한국지능시스템학회 논문지 2008, Vol. 18, No. 6, pp. 781-786

주가지수 관계와 유전자 알고리즘을 이용한 주식예측

Stock Forecasting using Stock Index Relation and Genetic Algorithm

김상호* · 김동현** · 한창희*** · 김원일**

Sangho Kim*, Donghyun Kim**, Changhee Han*** and Wonil Kim**

* 세종대학교 컴퓨터공학과

** 세종대학교 디지털콘텐츠학과

*** 육군사관학교 전자정보학과

요 약

이 논문에서 우리는 선형결합으로 표현된 여러 가지 주가지수의 관계를 찾아내어 주식의 등락을 예측하는 방법을 제안한다. 제안된 방법에서 중요한 점은 전체 주가지수 중에서 예측하는 지수와 관계가 있는 주가지수들을 선택하는 것과 그 주가지수의 적절한 관계를 찾아내는 것이다. 전체 주가지수와의 관계를 설정하는 것은 불가능하기 때문에 밀접한 관계가 있는 주가지수만을 이용하였고 주가지수의 관계를 찾는 방법으로 유전자 알고리즘(GA : Genetic Algorithm)을 사용하였다. 제안된 방법을 이용하여 2005년부터 2007년도까지의 실제 주가지수를 가지고 모의투자 시뮬레이션을 한 결과 모의 투자금액이 230% 증가하는 것을 확인하였다.

키워드 : 주식예측, 유전자 알고리즘, 선형 회귀 분석, 선형 결합

Abstract

In this paper, we propose a novel approach predicting the fluctuation of stock index by finding a relation in various stock indexes that are represented by linear combinations. The important points are to select stock indexes related to predicting indexes and to find the proper relations in them. Since it is unattainable to use entire stock indexes relation, we used only data that are closely associated with each other. We used Genetic Algorithm(GA) to find the most suitable stock-index relation. We simulated the investment in years from 2005 to 2007 with each real index. Finally we verified that the investment money increased 230 percents by the proposed method.

Key Words : Stock Forecasting, Genetic Algorithm, Linear Regression Analysis, Linear Combination

1. 서 론

컴퓨터 예측 기술이 점차 발전하면서 금융, 경제, 수요, 주식 등 다양한 예측 분야에서 컴퓨터의 활용도가 점점 늘어나고 있다. 그 예로, 한국은행에서 개발한 경제 예측 모델 BOKAM97과 BOK97[1][2], 주식예측모델인 자본 자산 가격 결정 모형과 차익 거래 가격 결정 모형 등 여러 분야에서 다양한 방법으로 사용되고 있으며 점차 범위가 확장되고 있다[3][4].

주식을 예측하는 방법은 지수의 등락·흐름, 환율, 금리, 매시간 변화하는 환경, 각 연관된 종목끼리의 관계 등 다양한 변수를 이용한다. 우리는 이중에서 연관된 종목끼리의 관계를 GA를 사용해서 효과적으로 주식의 등락을 예측하는 방법을 제안한다. 이 논문에서는 금융업의 주가지수를 은행, 보험, 증권과 같은 관계가 있는 주식들을 사용하여 주가지수를 예측하였다. 예측한 결과를 사용하여 모의 투자한 결과, 투자금액이 230% 증가하는 것을 확인하였고, 이 제안된 방법은 다른 관계된 종목들에게도 적용될 수 있다.

위의 주가지수의 예측방법을 제안하기 위해서 2장에서는 이전의 주식 예측을 하는 여러 연구에 대한 설명을 하고, 3장에서는 주식 예측의 사용될 일련의 과정들에 대한 기본적 지식과 그 사전 지식들을 토대로 예측 모델을 제안할 것이며, 4장에서는 실제 모의 투자를 하여 그 결과 값을 알아볼 것이다. 그리고 5장은 결론 부분이다.

접수일자 : 2008년 5월 29일
완료일자 : 2008년 8월 20일
본 연구는 서울시 산학연 협력사업의 지원에 의해 수행되었음.

2. 주식예측에 대한 이전의 연구

이전 연구들에 따르면 다양한 방법을 사용하여 주식을 예측하는 프로그램들이 많은 연구자들에 의해서 개발되어 왔다. Li는 필터 룰의 파라미터를 GA를 이용하여 주식예측을 하는 방법을 사용하였다[5]. 필터 룰은 주가가 최근의 저가에서 일정비율 x만큼 상승했을 때를 상승 국면에 진입한 것으로 보아 매입하고 반대로 최근의 고가에서 일정비율 x만큼 하락했을 때를 하락국면으로 보아 매도하는 것이다. 모든 항목의 주식 데이터의 값을 가지고 예측하기에는 데이터의 범위가 너무 광범위하므로 작은 지역 범위를 생성하여

그것을 가지고 예측한 후 실제 주식 데이터의 적용시키는 방법을 사용하였다. 그러나 필터 룰은 Fama의 효율적 시장 가설이 제기되면서 많은 학자들에 의해 유효성이 부정되었다[6]. 효율적 시장가설이란 모든 정보가 가격형성에 즉각적으로 반영되어 누구라도 계속적으로 우수한 투자성과를 올릴 수 없는 것을 말한다.

Almanza는 기회 발견을 이용하여 주식 예측을 하는 방법을 사용하였다[7]. 기회 발견이란 rare probability로 나타나는 사건이지만 중요한 영향을 미치는 요소들을 찾아내는 것을 말한다. 하지만 전체적으로 일어나는 사건 중에서의 중요한 영향을 미치는 요소들을 놓칠 수 있다는 단점이 있다.

Lertwachara는 매력 있는 주식을 선택하는 방법으로 금융정보, 추천 주식정보, 일상적으로 일어나는 재정 문제의 확률적 정보를 가지고 GA와 신경 회로망을 사용하여 예측하는 방법을 사용하였다[8]. 비선형 모델은 많은 파라미터들을 사용하고 그 선택의 범위가 넓기 때문에 선형 모델이 비선형 모델에 비해서 정확한 주식 관계를 찾아내는 데에 좀 더 효과적이라고 서술하였다.

Irwin은 Lertwachara와는 반대로 비선형 모델 방법으로 접근하였다[9]. 비대칭 비선형 모델인 마르코프 체인 방법과 GA의 결합을 통해 시계 열 변동률을 예측하는 방법을 사용하였다. 마르코프 체인은 어떤 상황에서의 가능성은 이전의 상황에 의해서만 영향을 받는다고 가정하는 것이다. 즉, 이벤트 간 변동률에 대한 예측을 이전 상태의 파라미터들과 GA의 결합을 통해 확률 값으로 예측하는 것을 말한다. 마르코프 체인은 시계 열 문제를 해결하는 데에 있어 현재시점에 적용이 가능한 이전 상태에 대한 파라미터들을 설정하는 데에 어려움이 많은 단점이 있다.

이와 같이 여러 연구들에 의해 주식예측을 위한 다양한 방법들이 제안되고 있다. 우리는 선택의 범위가 넓은 다양한 파라미터들을 사용해야 하는 비선형 모델보다는 좀 더 정확한 주식 관계를 찾아내는 선형모델을 이용한 주가지수 예측 방법을 제안한다.

3. 제안된 주식 예측 방법

본문에서의 궁극적인 목적은 주가지수의 선형관계를 GA를 이용하여 찾아내어 주가 등락을 효과적으로 예측하는 데에 있다. 먼저 우리는 예측모델들을 살펴본 뒤에 어떠한 방향으로 접근을 하여야 효과적인지에 대해서 알아보았다.

3.1 예측모델

예측 모델이란 알고 있는 하나 이상의 파라미터들을 가지고 모르는 여러 상황이나 현상을 예측하는 모델이다. 이것은 하나 이상의 파라미터들로 표현되어진 함수로 나타내어지며 그 결과 값으로 예측을 한다. 예측 율을 높이기 위해서는 예측모델과 목적함수의 적절한 선택, 그리고 그 구조에 알맞은 최적의 파라미터를 찾는 것이 중요하다.

예측모델에는 선형구조를 이용한 회귀 모델, 회귀 지역구분 모델, 통계적 방법을 이용한 예측 모델, 분류 예측모델 등이 있다[10]. 우리는 이들 중에서 해석하기 용이하고 파라미터의 선택이 쉬운 선형구조를 이용한 회귀 모델의 방식으로 접근을 하였다.

회귀 분석은 변수간의 관계를 분석하여, 알고 있는 변수를 기초로 하여 알려지지 않은 변수의 값을 예측하는 통계적 분석 방법을 말한다. 여기서 단순히 변수간의 밀접한 정도를 분석하는 것이 상관관계 분석이다[11]. 그림 1은 선형회귀의 한 예를 그래프로 표현한 것으로서 임의의 점들을 선택하여 각 점들과 한 선분과의 길이가 가장 짧은 선분을 선택하는 것을 말한다.

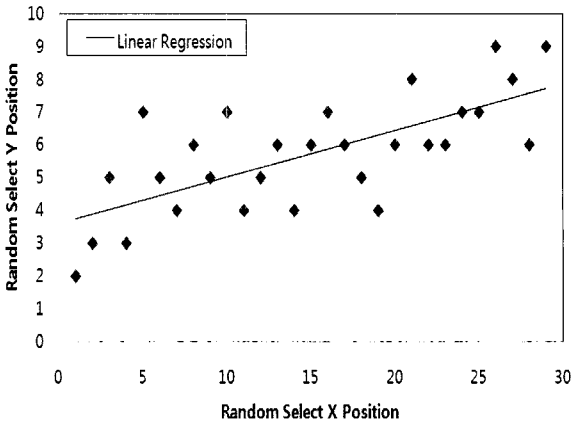

그림 1. 선형 회귀 분석의 한 예
Fig. 1. Illustration of linear regression on a data set

우리가 제안하는 방법이 여러 주식들 간의 관계를 구하는 것이기 때문에 위의 회귀 분석을 이용하여 최적 선형 결합을 찾아내는 방법을 사용하였다.

3.2 관련된 주가지수 검색

실제로 존재하는 모든 주식 데이터의 관계를 구하기에는 한계가 있다. 그렇기 때문에 구하고자 하는 값과 가장 관련이 높은 주식 데이터들만 추려내어 사용한다.

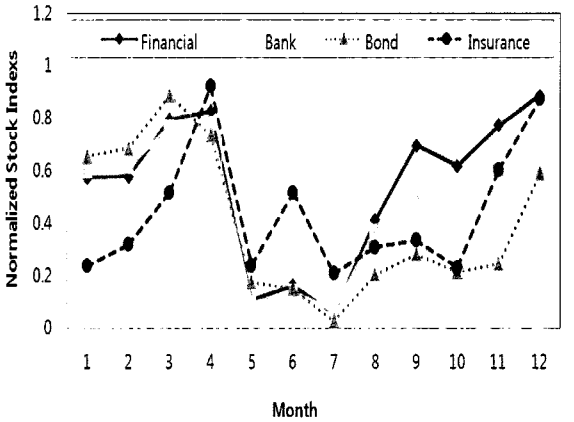

그림 2. 2004년 항목별 정규화한 주가지수
Fig. 2. Itemized Normalize Stock Indexes for 2004 years

그림 2는 각각의 주가지수의 변동범위 및 등락 값이 다르기 때문에 정규화를 사용하여 항목별 주가지수가 미치는 영향을 같게 표현된 것이다. 그림 2에서 알 수 있듯이 금융업, 은행, 증권, 보험의 주가지수는 서로 밀접한 관계가 있다고 말할 수 있다.

3.3 GA

GA는 생물학적 진화모델을 기초로 한 강력한 학습 알고리즘이다. 즉, 자연계의 법칙에 착안하여 주어진 환경에 잘 적응하는 우성인자만을 선택, 교배, 돌연변이 등 의 인위적 조작을 함으로써 최적 해를 찾아내는 알고리즘이다. GA는 NP-complete 최적화 문제와 같은 광대하고 복잡한 문제들의 최적 해를 구하는데 효과적으로 쓰이고 있다[12].

표. 1. 유전자 구성의 한 예
Table. 1. Organization of chromosome on a data set

	유전자 구성	변환값
은행	1100100011010111	0.51415
보험	0011100101011100	0.14984
증권	0100011100010101	0.18314

본 논문에서의 유전자의 구성은 표 1과 같다. GA에서 가장 중요한 점 중의 하나는 유전자의 데이터 표현을 문제의 특성을 최대한 반영하여 목적함수에 더 적합한 해를 산출하도록 설정하는 것이다. 따라서, 좀 더 자세하게 하기 위하여 유전자의 구성을 은행, 보험, 증권 각각 16바이트씩 2진수 방식으로 설정하였다. 생성된 유전자를 10진수로 변형한 뒤 0.00001의 값을 곱하여 소수점 5자리의 변환값을 가짐으로써 좀 더 정확한 예측을 하도록 구성한 것이다.

이렇게 생성한 유전자에서 재생산을 통하여 우성인자들을 선택한다. 재생산이란 적합성 함수에 따라 유전자들을 복제하는 과정으로서, 개체군내에서 적합성이 큰 개체에게 더 많은 기회를 주고 적합성이 약한 개체는 퇴화시키는 과정이다. 이때 적합성이 약한 개체를 무조건 삭제하게 된다면 최적해에 도달하기 힘들게 될 수도 있다. 열성유전자도 교배와 돌연변이를 통해 최적해에 도달할 수 있기 때문이다. 따라서, 적합성이 큰 개체를 작은 개체보다 교배와 돌연변이 과정에 확률적으로 좀 더 많이 선택되도록 하는 룰렛 방식을 사용한다. 이러한 연산을 통하여 살아남은 개체군들은 교배와 돌연변이를 하게 된다.

그림 3과 같은 단순 교배를 적용하였는데, 단순 교배란 새로운 유전자 하나를 만들기 위해서 다른 유전자 두 개(P1, P2)와 임의에 교배 포인트를 선택하여 교배 포인트 이전의 P1 유전자, 이후의 P2 유전자를 합쳐서 새로운 유전자 C1, C2를 생성하는 교배를 말한다.

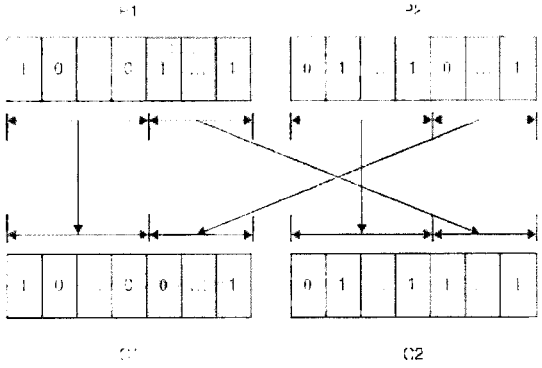

그림 3. 단순 교배
Fig. 3. Crossover Operation

또다른 GA 명령 중에 하나인 돌연변이는 유전자의 한 바이트를 변형시키는 것이다. 일반적으로 돌연변이의 확률을 너무 높게 설정하면 최적해에 도달하기 어렵기 때문에 전체 유전자의 10%만을 적용하였다.

3.4 목적함수

우리는 먼저 연관된 항목끼리의 관계를 해결하는 방법으로 선형 결합을 선택하였다. 식 (1)인 선형 결합은 벡터 공간에서 몇 개의 벡터에 대해, 스칼라 배의 합으로 구하여 얻어지는 원소를 말한다.

$$x = c_1x_1 + c_2x_2 + \ldots + c_kx_k (c_1, c_2, \ldots, c_k \in R) \quad (1)$$

c_1, c_2, c_3을 은행, 증권, 보험 주가지수로 사용하고 선형 공간의 원소 $x_1, x_2, \cdots, x_k$는 임의로 생성되어진 유전자 i의 $\alpha_i, \beta_i, \gamma_i$로 정의한다. $p(i)$를 예측 금융업 지수로 사용하기로 하고 $0<\alpha,\beta,\gamma,<1$의 범위 안에서 (2)와 같은 함수가 정의된다.

$$p(i) = \alpha_i \times c_1 + \beta_i \times c_2 + \gamma_i \times c_3 \quad (2)$$

이렇게 구한 예측 금융업 지수와 실제 금융업 주가지수를 사용해 예측확률을 구하여 $\alpha_i, \beta_i, \gamma_i$의 최적 해를 구하는 것이 연구의 목적이다. 식 (3)는 적합성 함수 $fitness(i)$로서 실제 금융업 주가지수 f_d와 예측 주가지수 $p(i)$의 차이를 더한 함수로 정의한다.

$$fitness(i) = \sum_{d=1}^{o} (p(i) - f_d) \quad (3)$$

목적 함수인 예측확률은 모든 유전자의 적합성 값 중에서 가장 작은 유전자 v를 선택한 뒤 그 유전자의 예측 값과 실제 주가지수를 비교하여 올바르게 예측한 날짜 k를 1년 중 평일의 날짜인 o로 나눈 값이다.

$$k \begin{cases} 0 & If\ d=0 \\ ++ & If\ f_d - f_{d+1} > 0 \text{ and } f_d - p(v) > 0 \\ & If\ f_d - f_{d+1} < 0 \text{ and } f_d - p(v) < 0 \end{cases} \quad (4)$$

$$objective\ function = k/o \quad (5)$$

4. 시뮬레이션

2004년의 주가지수를 사용하여 Generation을 100번 수행한 뒤 예측 금융업지수와 실제 금융업지수의 차이가 가장 작은 최적해로 예측확률을 구하여 보았다. 여기서 임의의 선택(Random Select)이란, 임의로 주가지수가 "상승", "하향" 한다고 가정하였을 때의 예측확률이다.

먼저 정규화전의 지수를 사용 하였을 때의 예측확률은 그림 4처럼 어느 한곳으로 수렴하지도 않았을 뿐만 아니라 예측확률도 높지 않아서 원하는 결과에 근접하지 않았다.

그러나 정규화의 도입으로 그림 5와 같이 예측확률의 상승을 높이고, GA의 특성중의 하나인 최적해 수렴에 근접하게 되었다.

그림 5는 정규화 된 2004년 주가지수를 사용한 선형 결합을 이용하여 제한된 기법으로 각각 예측확률을 구한 그래프이다. 제안된 방법과는 달리 임의 선택의 결과는 예측확률이 현저히 떨어지고 불규칙하게 나타났다.

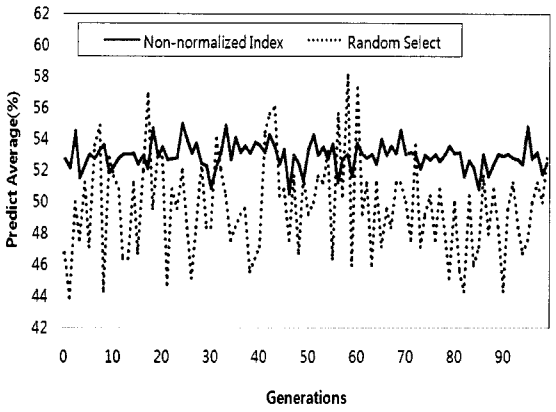

그림 4. 비정규화 주가지수로 구한 예측확률
Fig. 4. Financial Predict Average before Normalized Index

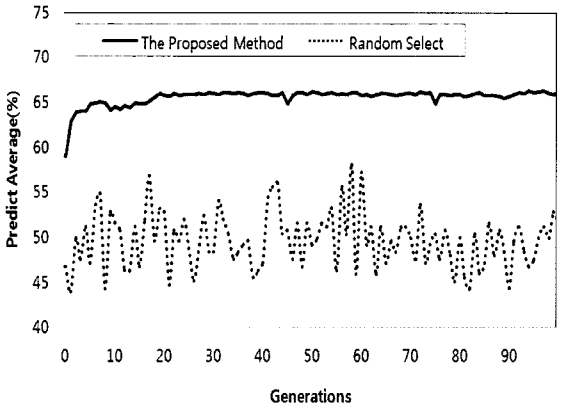

그림 5. 정규화 된 주가지수로 구한 예측확률
Fig. 5. Normalize Financial Predict Average

이렇게 구한 예측 방법으로 2005년부터 2007년까지의 모의투자를 해보았다. 정의 (6)은 투자 규칙으로서 예측지수가 전날 금융업주가지수보다 높을 경우 투자를 하고 작을 경우에는 투자를 하지 않는 것으로 정하였다. 정의 (7)는 투자금액 m이 주가지수가 증가한 비율인 r만큼 손해 또는 이득을 본다는 규칙으로 정하였다.

$$\begin{aligned} Rule: & \; If\, p(v) > f_{d-1}\, then\; Buy \\ & If\, p(v) < f_{d-1} then\; Sell \end{aligned} \tag{6}$$

$$\begin{aligned} Rule: & \; If\; p(v) > f_d \text{ and } Buy\; then\; m = m + m \times r \\ & If\; p(v) < f_d \text{ and } Buy\; then\; m = m - m \times r \end{aligned} \tag{7}$$

정의 (6)와 (7)의 규칙을 적용하여 모의투자 프로그램을 시뮬레이션 한 결과 그래프는 그림 6, 7, 8과 같다. 임의 선택 투자 평균(Random Investment Average)이란 지수의 "상향", "하향"을 임의로 예측한 10번의 모의투자 평균을 구한 것이다.

2005년에는 금융업 주가지수가 87% 증가할 동안 투자금액이 50% 증가하였다. 또한 주가지수와 투자금액의 증감비율의 곡선그래프가 비슷한 성향을 나타내는 것을 알 수 있다. 임의로 선택 투자하여 투자하였을 때는 투자금액이 증가하였지만 주가지수가 증가한 비율에 비해서는 적게 증가하였다.

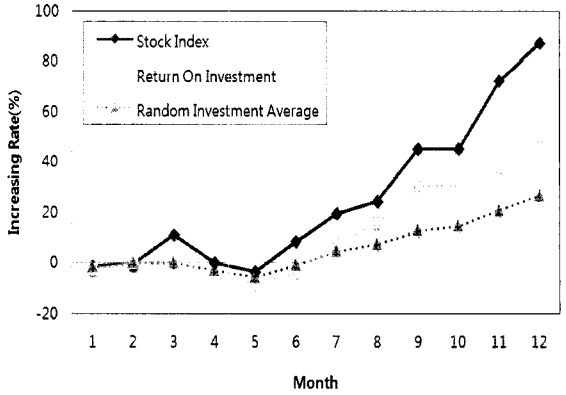

그림 6. 2005년 주가지수와 모의투자금액의 증감비율
Fig. 6. Financial Stock Index, Investment Money Return Rate for 2005 years

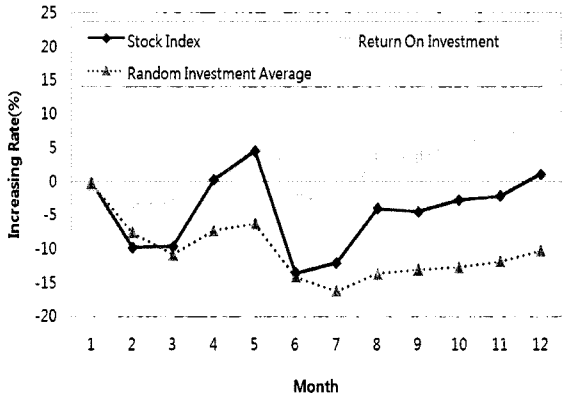

그림 7. 2006년 주가지수와 모의투자금액의 증감비율
Fig. 7. Financial Stock Index, Investment Money Return Rate for 2006 years

2006년에는 금융업 주가지수가 1% 증가할 동안 투자금액은 8% 증가하였다. 이 시기의 전체 주가지수 또한 1400point에서 1440point로 2% 밖에 증가하지 못하였다. 실제 주가지수가 오르지 못한 만큼 투자금액 또한 그다지 증가하지 못한걸 알 수 있었지만, 임의의 선택투자 결과인 투자금액이 감소한 현상보다는 좋은 결과를 나타내었다.

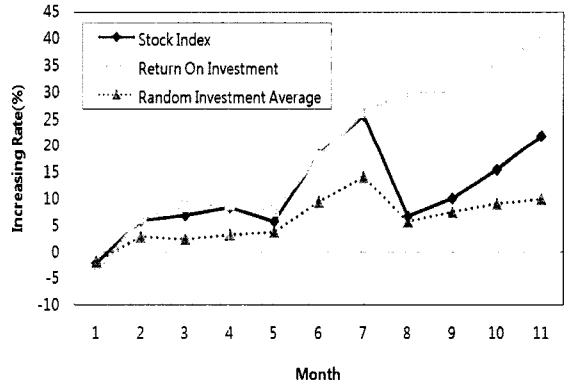

그림 8. 2007년 주가지수와 모의투자금액의 증감비율
Fig. 8. Financial Stock Index, Investment Money Return Rate for 2007 years

2007년에는, 금융업 주가지수가 21% 증가하는 동안 투자금액은 40% 증가하였다. 임의로 투자하였을 때는 투자금액이 증가하긴 하였으나 제안된 방법을 이용해서 예측한 투자 시뮬레이션보다는 낮은 증가율을 보였다.

위의 결과처럼 각 종목들 간의 관계를 가지고 선형 결합의 최적 해를 구하면 투자금액의 상승을 목적으로 하는 예측을 할 수 있게 된다.

5. 결 론

우리는 이 논문에서 주가지수들의 관계를 GA를 사용하여 구한 뒤 주식을 예측하여 2005년부터 2007년까지 주식 모의투자를 해보았다. 2005년에는 주가지수가 상당히 증가하여 투자금액의 증가율이 이에 따라가지는 못하였지만 투자금액이 증가함을 보였고 2006년에는 주가지수가 거의 증가하지 못했음에도 불구하고 투자이익을 보였다. 2007년의 주가지수의 흐름이 비교적 정상적이었고 이에 따른 투자 금액의 증가율이 매우 고무적이었다. 결과적으로 2005년부터 2007년까지의 금융업 주가지수가 260% 증가할 동안 모의투자금액이 230% 증가한 결과를 확인하였다. 이는 일반적 투자인 임의로 결정한 투자로 인해서 얻은 121%의 증가율보다는 훨씬 높은 수치로서 우리의 연구가 효과적이라는 것을 보여준다. 앞으로 점차 종목을 확대해서 각 종목들 간의 관계를 구하는 방향과 종목이 아닌 회사의 주가지수를 가지고 예측하는 방향으로 연구할 것이다. 또한 각 항목별 주식 변동 상황을 분류자로 정하여 이에 따른 규칙을 학습시키는 분류자 시스템을 사용하여 예측하는 방향도 연구 할 것이다.

참 고 문 헌

[1] 김양우, 이긍희, "새로운 연간거시계량경제모형 -BOKAM97," *경제분석*, 제 4권 제 1호, 한국은행, 1998. 3.

[2] 김양우, 장동구, 이긍희, "우리 나라 거시계량경제모형-BOK97," *경제분석*, 제 3권 제 2호, 한국은행, 1997. 5.

[3] F. Craig. "The Treynor Capital Asset Pricing Model," *Journal of Investment Management*, Vol. 1, No. 2, pp. 60-72, 2003.

[4] R. Stephen, "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, pp. 341-360, 1976.

[5] L. Li, C. Longbing, J. Wang, Z. Chengqi, "The Applications of Genetic Algorithms in Stock Market Data Mining Optimization," *The 5th International Conference on Data Mining, Text Mining and their Business Applications*, pp. 273-280, 2004.

[6] F. Fama, "Efficient Capital Market : A Review of Theory and Empirical Work," *Journal of Finance,* Vol.25, pp. 383-417, 1970. 5.

[7] A. Almanza, E. Tsang, "Forecasting stock prices Genetic Programming and Chance Discovery," *In 12th International Conference On Computing In Economics And Finance*, pp. 489, 2006. 1.

[8] K. Lertwachara. "Selecting Stocks Using a Genetic Algorithm : A Case if Real Estate Investment Trusts," *International Journal of The Computer, the Internet and Management,* Vol. 15, pp. 20-31, 2007. 5.

[9] M. Irwin, W. Tony, S. Thiagas, L. Lisa, "Volatility Forecast by Discrete Stochastic Optimization and Genetic Algorithms,", *IEEE International Conference on systems*, pp. 5824 - 5829, 2004.

[10] D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, A Bradford Book The Mit Press Cambridge, pp. 367-398. 2001.

[11] A. Berk, *Regression Analysis : A Constructive Critique*, Sage Publications, 2004.

[12] G. David, *Genetic Algorithms in Search Optimization & Machine Learning*, Addison-Wesley Publishing Company. pp. 1-23, 1989

저 자 소 개

김상호(Sangho Kim)
2002년 ~ 현재 : 세종대학교 컴퓨터공학과

관심분야 : 유전자알고리즘, 분류자 시스템, 임베디드 시스템
Phone : 02-3408-3667
Fax : 02-3408-4339
E-mail : kimsangho@sju.ac.kr

김동현(Donghyun Kim)
1983년 : 연세대 건축공학과 졸업
1991년 : 일본 오사카대 환경공학과 석사
1998년 : 일본 오사카대 환경공학과 박사

관심분야 : 디지털 콘텐츠
Phone : 02-3408-3795
Fax : 02-3408-4339
E-mail : mustache@sejong.ac.kr

한창희(Changhee Han)
1990년 : 육군사관학교 물리학과 졸업
1994년 : 미국 시라큐스대 전산학과 석사
2004년 : 미국 남가주대 전산학과 박사

관심분야 : 멀티메디아 컨텐츠, 가상 휴먼 모델링과 시뮬레이션, 인공지능
Phone : 02-2197-2882
Fax : 02-972-8179
E-mail : chhan@kma.ac.kr

김원일(Wonil Kim)
1982년 : 한양대 금속공학과 졸업.
1987년 : 미국 남일리노이대 전산학과 졸업
1990년 : 미국 남일리노이대 전산학과 석사
2000년 : 미국 시라큐스대 전산정보학과 박사

관심분야 : 인공지능, 디지털콘텐츠, 컴퓨터 보안, 유비쿼터스
Phone : 02-3408-3795
Fax : 02-3408-4339
E-mail : wikim@sejong.ac.kr