson, 1980; Bell et al., 1992; Fanning and Cogger, 1994; Barniv et al., 1997; Zhang et al., 1999 참고).

## 2.2 인공신경망(ANN)

인공신경망은 생물학적인 뇌의 작동 원리를 그대로 모방하는 방법으로, 그 강력한 예측력과 범용성으로 인해 오늘날 예측, 분류, 클러스터링 등 다양한 용도에 적용되고 있는 인공지능 기법이다. 인공 신경망 기법의 원리를 간략하게 살펴보면 다음과 같다.

인공신경망은 생물학적 뇌에 있어서 뉴런 (Neuron) 에 해당하는 처리 요소들 (Processing Elements) 로 구성된다. 노드 (node) 혹은 유닛 (unit) 라는 표현으로도 불리는 이들 처리 요소들은 입력, 순 입력, 출력, 가중치, 전이 함수 등 5개의 기본 요소로 이루어진다.

일반적인 인공신경망 모형에서는 입력층, 출력층, 은닉층의 3가지 층을 갖게 된다. 이 중 은닉층은 인공 신경망이 비선형 문제의 해결을 가능케 하는 핵심 요소인데, 인공 신경망의 모형에 따라 없을 수도 있고, 하나 혹은 그 이상의 은닉층을 가질 수도 있다. 이 중 가장 널리 활용되는 것은 은닉층이 1개인 3층 네트워크 모형이다.

이상 소개한 인공 신경망 기법은 (1) 복잡하고, 비선형적인 자료에서 지식이나 패턴을 추출할 수 있고, (2) 입력-출력 맵핑 기법이라서 자료에 대한 통계적인 분석 없이 결정을 수행할 수 있으며, (3) 상대적으로 적응력이 뛰어나고 견고한 모델이라는 점에서 다양한 분야에서 널리 적용되고 있다. 특히 경영학 분야의 문제해결과 관련하여, 90년대 이후 많은 연구들이 인공신경망을 예측모형으로 활용하였다 (Odom and Sharda, 1990; Tam and Kiang, 1992; Fletcher and Goss, 1993; Kim et al., 2002; Kim and Lee, 2004 등 참고)

하지만 모델이 제시하는 결과에 대해서 왜 그런 결과가 나오는지에 대한 원인을 명쾌하게 설명할 수 없다는 점과 과도하게 학습을 진행할 경우, 전체적인 관점에서의 최적해가 아닌 지역 내 최적해가 선택될 수 있다는 과적합화(overfitting) 문제는 인공 신경망 기법의 치명적인 단점이라고 할 수 있다 (Berry and Linoff, 1997).

## 2.3. Support Vector Machines (SVM)

SVM은 1998년 통계학자인 Vapnik에 의해 개발된 학습기법으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간의 선형문제로 대응시켜 나타내기

때문에 수학적으로 분석하는 것이 수월하다 (Hearst et al., 1998). 또한, SVM은 조정해야 할 파라미터의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다. 그리고 구조적위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 글로벌 최적해를 구할 수 있다는 점에서 ANN보다 성능이 좋은 기계학습기법으로 주목 받고 있다.

최근 몇 년간 SVM을 사용한 다양한 연구가 진행되었다. 그 예로서 SVM은 문서분류, 영상인식, 문자인식 등에서 뛰어난 일반화 성능을 보여주었다 (Joachims, 1998; Osuna et al., 1997). 또한, SVM을 재무분야에 적용한 연구도 있는데, 주로 시계열 예측 및 분류에 관한 것이다 (Tay & Cao, 2002; Kim, 2003). 또한 고객의 구매행동예측에도 SVM이 적용된 사례가 있다 (안현철 외, 2005).

## 2.4. 기법간 결합모형

모형간 결합을 통해, 예측력의 향상을 도모하고자 한 연구는 ‘패턴 인식’ 분야에서 먼저 활발하게 개척되었다. 문자인식에서 결합모형을 시도한 Ho et al.의 연구(1994)나 Huang and Suen (1995)의 연구, 안면인식에 결합모형을 시도한 Brunelli & Falavigna (1995)의 연구, 그리고 원격 감지(remote sensing)에 결합모형을 적용한 Benediktsson et al. (1997)의 연구 등이 패턴 인식분야에서 결합모형을 시도한 대표적인 사례가 되겠다.

이러한 모형간 결합은 경영학 분야의 문제해결을 위해서도 종종 적용되어 왔다. 우선 이재식 등 (2000)은 주가지수 선물옵션 거래를 위한 KOSPI 200지수의 등락예측모형을 서로 다른 설계의 인공신경망 모형을 결합하는 이른바 앙상블(ensemble) 기법을 제안하였다. Kim et al. (2002)은 인터넷 쇼핑몰의 고객행동을 예측하는 문제에 서로 다른 원리의 인공신경망 모형 3개를 결합하는 결합모형을 제시하였다. 한편 Kim et al. (2006)은 주가지수 예측에 있어, 서로 다른 예측기법들을 결합하는 결합모형을 제시하였다. 이 연구는 결합의 대상이 되는 결합모형이 모두 기계학습 모형이 아니라, ‘기계학습(인공신경망 모형)의 결과’, ‘전문가의 판단결과’ 그리고 ‘사용자의 판단결과’를 동시에 결합하도록 설계되어 있다는 점에서 그 특징이 차별화된다고 할 수 있다.

Lunga and Marwala (2006)는 Dow Jones 지수의 등락예측을 위한 모형으로, Learn++라 명명한 새로운 알고리즘을 제안하였다. Learn++는 유명한 앙상블 기법인 AdaBoost의 변형모형으로서, MLP(multi-layer perceptron) 모형에 앙상블을 접목한 방법론이라고 할 수 있다.

이렇듯, 지금까지 소개된 대다수의 방법론들은 동일한 종류의 기법, 그 중에서도 특히 설계가 복잡하고, 설계형태에 따라 예측성이나 오차의 형태가 다르게 나타날 수 있는 인공지능망 기법에 대해 ‘앙상블’ 형태로 결합한 경우가 대부분으로, 이종 기계학습 모형을 종합적으로 결합하고자 한 시도는 ‘경영학 분야’에서 그간 많이 소개되지 않았다. 이러한 이종 기계학습모형을 결합하기 위한 시도 중 하나로 최근 발표된 Huang et al. (2008)의 논문을 들 수 있다. 이 연구에서는 한국과 대만의 주가지수 등락 예측을 위하여, SVM, 사례기반추론(k-NN), LOGIT, ANN(BP, Back Propagation), 의사결정나무(C4.5) 등 5가지 인공지능 기법의 결합모형을 제시하고 있다. 하지만, 결합의 과정이 세련되지 못하고, 단순히 기법들의 투표(voting)만을 활용해 결합을 시도하고 있다는 점에서 예측력 향상에 한계가 있다고 할 수 있다.

### 3. 유전자 알고리즘 기반의 인공지능 예측기법간 결합모형

지금까지 문헌 연구를 통해 살펴본 바와 같이, 경영학 문제, 특히 ‘주식시장의 예측’과 관련하여, 이종 인공지능 기법들을 종합적으로 결합하고자 시도한 연구는 거의 없었다. 이에 본 연구에서는 유전자 알고리즘을 이용해 연속된 예측값을 생성할 수 있는 3가지 기법 - 즉 LOGIT, ANN, SVM<sup>1</sup> - 의 예측결과를 결합하여, 보다 우수한 예측결과를 생성시킬 수 있는 새로운 예측모형을 제안하고자 한다. 특히 본 연구에서는 각 예측기법별 최적 결합 가중치를 탐색하게 되는데, 결합 가중치 외에도 등락을 구분하는 기준 임계치(cut-off value, 일반적으로는 0.5로 활용)도 함께 최적화함으로써, 보다 정밀한 결합모형을 도출할 수 있도록

설계하였다. 본 연구의 제안모형은 다음의 3단계 과정에 의해 진행된다.

**1단계.** 이 단계에서는 최적화 대상이 되는 총 4개의 변수들에 대해, 유전자 알고리즘에 의해 탐색될 초기 개체군(population)이 생성되게 된다. 여기서 4개의 변수란 각 기법, 즉 LOGIT, ANN, SVM에 대한 상대적 가중치와 등락을 구분하는 기준 임계치 변수가 된다. 각 기법별 가중치의 경우 상대적 가중치를 도출해야 하므로, 각각 1에서 128사이의 값을 갖도록 설계하였다. 그래서,  $V_i$ 가 유전자 알고리즘에 의해 탐색되는 각 가중치 변수를 의미한다고 할 때, 실제 상대적 가중치  $v_i$ 는 다음의 수식에 의해 도출된다.

$$v_i = \frac{V_i}{\sum_i V_i} \tag{3}$$

128의 경우,  $2^7$  이므로, 각 기법의 가중치 변수를 위한 염색체 (chromosome)에는 각각 7비트씩 할당하였다. 한 편, 등락을 구분하는 기준 임계치는 이론상 0에서 1사이의 값을 갖는 실수인데, 이 변수의 경우 14비트를 할당하였다. 이를 통해, 1/10,000 이상의 정밀도를 가질 수 있도록 하였다. 다음의 그림 1은 본 연구에서 제안하고 있는 모형에서 적용하고 있는 유전자 알고리즘의 전체 염색체 구조를 도식으로 표현하고 있다.

**2단계.** 앞의 단계에서, 초기 개체군이 형성되면, GA는 초기 설정된 염색체에 대해 교배 (crossover), 돌연변이 (mutation) 등 다양한 과정을 적용하여, 계속 새로운 각 기법별 가중치 및 등락 임계치 후보들을 생성해 다시 학습용 데이터에 적용해 보게 되며, 중지 조건 (stopping condition)이 만족될 때까지 상기 작업을 계속 반복하게 된다. 이를 통해 학습용 데이터에 가장 우수한 성과를 보이는 4개의 변수값을 탐색하게 된다.

**3단계.** 앞의 1, 2단계를 반복하다보면, GA의 중지 조건이 만족되는 시점에서 최적 혹은 최적에 근접한 4개의 변수값이 도출되게 된다. 3단계에서는 이렇게 도출된 각 기법별 가중치와 등락 임계치를 모형 구축에 사용하지 않은 검증용 데이터(hold-out data)에 적용하여, 그 결과를 살펴보게 된다.

<sup>1</sup> 본래 SVM은 두 집단에 대해 소속여부를 판단하는 이분류(binary classification) 모형으로 설계되어 있다. 하지만, 각 집단에 속할 확률이 얼마나 되는지를 추정할 수 있는 방법이 이미 제시되어 있기 때문에(홍태호, 신택수, 2005), 본 연구에서는 이 추정확률(estimated probability)을 활용하여, SVM의 산출결과로 활용하였다.

Population (A generation)	A chromosome																		
	LOGIT 가중치				ANN 가중치				SVM 가중치				등록 기준 임계치						
	V <sub>11</sub>	V <sub>12</sub>	...	V <sub>17</sub>	V <sub>21</sub>	V <sub>22</sub>	...	V <sub>27</sub>	V <sub>31</sub>	V <sub>32</sub>	...	V <sub>37</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	...	T <sub>12</sub>	T <sub>13</sub>	T <sub>14</sub>
Chromosome 1	1	0	...	1	1	1	...	0	0	1	...	1	1	0	0	...	1	1	0
Chromosome 2	0	1	...	1	0	1	...	1	0	0	...	1	0	0	1	...	0	1	0
Chromosome 3	1	1	...	0	0	1	...	0	1	0	...	0	1	0	1	...	1	0	0
⋮																			
Chromosome m	0	0	...	1	0	1	...	0	1	1	...	1	1	1	0	...	1	1	1

그림 1- 제안모형의 염색체 구조

때문에, 모형 구축에 사용되지 않은 검증용 데이터에 적용해, 이 같은 이른바 ‘과잉학습(overlearning)’ 문제가 발생한 상황인지 아닌지를 최종적으로 확인해 볼 필요가 있다.

#### 4. 실험설계

##### 4.1. 실험 데이터

제안된 연구모형의 유용성을 검증하기 위해 한국종합주가지수(KOSPI)의 과거 자료를 이용하여 예측모형을 구축한다. 모형 구축에 사용된 표본은 한국종합주가지수의 1999년부터 2006년까지의 한국종합주가지수의 일별 종가자료이다. 표본의 추출은 임의추출방식에 의해 수행되었으며 총 표본의 크기는 1,826개이다. 수집된 전체 데이터 중 80%에 해당되는 1,460건의 데이터를 학습용으로 활용하였고, 나머지 20%를 검증용으로 사용하였다. 각 데이터셋의 크기와 상대비율은 표 1과 같다.

표 1. 데이터셋 별 데이터 크기와 비율

데이터	데이터 개수	전체 데이터에 대한 비율
실험용	1,460	80%
검증용	366	20%
전체	1,826	100%

본 연구에서는 모형의 구축을 위해 일반적으로 주가지수 예측에 많이 사용되는 기술적 지표들 중심으로 입력변수를 선정하였다. 본 연구에서는 Kim (2003)의 연구에서 활용한 12개의 지표를 입력변수로 활용하였는데, 그 내역과 정의가 다음의 표 2에 제시되어 있다.

표 2. 선택된 변수와 관련 선행연구

변수명	산식	관련 연구
Stochastic %K	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$	Achelis (1995)
Stochastic %D	$\frac{\sum_{i=0}^{n-1} \% K_{t-i}}{n}$	Achelis (1995)
Stochastic Slow %D	$\frac{\sum_{i=0}^{n-1} \% D_{t-i}}{n}$	Gifford (1995)
Momentum	$C_t - C_{t-4}$	Chang et al.(1996)
ROC	$\frac{C_t}{C_{t-n}} \times 100$	Murphy (1986)
Williams' %R	$\frac{H_n - C_t}{H_n - L_n} \times 100$	Achelis (1995)
A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$	Chang et al.(1996)
Disparity5	$\frac{C_t}{MA_5} \times 100$	Choi (1995)
Disparity10	$\frac{C_t}{MA_{10}} \times 100$	Choi (1995)
OSCP	$\frac{MA_5 - MA_{10}}{MA_5}$	Achelis (1995)
CCI	$\frac{(M_t - SM_t)}{(0.015 \times D_t)}$	Achelis (1995) Chang et al.(1996)
RSI	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} Up_{t-i}/n}{\sum_{i=0}^{n-1} Dw_{t-i}/n}}$	Achelis (1995)

## 4.2. 실험 설계

GA 탐색을 위한 제어 파라미터들과 관련해서는 모집단을 100개체(organisms)로 설정하였으며, 교배 및 돌연변이 비율에 대해서는 각각 0.7, 0.1로 설정하였다. 아울러 중지 조건으로는 5000회 반복, 즉 50세대만큼 탐색을 반복하도록 설정하였다.

제안모형의 성과를 보다 정밀하게 검증하기 위해, 본 연구에서는 다음의 표 3과 같은 다양한 비교모형들을 설정하였다. 아래 표에서 볼 수 있듯이, 현재 본 연구의 제안모형은 WAVG+T로 명명된 모형으로서, 가중치 최적화와 임계치 최적화를 모두 GA에 의해 수행하는 모형이 된다.

표 3. 비교모형 및 제안모형

모형	모형 결합	가중치 최적화	임계치 최적화	비고
LOGIT	없음	해당없음	없음	비교모형
ANN	"	"	"	"
SVM	"	"	"	"
SAVG1	결합	단순평균	"	"
SAVG2	"	가중평균*	"	"
WAVG	"	GA 탐색	"	"
SAVG+T	"	가중평균*	GA 탐색	"
WAVG+T	"	GA 탐색	GA 탐색	제안모형

\* 가중평균의 경우, 학습용 데이터에 대한 각 기법별 예측정확도를 가중치로 활용하였음

## 5. 실험 결과

전체적인 실험결과가 다음의 표 4에 제시되어

표 4. 실험 결과

모형	학습용 데이터 적중율	검증용 데이터 적중율	LOGIT 가중치	ANN 가중치	SVM 가중치	등락 임계치	최적 모형의 설정값
LOGIT	54.11%	52.46%	1	0	0	0.5	Stepwise: Backward-Conditional
ANN	54.66%	51.64%	0	1	0	0.5	# of nodes in the hidden layer = 24
SVM	66.23%	53.83%	0	0	1	0.5	RBF Kernel, C=100, $\delta^2=25$
SAVG1	55.75%	51.09%	1	1	1	0.5	
SAVG2	56.37%	51.09%	0.5411	0.5466	0.6623	0.5	
WAVG	65.82%	55.46%	14	1	109	0.5	
SAVG+T	55.55%	54.64%	1	1	1	0.5178	
WAVG+T	66.71%	55.74%	8	1	123	0.5024	

있다. 이 표의 결과를 통해 알 수 있듯이, 본 연구의 제안모형인 WAVG+T를 적용할 경우, 학습용-검증용 데이터 모두에서 가장 우수한 예측력을 얻을 수 있음을 확인할 수 있다.

## 6. 결론 및 향후 연구방향

본 논문에서는 유전자 알고리즘 기반의 이중 인공지능기법간 결합모형을 제안하였다. 본 연구의 제안모형은 각 기법별 가중치만 최적화하는 것이 아니라, 등락여부를 판단하는 기준 임계치까지 동시에 최적화함으로써, 보다 정밀한 예측이 가능하도록 설계되었다. 연구모형의 유용성을 확인하기 위해, 약 10여년간 수집된 국내 KOSPI지수 등락예측에 제안모형을 적용해 본 결과, 단일 모형이나 기존에 제시된 다른 결합모형을 적용하는 것에 비해, 제안 모형이 더 높은 예측성능을 제공함을 확인할 수 있었다.

본 연구결과를 시금석으로 하여, 향후 본 연구진은 현재 제시된 연구모형을 보다 확장, 발전시킬 계획이다. 우선, 현재 단일값으로 설정되어 있는 임계치를 2개의 임계치로 확장, 발전시키고자 한다. 이 경우, 종합된 예측값이 상단 임계치보다 크면, 주가가 오를 것으로 예측하고, 하단 임계치보다 작으면, 주가가 내릴 것으로 예측하지만, 예측값이 상단 임계치와 하단 임계치 사이의 값을 가질 경우에는 ‘판단 보류’로 예측하도록 설계된다. 이는 투자자들이 실제 투자를 수행할 경우, 주식에 대해 매입(Buy), 매도(Sell), 보류(Hold)의 3가지 중 하나로 의사를 결정한다는 점을 고려할 때, 투자 의사결정 관점에서 보다 현실을 반영한 예측모형을 구축할 수 있다는 장점을 갖게 된다 (이재식 등, 2000).

아울러, 학습용 데이터에 대한 최적 샘플 선택, 즉 instance selection 역시 연구설계에 반영할 수 있다. 학습용 데이터가 이상치(outlier)에 속할 경우, 이는 전체적인 기법별 가중치를 도출할 때, 왜곡을 유발할 수 있다. 때문에, 연구모형에 최적 학습용 사례를 선택하는 요소를 추가하는 것 역시 성과 개선에 도움이 될 것으로 예상된다.

세번째, 연구결과의 가치를 사후분석하는데 있어, 단순히 적중율만 고려하는 것이 아니라, 투자수익율도 함께 고려해 보고자 한다. 특히 앞서 제시한 2개의 임계치 도입을 통해, ‘보류’ 의사결정도 가능해진다면, 최적화의 대상 자체를 ‘적중율’보다는 ‘투자수익율’로 설정하는 것이 더 효과적일 것으로 생각된다.

마지막으로 현재 본 연구는 제안모형을 주가지수 등락예측에 적용하였지만, 모형 자체는 다른 경영분야의 의사결정문제에도 얼마든지 적용될 수 있다. 때문에, 향후 연구를 통해 고객의 구매예측모형이나 기업부도 예측 등 다른 분야에도 제안모형을 적용해 보고, 그 유용성을 검증해 볼 계획이다.

## 참고문헌

- [1] 안현철, 김경재, 한인구 (2005). Support Vector Machine을 이용한 고객구매예측모형. *한국지능정보시스템학회 논문지 11(3)*, 69-82.
- [2] 이재식, 송영균, 허성희 (2000). 인공신경망 앙상블을 이용한 옵션 투자예측 시스템. *한국지능정보시스템학회 2000년 학술대회 제2권*, 489-497.
- [3] 홍태호, 신태수 (2005). Using Estimated Probability from Support Vector Machines for Credit Rating in IT Industry, *한국지능정보시스템학회 2005년 추계학술대회논문집*, 509-515.
- [4] Achelis, S.B. (1995). *Technical analysis from A to Z*. Probus Publishing: Chicago.
- [5] Barniv, R. et al. (1997). Predicting the outcome following bankruptcy filing: a three-state classification using neural networks. *International Journal of Intelligent Systems in Accounting, Finance and Management 6(3)*, 177-194.
- [6] Bell, T. et al. (1990). Neural nets vs. logistic regression: a comparison of each model's ability to predict commercial bank failures. *Proceedings of the 1990 Deloitte & Touche/University of Kansas Symposium on Auditing Problems*, 29-58.
- [7] Benediktsson, J.A., Sveinsson, J.R., Ersoy, O.K. & Swain, P.H. (1997). Parallel consensual neural networks. *IEEE Transactions on Neural Networks 8(1)*, 540-564.
- [8] Brunelli, R. & Falavigna, D. (1995). Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17(10)*, 955-966.
- [9] Carpenter, G., Grossberg, S., Marhuzon, N., Reynolds, J. & Rosen, D. (1992). Artmap: A neural network architecture for incremental learning supervised learning of analog multidimensional maps. *IEEE Transactions in Neural Networks 3*, 678-713.
- [10] Chang, J., Jung, Y., Yeon, K., Jun, J., Shin, D., & Kim, H. (1996). *Technical indicators and analysis methods*. Jinritamgu Publishing: Seoul.
- [11] Choi, J., (1995). *Technical indicators*, Jinritamgu Publishing: Seoul.
- [12] Fanning, K. & Cogger, K., (1994). A comparative analysis of artificial neural networks using financial distress prediction. *International Journal of Intelligent Systems in Accounting, Finance and Management 3(3)*, 241-252.
- [13] Fletcher, D. & Goss, E., (1993). Forecasting with neural networks and application using bankruptcy data. *Information & Management 24*, 159-167.
- [14] Gifford, E. (1995). *Investor's guide to technical analysis: Predicting price action in the markets*. Pitman Publishing: London.
- [15] Hearst, M.A. et al. (1998). Support vector machines. *IEEE Intelligent System 13(4)*, 18-28.
- [16] Ho, T.K., Hull, J.J. & Srikari, S.N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis, Machine Intelligence 16(1)*, 66-75.
- [17] Huang, C.-J., Yang, D.-X. & Chuang, Y.-T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications 34(4)*, 2870-2878.
- [18] Huang, Y.S. & Suen, C.Y. (1995). A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE*

- Transaction on Pattern Analysis and Machine Intelligence* 17(1), 90–94.
- [19] Joachims, T. (1998). Text categorization with support vector machines. *Proceedings of the European Conference on Machine Learning (ECML)*, 137-142.
- [20] Kim, E., Kim, W. & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems* 34(2), 167-175.
- [21] Kim, K.J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2), 307-319.
- [22] Kim, K.J. & Lee, W.B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation, *Neural Computing & Applications* 13(3), 255-260.
- [23] Kim, M.-J., Min, S.-H. & Han, I. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications* 31(2), 241-247.
- [24] Kim, T.S., Yoon, J.H., & Lee, H.K. (2002). Performance of a nonparametric multivariate nearest neighbor model in the prediction of stock index returns. *Asia Pacific Management Review* 7, 107-118.
- [25] Lunga, D. & Marwala, T. (2006). Online Forecasting of Stock Market Movement Direction Using the Improved Incremental Algorithm. *Lecture Notes in Computer Science* 4234, 440-449.
- [26] McNelis, P.D. (2005). *Neural Networks in Finance: Gaining the predictive edge in the market*, Elsevier Academic Press: Oxford-UK.
- [27] Murphy, J.J. (1986). *Technical analysis of the futures markets: A comprehensive guide to trading methods and applications*. Prentice-Hall: New York.
- [28] Odom, M. & Sharda, R., (1990). A neural network model for bankruptcy prediction. *Proceedings of the International Joint Conference on Neural networks*, II-163-II-168.
- [29] Ohlson, J.A. (1980). Financial Ratios and Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18(1), 109-131
- [30] Osuna, E. et al. (1997). Training support vector machines: an application to face detection. *Proceedings of Computer Vision and Pattern Recognition*, 130-136.
- [31] Tam, K. & Kiang, M. (1992). Managerial applications of neural networks: the case of bank failure prediction. *Management Science* 38(7), 926-947.
- [32] Tay, F.E.J., & Cao, L.J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing* 48, 847-861.
- [33] Zhang, G. et al. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research* 116, 16-32.