「정보시스템연구」 제26권 제3호
한국정보시스템학회
2017년 9월, pp. 63~90

http://dx.doi.org/10.5859/KAIS.2017.26.3.63

절대 유사 임계값 기반 사례기반추론과 유전자 알고리즘을 활용한 시스템 트레이딩

한현웅* · 안현철**

〈목 차〉

Ⅰ. 서 론

다양한 예측 분야 중에서 주가지수 예측과 같은 시계열 예측은 가장 복잡하고 난해한 것으로 알려져 있다. 특히, Fama(1970)와 Malkiel(1999)의 연구에서 효율적인 시장은 예측이 불가능하다는 주장이 제기되었으며 Elton & Gruber(1984)는 과거 발생한 주가의 결합 기법이 무수히 많기 때문에 제한된 특정 기법을 활용하여 시장에서 효율적인 기법임을 주장하는 것은 불가능하다고 하였다. 이로 인해 오랜 기간 동안 정확한 시계열 예측이 어려운 것으로 여겨졌다. 그러나 Lo & Mackinlay(1988), Fuller & King (1990), Brock *et al.*(1992) 등의 선행 연구에서 시계열 예측이 어느 정도 가능하다는 주장을 제기하였다. 선행 연구에 이어 시계열 예측을 위해 많은 연구들이 이루어 졌으나 정확한 예측은 지금까지도 어려운 문제로 인식되고 있다. 이러한 시계열 예측에 있어서 어려운 원인 중 하나는 기존의 선형 모형들이 시계열의 움직임을 제대로 예측하지 못한다는 약점이었다. 최근 연구에서는 시계열의 비선형성을 극복하여 보다 정확한 예측 모형을 구축하기 위하여 인공신경망(artificial neural network; ANN)과 자기회귀 이분산모형(autoregressive conditional heteroscedasticity; ARCH), 그리고

* 국민대학교 비즈니스IT전문대학원 박사과정, hyunwoong74@nate.com(주저자)
** 국민대학교 비즈니스IT전문대학원 부교수, hcahn@kookmin.ac.kr(교신저자)

사례기반추론(case-based reasoning; CBR) 등을 활용하는 연구가 이루어지고 있다. 이 중 인공신경망과 자기회귀 이분산모형은 비교적 많은 연구가 되고 있으나 상대적으로 k-Nearest Neighbor Algorithm(k-NN)과 같은 사례기반추론은 시계열 예측에서 활발하게 활용되지 못하였다(Chun & Park, 2005; Donaldson & Kamstra, 1997; Kim & Han, 2000; Kim & Han, 2001; Kim & Lee, 2004; Poon & Taylor, 1992; Silvapulle & Choi, 1999).

사례기반추론은 의사 결정 수행을 위하여 논리적인 의사결정 과정을 모형화하여 문제를 해결하는 기법이다. 사례기반추론은 비구조화 되고 복잡한 의사결정 문제에 특별히 잘 응용 할 수 있는 장점이 있어서 다양한 분야의 의사결정문제에 활용되어 왔다. 그러나 많은 장점을 가지고 있음에도 불구하고 효과적인 사례기반추론 모형을 설계하고 구축하기 위해서는 연구자가 해결해야 하는 여러 문제들이 존재한다. 여러 문제 중에서도 사례기반추론은 적절한 유사도 측정 기법과 유사 사례 선택 방법, 유사사례들의 결합 방법 등과 같은 문제에 대하여 보편화되고 긍정적으로 평가된 방법론이나 원리를 제공해 주지 못하고 있는 실정이다. 이러한 점 때문에 사례기반추론 모형의 다양한 설계요소들을 결정하는데 있어서 실험자 혹은 사용자가 자신의 경험이나 직관으로 해야 하는 문제가 있었다. 이러한 어려움은 주가지수 예측과 같은 정교한 모형을 요구하는 문제 상황에서 부정확한 예측성과를 제시하기도 했었다(Kim, 2004). 이에 지금까지 다수의 연구자들에 의해 최적의 사례 간 유사도 측정 방법이나 특정 사례를 대표하는 변수군의 최적 선택 방법, 근접 사례 결합 시 적용되는 가중치의 값을 최적화 하는 방법 등이 연구되어 왔다(Chiu et al., 2003; Kim & Han, 2001; Shin & Han, 1999; Wang & Ishii, 1997).

이러한 사례기반추론의 최적화 연구들 중에서 이훈영과 박기남(1999), i Guiu *et al*.(1999), Jarmulak *et al*.(2000) 등은 사례기반추론 모형에서 성과를 향상시키기 위해서 결합되는 유사 이웃 사례의 수를 최적화 하는 것이 성능 제고에 지대한 영향을 미칠 수 있음을 제시하였다. 특히 안현철(2013)은 절대값 방식의 임계값에 기반하여 사례기반추론의 유사 이웃사례를 선택할 경우, 획기적으로 사례기반추론의 예측정확도가 개선될 수 있음을 제시한 바 있다.

이에 본 연구에서는 기존 연구에서 제안된 절대 유사 임계값에 기반한 사례기반추론 모형을 주식시장 예측에 적용하여, 투자 수익률을 효과적으로 제고시킬 수 있는 시스템 트레이딩(system trading) 기법을 제안한다. 구체적으로 본 연구의 제안 기법은 사례기반추론이 확실한 매수(buy) 혹은 매도(sell) 신호가 발생하였을 경우에만 매매 거래를 하도록 하고, 익일 주식시장의 방향이 불확실한 경우 의사 결정을 보류(hold)하도록 설계되어, 거래 건수를 줄이면서 동시에 투자수익은 높일 수 있도록 설계되었다. 아울러, 본 연구에서는 사례기반추론의 성능 제고와 밀접한 관련이 있는 것으로 알려진 최적의 특징변수 선택(feature selection)과 최적의 학습사례 선택(instance selection)도 함께 고려하였으며, 이러한 특징변수 및 학습사례 선택 그리고 절대 유사 임계값을 동시에 최적화하기 위한 기법으로 유전자 알고리즘(Genetic Algorithm; GA)을 사용하였다. 실제

로 제안된 기법이 유의미한 수익을 발생시키는지 검증하기 위해, 본 연구는 2009년부터 2016년 사이의 KOSPI 200 데이터를 활용하여 연구 모형의 실무적 활용 가능성을 검증하였다.

기존에 인공지능 혹은 데이터마이닝 기법을 활용해 주식시장의 등락을 예측하는 연구들의 경우, 주식시장의 등락을 얼마나 정확하게 예측하는가에 초점을 맞추어 개발되므로 실제 투자 수익과는 다소 괴리가 발생하는 경우도 종종 발생하였다(이재식 외, 2000; 안현철과 이형용, 2009; 김선웅과 안현철, 2010). 하지만 앞서 설명했듯이, 본 연구의 제안 기법은 주식시장의 등락 여부에 대한 정확한 예측이 아닌 투자 수익률 자체를 극대화할 수 있도록 설계되어 있어, 보다 효과적인 시스템 트레이딩을 구현하는데 실무적으로 기여할 수 있을 것으로 예상된다.

이후 본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 연구와 관련된 이론 및 과거 문헌들을 고찰해 보고, 3장에서는 유전자 알고리즘을 이용해 절대 유사 임계값와 특징변수, 그리고 참조사례 선택을 수행하도록 설계된 본 연구의 제안 모형을 소개한다. 4장에서는 제안 모형에서 사용된 실험 데이터 설명과 모형의 유용성을 확인하기 위한 실험 설계 및 실험 결과가 소개되며, 마지막 5장에서는 본 연구의 결론과 의의를 정리하고 한계점을 제시하고자 한다.

Ⅱ. 이론적 배경

본 연구에서는 사례기반추론과 유전자 알고리즘을 결합시킨 새로운 시스템 트레이딩 기법을 제시하였다. 이에 본 장에서는 우선 시스템 트레이딩과 관련한 기존 문헌들을 고찰하고, 이어 제안기법의 근간을 이루는 사례기반추론과 사례기반추론의 성과를 개선하기 위하여 시도된 다양한 선행 연구들을 살펴본다. 끝으로 사례기반추론 모형의 여러 설계 요소들을 최적화하기 위해 사용된 유전자 알고리즘의 기본적인 개념에 대해 살펴본다.

2.1 시스템 트레이딩

시스템 트레이딩이란 체계적인 거래(systematic trading)를 위해 사용되는 거래 전략을 말한다. 시스템 트레이딩에서는 거래자 개인의 편견 및 자의적 판단 등을 배제하고 순수하게 기술적 분석만으로 거래 전략을 도출하며, 이를 과거 일정 기간 동안의 시장가격 자료를 바탕으로 실험하여 여러 가지 방법으로 평가한 후, 평가 된 시스템 트레이딩에서 발생하는 매수 혹은 매도 신호 등에 전적으로 의존하여 거래가 이루어지게 된다(Vince, 1990).

이러한 시스템 트레이딩의 장점은 다음과 같이 크게 두 가지로 나누어 볼 수 있다. 첫째는 실제 거래에 있어서 수익 실현의 가능성에 대한 거래에 대하여 실질적인 자본을 투입하지 않고 손익을 미리 실험을 통하여 검증해 볼 수 있다는 것이며, 둘째는 발생할 수 있는 위험과 잠재적인 수익에 대하여 검증을 함으로써, 시스템 트레이딩에 대한 신뢰도를 높일 수 있다는 것이다.

시스템 트레이딩에 관한 연구는 실무 현장에서의 높은 관심을 받는 반면에 학계에서는 큰 주목을 받지 못하였다. 1960년대에 정립되기

시작하여 큰 호응을 얻기 시작한 효율적 시장 가설이 있었기 때문이다. Alexander(1964)는 주가가 저점에서 일정 비율까지 오르면 매수하고 반대로 고점에서 일정 비율까지 하락하게 되면 매도하는 필터 기법(filter rule)을 활용한 결과로 투자금액 대비 추가 수익을 내기 어렵다는 것을 증명하였다. Fama(1965)는 주가의 런 검정(runs test)과 상관관계분석(serial correlation)을 활용하여 주식시장이 효율적임을 주장하였다. 이어지는 효율적 시장 가설에 대한 연구들의 결과는 주로 가설을 지지한 결론을 보여주고 있다.

그러나 Granger(1981)의 연구에 의해서 주가와 같은 변화가 많은 시계열자료 분석에 선형 회귀분석을 활용하는 것이 적절치 않음이 밝혀진 이후로는 인과관계분석(causality) 기법이 등장하여 계량경제학적 주가예측모형에 대한 연구에 활발히 활용되어지기 시작하였다. Caporale & Pittis(1998)는 공적분 검정(co-integration test)을 활용하여 시장에서 부분적으로는 주가를 예측할 수 있다는 것을 증명하였다. 그리고 McMillan(2007)은 영국 외 3개 국가의 주식시장 거래량을 투입 변수로 입력하여 비선형 회귀모형을 적용한 결과 선형모형보다 보다 좋은 투자성과를 얻을 수 있다는 것을 보여주었다.

최근에는 주가 예측에 인공지능기법을 활용한 연구가 활발하게 진행되고 있다. 주가 예측을 위한 입력 변수로는 기술적 분석(technical analysis)과 기본적 분석(fundamental analysis)의 지표들이 주로 사용되고 있다. 여기서 기술적 지표는 과거의 주가나 거래량 데이터를 활용하여 이동평균(Moving Average; MA), 이동평균 수렴·확산지수(Moving Average Convergence and Divergence; MACD), 스토캐스틱(stochastic), 모멘텀(momentum) 등으로 변환하여 생성한 지표들이고, 기본적 지표는 주가수익비율(Price/Earnings Ratio; PER), 경제성장율(Rate of Economic growth), 환율, 금리, 기업의 부채비율, 배당률 등과 같은 주가에 직접적인 영향을 미치는 경제 지표 변수들이다. 기술적 지표는 주가 데이터에서 쉽게 산출해 낼 수 있는데 단기적으로 주가의 움직임을 주시하고 포착하는 데 적절한 지표이며, 오래 전부터 시장 참여자들에 의하여 투자에 실제로 많이 활용되고 있다. 반면 기본적 분석은 주가를 결정하는데 있어서 본질적 정보를 활용하는 분석 기법인데, 기본적 분석을 통해 산출되는 지표 변수들은 주가에 장기적으로 영향을 주고 있기 때문에 단기적인 거래를 지향하는 시스템 트레이딩에서 활용하기에는 대체로 적절하지 못하다.

한편, 목표 함수는 통계적인 기준과 비통계적인 기준으로 나누어지는데, 예측값의 정확도를 평가하기 위한 통계적인 기준으로는 평균제곱근편차(Root Mean Square Error; RMSE)와 평균절대편차(Mean Absolute Error; MAE) 등과 같은 지표들이 활용되며, 비통계적 기준으로는 적중률(hit ratio)이나 수익률(rate of return)이 주로 사용된다.

주가지수 예측과 관련된 연구들을 살펴보면, Yudong & Lenan(2009)은 BP-ANN을 활용하여 S&P 500 주가지수를 예측하였다. 이 연구에서 입력 변수로는 S&P 500 주가 지수의 기술지표 10개를 활용하였으며, 예측값의 목표 함수로 평균제곱오차(Mean-Squared Error; MSE)

를 활용하였다. Schulmeister(2009)는 기술적 지표 2,580개를 활용하여 S&P 500 주가를 분석하여 거래별 수익률을 단순하게 덧셈하는 총수익 관점에서 주가 예측 모형을 제안하였는데, 이 연구의 결과로 1990년대에 들어서면서 기술적 지표의 수익률이 하락하는 것을 확인하였다. Atsalakis & Valavanis(2009a)는 입력변수로 주가의 5일 동안의 이동평균 값을 투입하는 Neuro-Fuzzy 모형을 활용하여 익일의 상승 및 하락을 예측하였으며, 그 결과를 적중률과 단순 수익률 기준으로 선행 연구들의 예측 방법들과 비교 분석한 결과, 제안 모형의 성과가 우수함을 확인 할 수 있었다.

시스템 트레이딩과 관련한 연구로는 대표적으로 Nunez-Letamendia(2007), Bao & Yang(2008), Chavarnakul & Enke (2009), 안현철과 이형용(2009), 김선웅과 안현철(2010) 등의 연구가 있다. 우선 Nunez-Letamendia(2007)는 유전자 알고리즘을 활용하여 기술적 시스템 트레이딩의 최적화를 시도하였다. 그러나 입력 변수로 가장 널리 알려진 기술적 지표인 두 이동평균선의 교차만을 활용하였으며, 적합도 함수(fitness function)로 누적 수익률함수를 사용하였지만, 연구의 주목적이 최적의 기술적 규칙을 찾아내는 목적보다는 통제 파라미터 값에 대한 유전자 알고리즘의 강인성 검정(robustness test) 중심의 연구가 되었다는 한계가 있다. 또한, 이 연구에서 제안된 시스템은 매일 지속적인 매수 혹은 매도신호를 발생하도록 설계되어 있다. 이러한 이유로 실제 투자에 활용하는데 있어서, 다소 무리가 있다는 한계가 있다. Bao & Yang(2008)은 고차원 표현(high-level representation)과 확률 모형을 결합하는 인공지능 시스템 트레이딩을 개발하였다. 입력 변수로는 과거의 주가 데이터와 4개의 기술적 지표를 활용하였는데, 비가격 변수에 대해서는 전혀 고려하지 않았다는 한계가 있다.

Chavarnakul & Enke(2009)는 인공신경망과 퍼지 로직, 유전자 알고리즘을 활용한 시스템 트레이딩을 제안하였다. 이들의 연구에서 입력변수는 과거의 주가 데이터와 거래량 데이터를 결합하여 활용하였고, 매수-매도 신호에도 매매가 이루어지지 않는 중립 지대를 두었지만 중립 지대를 결정하는 임계값의 범위를 +0.5와 -0.5로 비교적 큰 값으로 주었으며, 이 값을 주관적으로 결정하여 사용하였다.

안현철과 이형용(2009)은 이중 임계값 모형을 사용하여 익일의 주가패턴이 불확실 할 경우 매매를 보류 가능하도록 중립 지대를 설정하여 실제 투자에서 유용하게 사용할 수 있도록 제안하였다. 이 연구 역시 입력변수로 12개의 기술적 지표를 활용하고 있는데, Bao & Yang(2008)의 연구와 같이 비가격 변수에 대하여 전혀 고려하지 않은 한계가 존재한다.

Atsalakis & Valavanis(2009b)는 주식시장 예측에 다양한 인공지능기법을 사용한 100편 이상의 연구 논문을 조사하여 성과측정방법(performance measures), 예측방법(forecasting methodology), 입력변수(input variables) 등의 기준으로 분류작업을 시도하였다. 여러 논문에서 분석대상이 되는 주식시장은 미국 뿐만 아니라 아시아와 유럽, 남미까지 24개 이상의 주식시장으로 골고루 분포하고 있다. 입력변수의 수는 대부분 4~10개 사이이지만 2개를 사용한 연구부터 최대 61개를 사용한 연구까지 다양하게 구성하고 있다. 입력 변수로는 30% 이상이

주가지수의 과거 종가나 주가 데이터를 사용하고 있었다. 가장 많이 활용된 예측 방법은 인공신경망이었다. 전체 분석 대상의 90% 이상이 인공신경망이나 인공신경망의 변형 모형을 활용한 것으로 나타났다.

김선웅과 안현철(2010)의 연구는 Support Vector Machines(SVM)에 유전자 알고리즘을 결합한 시스템 트레이딩 모형을 제안하였다. 이 연구에서는 유전자 알고리즘을 활용하여 이중 임계값을 최적화하였으며, 비가격 변수들을 실험에 투입하여 비슷한 조건의 기술적 지표만 활용한 기존 연구들 보다 등락 예측에 있어서 개선이 이루어짐을 확인하였다. 그러나 한계점으로 학습 데이터에 적용한 무작위 추출로 인해 지수의 상승과 하락 사례의 균형화가 시계열 데이터를 왜곡시킬 가능성이 있다는 점을 지적할 수 있다.

송성환 외(2016)는 자연언어처리를 이용하여 소셜 빅데이터를 활용하여 감성을 추출하고 통계분석을 하여 추출된 감성의 흐름과 관련이 깊은 주식을 발굴하여 발굴된 종목들를 월별로 기계학습 및 평균회귀전략을 기반으로 주가예측을 시도하였으며 자산배분 모형으로는 Black-Litterman모형을 이용한 연구를 시도하였다. 이러한 모델을 적용하여 2011년 1월부터 2014년 10월까지 시뮬레이션을 실시한 결과, 코스피 수익률 대비 약 20%를 초과하는 성과를 확인하였다. 그리고 2015년 1월부터 2015년 8월까지 실전투자 결과에서 코스피 수익률 대비 12%를 초과하는 실적을 보이는 결과를 도출하였다. 이 연구를 통하여 빅데이터 분석을 활용한 트레이딩이 실제 시장지표 대비 높은 수익률을 보일 수 있는 가능성을 확인하였다.

최근에는 딥러닝(deep learning)을 주식시장 예측에 활용한 연구들이 소개되고 있다. 딥러닝은 인공 신경망(ANN: artificial neural network)을 기반으로 하는 기계 학습 방법으로 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화(abstraction)를 시도하는 일련의 기계학습 알고리즘으로 정의된다. 딥러닝은 이미지 혹은 동영상 속 사물 인식, 음성 인식 등에서 놀라운 성능을 보여주고 있는데, 최근 이우식(2017), 송유정과 이종우(2017) 등 일부 국내의 학자들이 이와 같은 딥러닝을 주식시장 예측에 적용하는 연구를 시도하고 있다.

이우식(2017)의 연구에서는 기술적 주가분석기법을 딥러닝 모형에 적용하여 코스피지수의 상승 및 하락을 예측하는 연구를 수행하였다. 이 연구에서는 전체 변수를 사용하지 않고 일부의 선택된 변수만으로 코스피 주가지수의 방향성을 예측할 수 있다는 결과를 도출하였다. 그러나 5종의 기술적 지표만을 활용한 딥러닝 기법이 비교 실험 모형인 의사결정나무모형과 SVM의 비교 실험 결과와 비슷한 예측력을 나타내어 미래 주가지수 움직임 예측에 대한 딥러닝의 한계를 보여주었다.

송유정과 이종우(2017)의 연구에서는 일별 시가, 고가, 저가, 종가, 거래량 등의 5가지 간단한 주가 데이터를 이용해 삼성전자 주식의 증감패턴을 고려한 심층신경망(Deep Neural Network, DNN)의 딥러닝 기반 학습 모형을 제시하였다. 실험 결과, 약 56% 정도의 예측 정확도를 갖는 것으로 나타나, 딥러닝이 주가 예측에 어느 정도 성능을 발휘할 수 있음을 확인하였다. 하지만, 이 연구는 단 한 종목만을 대상으로 실험이 진행되었고 기술적 지표가 아닌 단

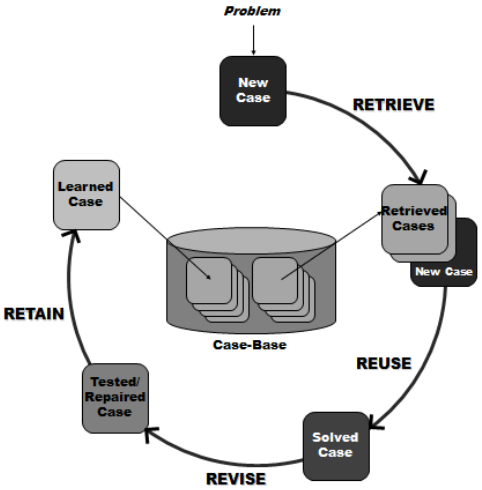
순 지표를 대상으로 실험이 진행되었다는 한계를 가지고 있다.

이상 소개한 시스템 트레이딩 관련 기존 연구들을 종합해 보면, 지금껏 효과적인 트레이딩 전략을 도출하기 위해 기계학습 기법을 응용하고자 한 다수의 시도들이 있었고, 그 중 일부는 유전자 알고리즘을 이용해 최적화를 함께 시도하였음을 확인할 수 있다. 하지만, 그 어떤 연구에서도 본 연구에서 채택한 사례기반추론을 예측을 위한 기본 기계학습 방법론으로 채택하고 있지 않았음을 알 수 있다. 이는 사례기반추론이 비록 많은 장점을 갖고 있지만, 예측 정확도가 좋지 않아 수익을 올리기 어렵다는 결정적인 한계에 기인한 현상인 것으로 풀이된다. 이에 본 연구에서는 사례기반추론의 이 같은 한계를 극복하고, 궁극적으로 수익률 극대화를 도모할 수 있는 새로운 사례기반추론 알고리즘인 절대 유사 임계값 기반 사례기반추론을 소개하고, 이를 유전자 알고리즘의 최적화 기능과 결합하여 본 연구만의 새로운 시스템 트레이딩 방법론을 제안하고자 한다.

2.2 사례기반추론

사례기반추론은 과거 경험이나 사례를 통하여 주어진 문제에 대한 해를 찾아내는 문제 해결 기법이다. 일반적인 인공지능기법은 문제와 해법 사이에서 일반적 관계를 도출한 결과를 기반으로 추론을 하는 원리로 이루어져 있어서 비교적 정형화된 문제 해결에는 적합하지만 구조적으로 지식의 지속적인 갱신이 어려운 한계를 지니고 있다. 그렇지만, 사례기반추론은 과거에 축적된 정보만 있으면, 어떤 문제든 해결이 가능하므로 복잡하거나 비구조화 되는 문제를 해결하는데 유리하며, 지식기반을 지속적으로 업데이트 할 수 있다는 측면에서 상대적으로 우수하다고 할 수 있다(Shin & Han, 1999).

사례기반추론은 다음 <그림 1>에 제시되어 있는 것과 같이, 이른바 4R(RETRIEVE, REUSE, REVISE, RETAIN)이라 불리는 4단계의 절차에 의해 이루어진다(Aamodt & Plaza, 1994).



<그림 1> 사례기반추론의 4R

이 중에서 첫 번째 단계인 RETRIEVE는 사례기반추론의 효과를 결정하는 가장 중요한 단계이다. 이 단계는 모형에서 주어진 문제 해결에 도움이 될 수 있을 것으로 추정되는 여러 사례를 선택하게 되는데, '어떤 방법으로 유사한 사례들을 선별하여, 선별된 사례들의 조합을 어떻게 구성하고, 추천 결과를 생성해 낼 것인가?'에 따라 사례기반추론 모형의 성능을 크게 좌우하기 때문이다. 그렇기 때문에 사례간의 유사

도를 어떻게 측정하고, 추천 결과 도출 시 유사 사례는 어떻게 결합할 것인가 하는 등의 문제는 전통적으로 주요 사례기반추론의 연구 과제로 자리매김 하여 왔다(Chiu, 2002).

입력 사례와 기존 참조사례들 간의 유사도를 측정하는 기법들에는 다양한 기법이 사용되어 왔는데, 그 중에서도 유클리드 거리(Euclidean distance)나 맨하탄 거리(Manhattan distance) 기법이 주로 활용되고 있다. 그리고 유사도 기준에 의해 입력 사례와 근접한 이웃 사례를 찾아내는 방법으로는 Nearest-Neighbor(NN) 기법이 가장 널리 활용되고 있다(Jarmulak *et al.*, 2000; Chiu, 2002).

Nearest-Neighbor 기법 중에서도 가장 유사한 하나의 사례를 찾아서, 그것을 기반으로 해를 찾는 기법을 'One Nearest-Neighbor(1-NN)' 기법이라고 한다. 하지만 일반적으로는 유사 사례를 하나가 아닌 여러 개를 선택하여 내삽법(interpolation)이나 투표(voting) 등의 기법을 활용하여 종합한 결과를 토대로 최종 해를 제시하는 방법이 더 많이 적용되고 있다. 이러한 방법을 'k Nearest-Neighbor(k-NN)' 기법이라고 하는데, 이 때 k는 해를 구하기 위하여 참조할 유사 이웃 사례의 개수를 의미한다. 이 k-NN 기법의 경우, 사례기반추론에 적용한 해가 과거의 여러 사례를 보다 다양하게 반영하여 일반화되고 잡음(noise)에 민감하지 않은 결과를 도출해 낼 수 있다는 면에서 이점이 있다. 하지만 k의 값이 너무 커지게 되는 경우에는 반대로 k-NN 기법이 사례기반추론의 성과에 악영향을 줄 수도 있다. k의 값이 필요 이상으로 커질 경우에는 선택된 유사사례들 중에서 유사하지 않은 사례를 포함할 가능성이 높아지기 때문이다. 이러한 문제 때문에 k-NN 기법에서 최적의 k를 찾아내는 것이 사례기반추론 시스템의 성과를 높이는데 있어서 매우 중요한 요소임을 알 수 있다.

사례 간의 유사도를 보다 정확하게 측정함으로써 사례기반추론의 예측 정확도를 개선하기 위한 또 다른 접근법으로 (1) 적절한 특징변수의 선정(feature selection)과 (2) 적절한 참조사례 선정(instance selection)이 오래 전부터 학자들에 의해 연구되어 왔다. 이 중 특징변수 선정은 유사도 측정과 관련이 높은 특징변수들만 선별하는 전처리 과정을 의미한다(안현철 외, 2005a). 이와 관련해 Siedlecki & Sklanski (1989)는 유전자 알고리즘을, Cardie(1993)는 의사결정나무를 특징변수 선정 방법론으로 제안하였다. 한편 Skalak(1994)과 Domingos (1997)는 각각 힐 클라이밍(hill climbing) 알고리즘과 군집 분석(clustering analysis)을 특징변수 선정을 위한 방법론으로 제안한 바 있다.

한편 참조사례 선정은 전체 사례기반에 저장된 참조사례 중에서 유사도 측정과 관련해 대표성이 높은 사례들만 선별하고, 그렇지 않은 사례는 참조 과정에서 배제하는 전처리 과정을 의미한다(안현철 외, 2005a). 이 분야의 초기 연구로는 Hart(1968)와 Wilson(1972)의 연구가 있는데, 이들은 간단한 정보이득(information gain) 개념에 기반하여 최적의 참조사례를 선정하는 방법론을 제안하였다. 참조사례 선정과 관련한 최신 연구들은 보다 고차원적인 수리적 기법이나 인공지능을 방법론으로 활용하고 있다(안현철 외, 2005a). 예를 들어, Sanchez et al.(1997)는 근접 그래프 접근법을, Lipowezky (1998)는 선형계획법을 이용해 참조사례 선정

을 수행하는 방안을 제안하였다. 인공지능 기법들도 사용되고 있는데, Huang et al.(2002)은 인공신경망을, Babu & Murty(2001)는 유전자 알고리즘을 참조사례 선정 기법으로 제안한 바 있다.

2.3 절대 유사 임계값 기반 사례기반추론

전통적인 사례기반추론 모형들은 모두 유사 이웃 사례 선정 시, 특정 수(k-NN의 k)의 이웃 사례나 유사도의 상대적인 비율에 근거한 이웃 사례를 선택하였다(Sun & Hui, 2006; 박윤주, 2006). 때문에 주어진 문제를 해결하는데 있어 참조하기 적합한 유사 사례가 아님에도 불구하고, 다른 사례들과 비교해 상대적으로 더 근접하다는 이유로 참조 대상에 포함될 수 있는 위험에 노출된다.

이러한 문제를 해결하기 위해 안현철(2013)의 연구에서는 결합사례의 유사도 측정이나 결합 유사 사례의 선택 기준으로 0부터 1사이의 값을 갖는 절대 유사 임계값을 적용하였다. 이 연구에 따르면, 모든 입력변수들에 최소-최대 정규화(min-max normalization)가 적용한 뒤, 유클리드 거리에 기반해 사례 간 유사도를 산출하게 되면, 해당 유사도값은 항상 0~1 사이의 값을 갖게 된다. 때문에 상대적 비율이 아닌 절대값 형태로 유사사례의 기준을 정할 수 있게 되는 것이다. 이처럼 절대 유사 임계값을 기준으로 이웃 사례를 결정하게 되면, 경우에 따라 유사한 이웃 사례가 하나도 나오지 않을 수도 있는데, 이 경우 예측결과를 생성하지 않고 '모름(don't know)'으로 결과를 회신할 수 있도록 하였다.

2.4 유전자 알고리즘(Genetic Algorithm; GA)

유전자 알고리즘은 찰스 다윈(Charles Darwin)의 적자생존의 원리와 멘델(Mendel)의 유전 법칙을 응용한 최적화 기법으로, 생물의 진화 과정을 응용하여 적응적으로 탐색 공간을 탐색하여 최적 또는 근접 최적의 해를 찾아내는 탐색 기법이다(홍승현과 신경식, 2003).

유전자 알고리즘은 점에 대한 탐색이 아닌 개체들이 모여 이루어진 군집에 대하여 병렬적으로 탐색이 된다는 점에서 기존의 최적화 알고리즘과 차별화 된다. 또한 탐색의 방향과 영역이 초기 값에 크게 의존하지 않으며, 세대에 따라 확률적으로 변화된다는 점에서 전역 최적화가 가능한 이점을 가진다(옥중경과 김경재, 2009).

유전자 알고리즘의 작동 프로세스는 다음과 같다. 먼저 본격적인 진화 과정에 앞서 유전자 알고리즘은 해결하고자 하는 문제의 해(solution)를 이진코드(binary code) 형태의 값으로 생성하는데, 이렇게 생성된 하나의 개체를 염색체(chromosome)라고 한다. 유전자 알고리즘은 전체의 탐색 공간 안에서 임의로 n개의 염색체들을 선택하여 이 값들을 계속 진화시켜 나가게 되는데, 이렇게 생성된 염색체들의 집합을 모집단(population)이라고 호칭한다. 유전자 알고리즘의 진화 과정이 본격적으로 시작되기 전에 모집단을 구성하는 모든 염색체들의 값은 임의의 값으로 초기화된다.

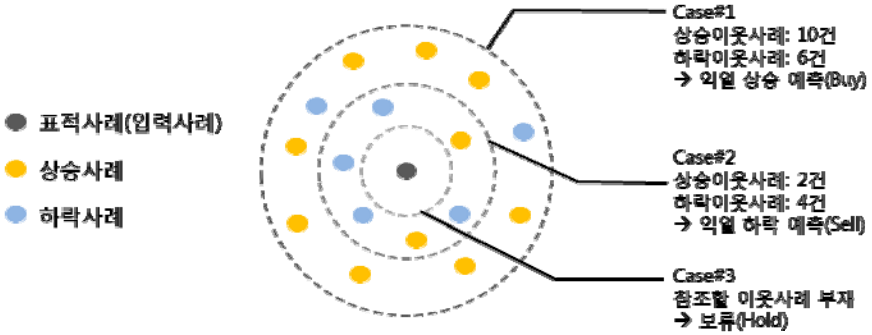
첫 번째 단계의 초기화 작업을 마치고 나면, 다음 단계에서는 생성된 모집단이 문제해결에 있어서 얼마나 적합한지를 평가하기 위하여 이

른바 적합도 함수(fitness function)를 활용하여 각 염색체(개체)의 적합도를 평가하게 된다. 이렇게 각 염색체의 적합도가 평가된 후 유전자 알고리즘은 평가된 염색체 집단을 확률적으로 선택(selection) 및 교배(crossover), 돌연변이(mutation) 등의 유전적인 조작을 실행하여 이전 세대에서 진화된 새로운 염색체들로 구성된 세대를 생성하게 된다. 이렇게 생성된 새로운 세대의 염색체 집단은 다시 적합도 함수에 의해 재평가되며, 그 결과에 의하여 다시 유전자 조작이 이루어지게 된다. 이러한 평가와 유전자 조작 작업은 목표로 하는 적합도의 수준이 도출되거나 사전에 정해진 최대 진화 수와 같은 종결 조건이 될 때까지 반복적으로 이루어지게 된다. 이렇게 유전자 알고리즘은 생성된 전체 세대 중에서 가장 최적의 적합도를 나타낸 개체를 최종 선택하여 그 결과를 전역 또는 유사 전역 최적해로 도출한다(안현철 외, 2005a).

이러한 유전자 알고리즘은 여러 이점으로 인해 지금까지의 선행된 많은 연구에서 인공지능 기법의 설계 요소를 최적화하기 위하여 적용되어 왔다(안현철 외, 2005a; Chiu, 2002; Chiu *et al.*, 2003; Kim & Han, 2001; 2003; Shin & Han, 1999; Siedlecki & Sklansky, 1989 등). 또한 시스템 트레이딩의 투자의사결정 시 참조되는 임계값을 최적화하기 위한 용도로 사용되기도 하였다(이형용, 2008; 안현철과 이형용, 2009). 본 연구에서도 절대 유사 임계값과 참조 사례의 선택, 그리고 특징변수의 선택을 위한 최적화 도구로 유전자 알고리즘을 적용한다.

Ⅲ. 제안 기법

본 연구에서는 안현철(2013)이 제안한 절대 유사 임계값 기반 사례기반추론에 기반한 시스템 트레이딩 기법을 제안한다. 시스템 트레이딩에 절대 유사 임계값을 기준으로 이웃 사례를 선택하는 사례기반추론을 적용하게 될 경우, <그림 2>와 같이 유사 이웃사례가 하나도 탐색되지 않을 경우에는 예측 결과를 도출하지 않고 '보류(Hold)'로 결과를 도출 할 수 있다. 즉, 주식시장을 예측하고자 시도했던 대부분의 선행 연구들처럼 '매수(Buy)' 또는 '매도(Sell)'의 두 가지 상태로만 결과를 예측하는 것이 아니라, '모름(Hold)'을 포함한 3가지 경우로 결과를 도출하기 때문에, 보다 효과적으로 투자의사



<그림 2> 제안 기법의 투자의사결정 판정 원리

결정을 내릴 수 있다.

이와 함께 사례기반추론의 성과에 가장 큰 영향을 미치는 요소들로 제시되어 온 최적 참조사례의 선택, 그리고 최적 특징변수의 선택도 함께 수행하도록 설계하여, 사례기반추론의 성능이 한층 더 제고될 수 있도록 하였다. 이 때, 최적의 절대 유사 임계값, 참조사례의 선택, 특징변수의 선택을 결정하기 위한 최적화 기법으로는 유전자 알고리즘을 적용하고자 하였다.

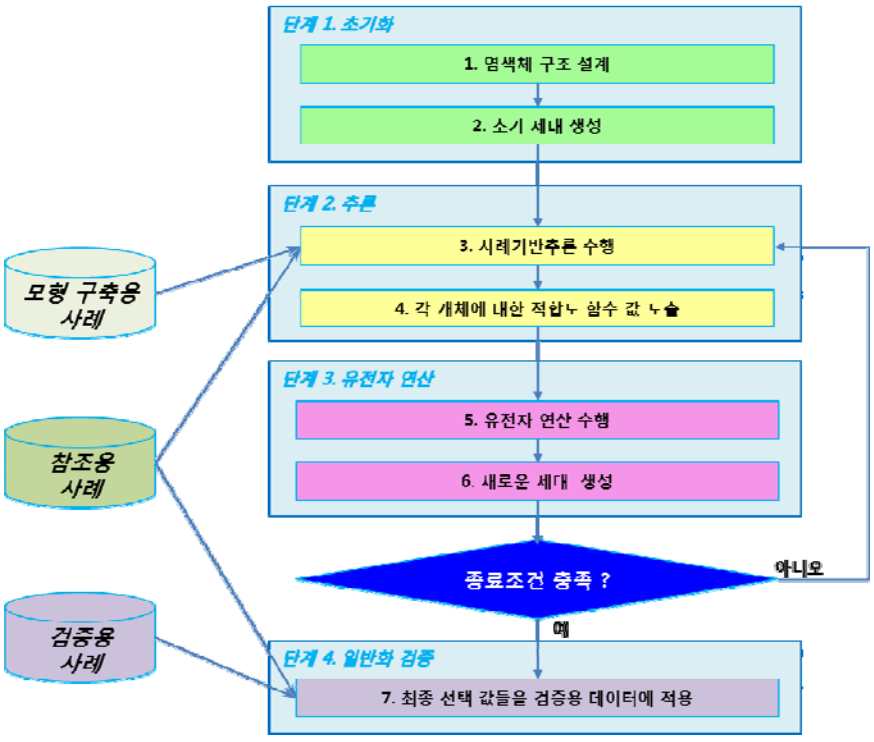
본 연구의 제안 기법은 다음의 <그림 3>에 제시된 것과 같이, 총 4단계의 절차에 따라 수행된다.

단계 1. 염색체 및 초기 개체군 생성

먼저 1 단계에서는 절대 유사 임계값, 특징변수 선택, 참조사례 선택을 최적화할 수 있도록 유전자 알고리즘의 염색체(chromosome)를 설계하고, 설계된 염색체 구조를 바탕으로 초기 개체군을 생성하는 작업이 이루어지게 된다. 이 때 염색체는 유전자 알고리즘의 유전 조작 활동이 쉽게 이루어질 수 있도록 이진코드(binary code)로 설계되며, 무작위 값으로 초기화된 염색체들을 100~300개 만들어 초기 세대 개체군(population)을 생성하게 된다.

단계 2. 현재 개체군에 대한 사례기반추론 실행

1 단계 과정을 통해 초기 세대 개체군이 생성되면, 2 단계에서는 개체군에 속한 염색체들을 대상으로 사례기반추론을 실행하게 된다. 즉,



<그림 3> 제안 기법의 구동 절차

각 염색체들이 나타내는 절대 유사 임계값, 특징변수 선택, 참조사례 선택 조건에 따라 사례기반추론을 실행한 뒤, 그 결과를 도출하게 되는 것이다.

전술했듯이, 절대 유사 임계값을 적용하기 위해서는 사례 간 유사도 값이 항상 0에서 1사이의 값을 갖도록 해야 한다. 이를 위해 2단계의 사례기반추론에서는 입력되는 모든 변수들(a_k)에 대한 최소-최대 정규화(Min-Max normalization) 변환 과정을 거친 뒤 아래 산식과 같이 유클리드 거리(Euclidean distance)를 기반으로 유사도를 산출한다. 이렇게 하면, 모든 유클리드 거리가 $0 \leq sim(U_0, U_i) \leq 1$을 만족하게 되어 어떠한 상황에서도 유사도의 값은 0부터 1사이의 값을 가지게 된다.

$$sim(U_0, U_i) = \frac{\sqrt{\sum_{k=1}^{m}(a_k^0 - a_k^i)^2}}{m} \quad [1]$$

where $U_i = (a_1^i, a_2^i \cdots, a_m^i), k = 1, 2, \cdots, m$

단계 3. 적합도 함수 값의 따른 유전자 알고리즘 수행

3 단계에서는 2 단계의 결과인 각 후보 유사 임계값, 특징변수 선택, 참조사례 선택 적용 시 도출된 적합도 함수(fitness function) 값에 의거하여, 유전자 조작을 수행하고 새로운 세대를 만드는 작업이 이루어지게 된다. 이 때, 제안 기법에 적용된 적합도 함수는 모형 구축용 데이터에 대한 평균 투자 수익률(rate of return)이다. 여러 유전인자 조작 기법 중에서 본 연구는 선택(selection)과 교배(crossover), 그리고 돌연변이(mutation)의 세 가지 기법을 적용하였으며, 사전에 설정한 중지 조건에 도달하기 전까지 2단계, 3단계 작업을 계속 반복하게 된다.

단계 4. 검증용 사례에 대한 예측 정확도 측정

유전자 알고리즘의 중지 조건이 도달하는 시점에서 앞의 단계들이 모두 끝나면, 최적에 근접하거나 최적 유사 임계값과 특징변수 및 참조사례의 선택이 도출되게 된다. 마지막 4단계에서는 이렇게 도출된 파라미터(parameter) 값들에 기반한 사례기반추론을 모형 구축에 활용하지 않았던 검증용 데이터에 적용하여, 그 성능을 최종 검증하게 된다. 본 단계를 통하여 연구의 제안모형이 실제 현실에 적용 가능성이 있는지를 확인해 볼 수 있다.

Ⅳ. 실증분석

4.1 실험 데이터

본 연구는 제안 기법의 유용성을 검증하기 위하여 국내 주식시장에 적용하여 보았다. 구체적으로 구글 파이낸스(Google Finance, https://www.google.com/finance)에서 제공하는 일별 KOSPI 200 주가 지수를 사용하여, 제안 기법의 성능을 확인해 보고자 하였다. 모형 구축에 사용된 표본 데이터는 2009년부터 2016년까지 총 8년간 1,986건 발생한 KOSPI 200 지수의 일별 종가 데이터이다. 지난 2007년 미국에서 시작된 서브프라임 모기지 사태(subprime mortgage crisis)는 전 세계로 파급되

어 국제 금융 시장의 신용 경색을 불러 일으켜 대규모의 금융위기 사태를 발생시켰다. 때문에 2007~2008년 사이의 주식시장 데이터는 그 이전이나 그 이후와 상당히 다른 패턴을 보이는 이상치에 가깝다. 때문에 본 연구에서는 글로벌 위기가 해소되기 시작한 2009년부터 시작되는 데이터를 구글 파이넌스를 이용하여 수집하고, 이를 기반으로 검증 작업을 수행하고자 하였다.

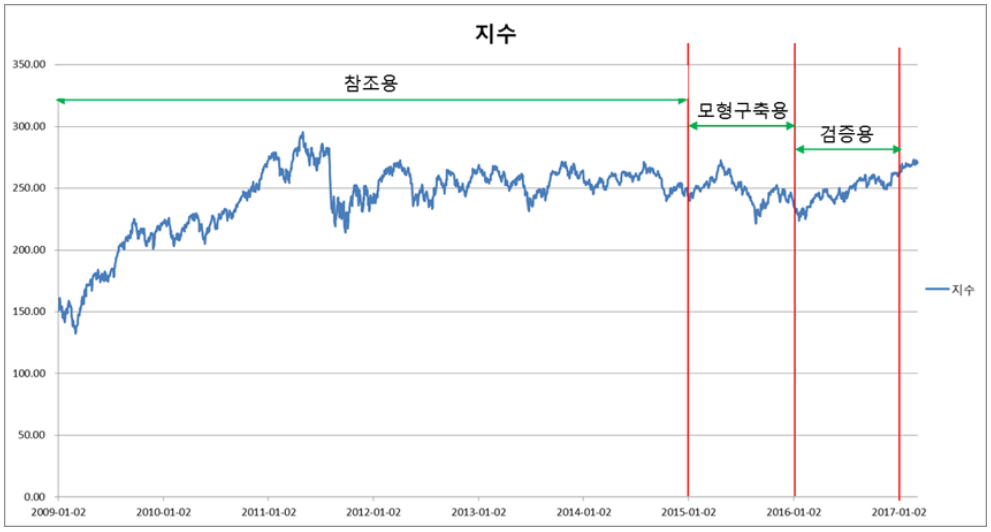
보통 사례기반추론 모형은 일반화 정도를 측정하기 위하여 데이터 셋을 학습용(참조용)과 검증용의 두 가지로 구성한다. 하지만, 본 연구에서 제안하는 모형은 구축 과정이 최적화 과정을 통하여 이루어 되므로 과도적합(overfitting) 문제를 최소화하기 위한 방편으로 데이터 셋을 참조용 데이터셋과 모형구축용 데이터셋, 그리고 검증용 데이터셋의 세 종류로 구분하였다. 모형에 사용된 데이터 셋의 선정 기준은 다음의 <그림 4>와 같다.

데이터의 구성은 <표 1>과 같이 약 75%를 차지하는 2009년부터 2014년까지의 1,492건의 데이터를 참조용으로 활용하였고 2015년의 248건(약 12.5%)의 데이터를 모형 구축용, 2016년의 246건(약 12.4%)를 검증용으로 활용하였다.

<표 1> 실험용 데이터셋의 구성

데이터 셋	데이터 수	구성 비율	기 간
참조용	1,492	75.12 %	2009년 ~ 2014년
모형 구축용	248	12.49 %	2015년
검증용	246	12.39 %	2016년
전 체	1,986	100.00 %	2009년 ~ 2016년

후보 특징변수로는 주가 지수 예측에 주로 사용되는 기술적 지표를 중심으로 <표 2>와 같은 변수들을 사용하였다. 이 기술적 지표들을 간략히 살펴보면 다음과 같은 특성이 있다 (Achelis, 1995; Chang et al., 1996; Wilder, 1985; 1986).



<그림 4> 데이터셋의 범위

<표 2> 후보 특징변수들

기술지표	산 식	관련 연구
Stochastic %D	$\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$	Achelis(1995) Kim & Ahn(2008; 2012) 김선웅과 안현철(2010) Dao & Ahn(2014)
Stochastic Slow %D	$\frac{\sum_{i=0}^{n-1} \%D_{t-i}}{n}$	Achelis(1995) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
Momentum	$C_t - C_{t-4}$	Chang et al.(1996) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
ROC	$\frac{C_t}{C_{t-n}} \times 100$	Achelis(1995) Kim & Ahn(2008; 2012) 김선웅과 안현철(2010) Dao & Ahn(2014)
Williams' %R	$\frac{H_n - C_t}{H_n - L_n} \times 100$	Achelis(1995) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$	Chang et al.(1996) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
Disparity 5	$\frac{C_t}{MA_5} \times 100$	Choi(1995) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
Disparity 10	$\frac{C_t}{MA_{10}} \times 100$	Choi(1995) Kim & Ahn(2008; 2012) 김선웅과 안현철(2010) Dao & Ahn(2014)
OSCP	$\frac{MA_5 - MA_{10}}{MA_5}$	Achelis(1995) Kim & Ahn(2008; 2012) 김선웅과 안현철(2010) Dao & Ahn(2014)
CCI	$\frac{M_t - SM_t}{0.05 \times D_t}$	Achelis(1995) Chang et al.(1996) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
RSI	$100 - \frac{100}{\left(1 + \frac{\frac{\sum_{i=0}^{n-1} Up_{t-i}}{n}}{\frac{\sum_{i=0}^{n-1} Dw_{t-i}}{n}}\right)}$	Wilder(1986) Achelis(1995) Kim & Ahn(2008; 2012) Dao & Ahn(2014)
Ultimate Oscillator	$\frac{(4 \times avg_7) + (2 \times avg_{14}) + (avg_{28})}{4+2+1} \times 100$	Williams(1985) Achelis(1995) 안현철과 이형용(2009)
SONAR	$\frac{MA_n - (MA_{n-1})}{MA_{n-1}} \times 100$	Achelis(1995) 안현철과 이형용(2009)
VHF	$\frac{HH_n - LL_n}{\sum_{t=1}^{n} \lvert (C_t - C_{t-1}) \rvert} \times 100$	Achelis(1995) 안현철과 이형용(2009)

주) C_t:t 거래일의 종가, L_t:t 거래일의 저가, H_t:t 거래일의 고가, MA_t:직전 t 거래일의 이동평균, LL_t:직전t 거래일의 최저가, HH_t:직전t 거래일의 최고가, UP_t:t 거래일에서의 상향가격변동치, DW_t:t 거래일에서의 하향가격변동치, $M_t: \frac{(H_t + L_t + C_t)}{3}$, $SM_t: \frac{(\sum_{i=1}^{n} m_{t-i+1})}{n}$, $D_t: \frac{\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert}{n}$

Stochastic(Fast or Slow)은 현재의 주가 수준이 일정 기간의 변동범위 안에서 상대적으로 어느 수준의 위치에 있는지를 가지고 판단하는 지표이다. 즉, %K의 값이 100이면 5일 동안에 형성된 시장가격 중 당일 종가가 최고 수준임을 나타내고, %K값이 0일 때 5일 동안 형성된 시장가격 중 당일 종가가 최저 가격을 의미한다. Stochastic의 활용 방법은 지표 %K가 80% 이상에서 %D를 뚫고 하락 할 경우 매도 시점을 파악하고, 지표 %K가 20% 이하에서 %D를 뚫고 상승 할 경우 매수 시점을 파악하면 된다. 이 지표는 단기 매매 지표로 이용된다.

Momentum은 단기간 주가가 변해온 양을 측정하는 지표이며, 가격을 이용한 추세판단 지표이다. 현 시점의 가격에서 일정 기간 전의 가격을 차감해서 산출해 내기 때문에 수식이 단순하고 가격의 방향에 따르기 않고 가격의 방향이 전환하기 전에 먼저 지표의 방향이 변함으로서 선도지표의 역할을 하는 특징이 있다.

ROC(Rate of Change)는 가격을 활용한 추세판단 지표이다. 과거의 일정시점 가격과 현재의 가격을 비교하여 현재 가격이 상승이나 하락 추세에 있는지를 판단하는 지표이다. ROC와 유사한 분석기법으로 Momentum이 있는데, Momentum은 가격 변동 비율을 나타내고 ROC는 이를 백분율로 표시한다는 점이 다소 다르다.

Williams' %R은 일정 기간 내에서 시장 가격의 고가와 저가 범위에서 당일의 종가가 어디 위치에 있는지를 나타내는 지표이며 Stochastics과 비슷하다. Stochastics보다는 강세장에서 높은 예측력을 보이고, 약세장에서는 예측력이 다소 떨어지는 특징이 있다.

A/D Oscillator(Accumulation/Distribution Oscillator)는 A/D Line을 변형한 지표이며, 산출식은 매우 단순하다. A/D Oscillator는 A/D Line값이 변동 폭이 지나치게 크게 움직이므로 이것을 Oscillator化하여 도식화하였다. 즉, A/D Line의 단기와 장기의 지수이동평균의 차이를 이용하여 필요 없는 움직임을 없애고 지표의 변동 폭을 줄인 것이다. 주로 주가나 지수의 과열과 침체를 판단하는데 사용된다.

이격도(Disparity)는 주가와 이동평균선 사이가 얼마나 벌어져 있는가(괴리율)를 보여준다. 당일 주가를 이동평균치로 나눈 값으로 단기 투자 시점을 포착하는 기술적 지표다.

OSCP(Price Oscillator)는 MACD와 유사한 지표이다. 즉, 두 종류의 이동평균선의 차이를 확인하여 주가의 변화 시점을 포착하려는 목적을 가지고 있다. 다만 MACD와 OSCP의 차이는 MACD가 12일 지수이동평균선과 26일 지수 이동평균선의 차이를 이용하여 수치로 표시되지만, OSCP는 수치와 비율로 모두 표시한다는 것이다. 또한, OSCP의 경우 이동평균선을 지수이동평균선과 단순이동평균선 모두 사용가능하다는 것이다. 추가로 MACD는 Signal선을 통해 분석하지만 OSCP의 경우는 방향성 및 Zero line의 교차점을 주요 분석 기법으로 사용하는 정도이다. MACD(Moving Average Convergence Divergence) 지표는 장기 이동평균선과 단기 이동평균선이 서로 벌어지면 결국 다시 좁아진다는 성질을 이용하여 두 이동평균선이 가장 멀리 떨어지는 지점을 찾는 것이다.

CCI(Commodity Channel Index)는 최근 가격이 이동평균가격이 얼마나 벌어져 있는가를 표시하여 추세의 방향과 강도를 표시하는 지표

이다. 따라서 CCI의 절대 값이 크면 추세는 강하다는 것을 나타나고 CCI 값의 부호가 (+)일 경우는 상승, (-)일 경우는 하락 추세로 이해하면 된다. CCI는 추세의 방향과 강도를 모두 보여줄 수 있어 추세추종 거래에 유용한 지표로 인식되고 있다.

RSI(Relative Strength Index)는 반추세지표(Countertrend Oscillator)의 대표 지표로서 가격의 상승압력과 하락압력의 상대적인 강도를 보여준다. ROC와 Stochastics 등이 과거 자료에 의해 지표가 왜곡될 수 있는 단점을 가지고 있지만, RSI는 이 단점을 개선한 지표이다. RSI는 일정 기간을 기준으로 그 기간의 가격 변동 중 상승분이 얼마나 비중을 차지하고 있는지를 나타내고, RSI가 0에 근접하다는 것은 그 기간 중 하락 강도가 강하다는 뜻이고 반대로 RSI가 100에 근접하다는 것은 그 기간 중 상승 강도가 강하다는 것을 의미한다.

Ultimate Oscillator는 다른 세주기의 평균을 가중하여 합산한 값을 지표로 사용한다. 주기는 보통 7일 평균, 14일 평균, 28일 평균을 적용한다.

SONAR는 핵심 기법을 Momentum을 이용하기 때문에 Sonar Momentum으로도 칭해진다. Momentum은 기하학적으로 곡선 상에 있는 점의 기울기를 계산하는 것으로 경제학에서 언급되는 한계변화율을 연상하면 될 것이다. 예로 주가에 한계변화율 개념을 적용하면 주가가 바닥에서 상승 시 그 상승폭(한계변화율)은 점차 커지면서 어느 수준에 다다르면 주가는 계속 상승하지만 그 상승폭이 점점 둔화되어 한계변화율이 감소하며, 결국에는 정점을 이루어 한계변화율은 0이 되서 주가는 하락 전환한다는 것이다. 즉, 주가의 상승기 및 하락기에 상승과 하락의 기울기를 계산하여 상승 및 하락의 강도를 계산해서, 그것에 따라 주가의 강도가 강할수록 이전의 흐름을 이어나갈 확률이 높아지지만 반대로 주가의 강도가 약화되면 이것은 주가의 변동가능성이 많아짐을 의미한다는 것에서 착안한 것이다.

VHF(Vertical Horizontal Filter)는 장이 추세적 장인지 비추세적 장인지 결정한다. MACD나 이동평균 같은 추세 추종형 지표들은 추세적 장에서는 정확히 표현하는 반면, 비추세적 시장에서는 상반된 결과를 내놓기도 한다. 또한 RSI나 Stochastic 같은 Oscillator들은 비추세적 장에서는 매매시점을 잘 포착하지만 추세적 장에서는 너무 성급한 신호를 내놓기도 한다. VHF는 시장의 추세 정도를 나타내어 사용 할 지표를 선택하는데 도움을 주는 지표이다.

한편 유전자 알고리즘 탐색을 위한 제어 파라미터들과 관련해서, 개체군 규모를 300개체로 설정하였으며, 교배 비율을 0.7로 설정하고 돌연변이 비율은 0.1로 설정하였다. 아울러 중지 조건으로는 6000회 반복, 즉 20세대만큼 탐색을 반복하도록 설정하여 실험하였다. 모형실험에 사용된 프로그램은 Microsoft Excel VBA(Visual Basic for Applications)를 활용하여 절대 유사 임계값 기반 사례기반추론을 구현하였으며, 유전자 알고리즘 최적화는 Palisade Software사의 Evolver Version 5.5를 이용해 구현하였다.

아울러, 제안 기법의 성과를 좀 심도 있게 분석하기 위하여 몇 가지 비교모형을 함께 실험해 보았다. 구체적으로 로지스틱 회귀모형(Logistic Regression; LOGIT)과 다중판별분석

(Multi Discriminant Analysis; MDA), 인공신경망(Artificial Neural Network; ANN), Support Vector Machines(SVM) 등의 기법을 추가로 실험해 보고, 해당 모형이 산출하는 신호에 따라 매매한 결과와 제안 기법의 수익성을 서로 비교해 보고자 하였다.

4.2 실험 결과

먼저 다음의 <표 3>에 비교모형들의 성능이 제시되어 있다. <표 3>에서 각 모형의 성능은 크게 3가지 지표로 측정되었다. 첫 번째는 '예측 정확도'이다. 예측 정확도는 금일의 주식시장 상태 데이터를 이용해, 익일의 주식시장 등락을 예측했을 때, 얼마나 정확하게 등락을 맞추었는지에 대한 비율값을 나타낸다. <표 3>에서 CBR의 경우, 학습용 표본에 대한 예측정확도가 제시되어 있지 않은데, 이는 CBR에서 학습용 표본은 모두 예측결과를 생성하기 위한 '사례기반'으로 사용되었기 때문이다.

두 번째는 수익률이다. 수익률은 익일 주식시장의 등락이 예측되었을 때, 그 결과에 따라 매매의사결정을 내릴 경우 거래 종료시점에 얻게 될 투자수익률을 의미한다. 예측 정확도가 높을 경우, 대체로 수익률이 높아지지만 항상 비례 관계에 있는 것은 아니다.

세 번째 지표인 거래건수는 앞서 소개한 수익률을 산출하는데 이루어진 거래 횟수이다. 만약 어제 시점에 오늘 주식시장이 오를 거라고 예측하여 시스템이 '매수'를 단행했는데, 오늘 시점에 내일 주식시장이 내릴 거라고 예측한다면 해당 시스템은 어제 매수한 주식을 오늘 '매도'하게 될 것이다. 반면 해당 시스템이 오늘 시점에 내일 주식시장이 오를 것으로 예측한다면, 이미 어제 '매수'를 단행한 상태이기 때문에 오늘은 그냥 가만히 있으면 된다. 즉, 이 예에서 전자의 경우에는 2번의 거래를 하게 되지만, 후자의 경우에는 1건의 거래만 발생하게 된다. 모

<표 3> 비교모형 실험결과

모형		LOGIT	MDA	CBR[2]	ANN	SVM
최적 설정값		*Backward*	*Enter*	$k=2$	$h=23$	*RBF* $C=1$ $\sigma^2=1$
예측 정확도	학습용	52.13%	52.87%		51.88%	58.74%
	모형구축용				51.61%	
	검증용	54.88%	53.66%	52.85%	57.21%	55.28%
수익률	학습용	196.28%	206.40%		179.17%	563.36%
	모형구축용				99.73%	
	검증용	117.06%	115.72%	110.16%	123.03%	115.97%
거래 건수[1]	학습용	207	352		96	241
	모형구축용				11	
	검증용	28	43	54	11	32

형에서 특정 방향이 연속적으로 예측될 경우, 거래건수는 감소하게 되며, 이 경우 거래비용을 절감할 수 있다. 즉, <표 3>에서 거래건수가 적게 나타난 모형(예. ANN)은 거래 기간 동안 특정 방향이 연속적으로 예측된 경우가 많은 모형이라 할 수 있다.

<표 3>에서 볼 수 있듯이 비교모형 중 검증용 데이터 셋에서의 수익률은 ANN > LOGIT > SVM > MDA > CBR 순으로 나타났다. 구체적으로 인공신경망(ANN)이 123.03%로 가장 높게 나왔고 전통적인 사례기반추론에서 가장 낮은 110.16%가 도출되었다. 벤치마크(benchmark) 수익률, 즉 검증용 기간으로 설정된 2016년 초에 매수하여, 2016년 말에 매도했을 때의 수익률이 110.82%로 도출되었는데, 전통적인 사례기반추론은 이 보다도 낮은 결과값을 보여 주식시장 예측에 있어 대단히 낮은 정확도를 보임을 알 수 있다.

한편 다음의 <표 4>는 제안 기법의 실험결과를 제시하고 있다. 본 연구에서는 절대 유사 임계값(Th)과 최적 특징변수 선택(FS) 그리고 최적 참조사례 선택(IS)을 고려하고자 하였는데, 절대 유사 임계값의 최적화는 기본으로 하되 여기에 FS나 IS를 추가로 수행했을 때 어떤 성능의 개선이 이루어지는지 확인해 보고자 하였다. 그 실험결과가 <표 4>에 모두 제시되어 있다.

우선 예측정확도의 경우, 제안모형은 55.69%~61.38% 사이의 성과를 나타내고 있는데, 이는 비교모형들과 비교해 상당히 높은 수치임을 알 수 있다. 특히 가장 우수한 성능을 보인 제안모형인 모형3(61.38%)은 비교모형 중 가장 높은 정확도를 보인 인공신경망(57.21%)에 비해 약 4% 이상 더 높은 정확도를 나타냄을 알 수 있다. 하지만, 이 둘을 직접적으로 비교하는 것은 학술적으로 큰 의미를 갖고 있지 못하다. 그 이유는 제안모형의 경우, 참조할 사례가 하나도 없었던 경우에 대해서는 예측을 수행하지 않았고, 따라서 모든 경우에 대해 예측을 수행한 것이 아니라, 일부 표본들이 예측정확도 산출 시 배제된 결과값이기 때문이다. 따라서, 제안모형의 성능은 비교모형과의 수익률 관점에서 비교해야 보다 정확하게 판단할 수 있다.

<표 4> 제안 기법 실험결과

유형		모형1: Th	모형2: Th + FS	모형3: Th + IS
최적 임계값		0.10806	0.12825	0.22354
최적 특징변수의 수		14	6	14
최적 참조사례의 수		1,492	1,492	770
예측 정확도	모형구축용	51.21%	55.65%	55.65%
	검증용	56.10%	55.69%	61.38%
수익률	모형구축용	110.08%	111.29%	115.87%
	검증용	120.73%	123.91%	131.14%
거래 건수	모형구축용	31	24	37
	검증용	30	24	47

수익률 관점에서 보면, 절대 유사 임계값만을 최적화한 모형(모형1, 120.73%)을 제외한 모든 모형(모형2, 모형3)에서 비교모형 중 가장 우수한 성능을 보였던 인공신경망(123.03%)보다 더 우수한 투자수익률을 나타냈음을 확인할 수 있다. 또한 특징변수 선택을 최적화 하려고 한 모형2(123.91%)에 비해, 참조사례 선택을 최적화 하고자 한 모형3(131.14%)에서 더 우수한 수익률을 나타냈는데, 이를 통해 최적의 시스템 트레이딩 기법을 도출하는데 있어 적절한 특징변수의 선택보다 유의미한 참조사례를 선택하는 것이 더 큰 영향을 미친다는 점을 알 수 있다.

이상의 실험결과들을 모두 종합해 보면, 기존의 사례기반추론 모형에서는 벤치마크 수익률 보다도 낮은 수익률을 보였는데, 제안된 사례기반추론 모형에서는 벤치마크 수익률은 물론 모든 비교모형보다도 더 높은 수익률을 나타냈다. 즉, 가장 낮은 성능을 보인 사례기반추론이 절대 유사 임계값 기반 사례기반추론으로 변형되자 성능이 향상된다는 사실을 확인할 수 있는데, 이를 통해 본 연구에서 제안한 절대 유사 임계값 기반 사례기반추론이 그만큼 시스템 트레이딩에 더 적합한 알고리즘임을 알 수 있다.

Ⅴ. 결 론

본 연구에서는 최근 금융 분야에서 주목 받고 있는 시스템 트레이딩과 관련하여, 사례기반추론과 유전자 알고리즘을 결합시킨 새로운 형태의 지능형 시스템 트레이딩을 제안하였다. 구체적으로 본 논문에서는 이른바 절대 유사 임계값 개념의 적용과 함께 특징변수 선택, 그리고 참조사례 선택 등을 반영하여 수익률을 높일 수 있는 새로운 개념의 사례기반추론 기법을 개발하고 제안하였다. 제안 기법의 유용성을 확인하기 위하여 2009년부터 2016년까지 총 8년간의 축적된 일별 KOSPI 200 지수에 적용하여 실험한 결과, 제안 기법이 기존의 연구에서 다뤄진 다른 모형들과 비교해 볼 때 더 향상된 수익률이 나오는 것을 확인 할 수 있었다.

본 연구의 시사점은 다음과 같이 요약 할 수 있다. 먼저, 본 연구는 변화가 많은 주식 시작에서 등락을 예측하기 위한 기법으로 선행 되었던 다른 예측 모형에 비해 절대 유사 임계값 기반 사례기반추론이 수익률 향상에 상당히 도움되는 것을 실증적으로 증명하고 있다. 또한 전통적인 사례기반추론과는 월등히 차이가 있는 것을 확인하여 유전자 알고리즘과 결합이 개선된 사례기반추론 모형 구축의 효용성을 증명하였다.

다음으로 본 연구의 제안 기법이 포함하고 있는 특징변수 및 참조사례 선택이 매매 수익률 증대 목표의 시스템 트레이딩에 유용하게 활용될 수 있다는 것을 시사하고 있다. 본 연구를 통해 수행된 실험에서 특징변수 및 참조사례 선택을 도입했을 때, 절대 유사 임계값만을 최적화한 모형보다 수익률이 더 높다는 것을 확인 할 수 있었다. 이런 것들로 미루어 보면 향후 시스템 트레이딩에서 참조사례와 특징변수의 선택은 다른 모형의 최적화 관련 연구에서도 적용되어 성능 개선에 도움이 될 것으로 기대된다.

또한 특징변수 선택보다는 참조사례의 선택

이 사례기반추론의 성능 개선에 더 지대한 영향을 미친다는 것을 확인한 점 역시 주목할 만한 결과이다. 이는 부도예측이나 고객관계관리 등 다른 경영분야의 사례기반추론 연구들(예. Ahn et al., 2007; Ahn & Kim, 2009 등)에서 나타난 결과와 동일한 양상이라는 점이라서, 더 의미 있는 발견이라고 할 수 있다.

실무적 관점에서 본 연구는 기존 기법 대비 향상된 수익률을 가져다 줄 수 있는 새로운 알고리즘을 제안하였다는 점에서, 수익성 개선에 지대한 관심을 갖고 있는 투자금융 관련 기업들에게 유용한 지침이 될 수 있을 것으로 기대된다. 특히 본 연구의 제안 기법은 기본적으로 사례기반추론에 기반하고 있기 때문에, (1) 과거 어떤 사례들을 참조하여 특정 투자의사결정을 내리게 되었는지에 대한 이유(Why)를 설명할 수 있으며, (2) 사례기반을 실시간으로 갱신하는 것이 용이하여, 주기적인 재학습 과정을 필요로 하지 않는 등의 장점을 추가로 갖고 있다. 때문에 실무적 활용가치가 대단히 높은 기법이라고 할 수 있다.

학술적 관점에서 본 연구는 기존 연구에서 제시된 '절대 유사 임계치 기반 사례기반추론'을 시스템 트레이딩이라고 하는 새로운 분야에 적용함으로써, 그 우수성을 다시금 확인했다는 의의가 있다. 해당 기법을 처음 제안한 안현철(2013)의 연구에서는 제안모형을 국내 한 온라인 쇼핑몰의 표적 마케팅 대상 선정에 적용하여 그 성능을 검증하였다. 하지만, 이 때는 커버리지(coverage)라는 새로운 조절변수를 사용해야 하는 등 적용에 다소 불편함이 있었다. 하지만 본 연구에서 적용한 시스템 트레이딩 분야의 경우, 수익률을 극대화하도록 모형을 설계하는 것이 가능하여 그 어떤 조절변수의 도입도 필요치 않은 장점이 있었다. 이처럼 절대 유사 임계치 기반 사례기반추론이 가장 효과적으로 활용될 수 있는 새로운 응용 분야를 개척했다는 점이 본 연구의 또 다른 학술적 의의라 하겠다.

본 연구의 한계점은 다음과 같다. 먼저 실제 매매에 있어서 거래에 대한 비용 지불이 발생함에도 본 연구에서는 거래 비용에 대한 감안이 없이 실험을 수행을 했다는 점이 있다. 거래비용에 대한 추정이 어려울 수 있지만, 향후 연구에서는 거래비용까지 반영하여 제안 시스템의 성능을 더욱 정밀하게 측정하는 노력이 필요할 것이다.

두 번째로 실험 데이터에 있어서 다양한 사례의 데이터를 좀 더 확보해야 할 필요성이 있다는 것을 지적 할 수 있다. 현재 연구에서는 2009년부터 2016년까지의 일별 KOSPI 200 지수를 검증용 데이터로 활용하였는데, 검증이 단 1년간의 수익률로 판단되었다는 점은 본 연구의 한계라 할 수 있다. 향후 데이터가 더 누적되었을 때, 보다 방대한 데이터를 바탕으로 제안기법의 성능을 재검증할 필요가 있을 것으로 사료된다.

실증분석의 대상이 KOSPI 200 지수에 한정되었다는 점 역시 본 연구의 주요한 한계점이 될 것이다. 개별 종목의 주가나 종합주가지수 외의 다양한 펀드 등으로 분석대상이 확대된다면, 제안모형의 실용성을 한층 더 정밀하게 확인하는데 도움이 될 수 있을 것으로 예상된다.

또한 본 논문에서 제안된 새로운 사례기반추론 모형은 주가 예측뿐만 아니라 정밀한 예측을 요구하는 다른 분야에도 적용이 가능할 것

이다. 그렇기 때문에 제안 기법을 다른 경영분야의 예측 문제에도 적용해 보고, 과연 그 분야에서도 동일한 성과 개선이 이루어지는지 향후 연구에서 확인해 볼 필요가 있다.

끝으로 본 연구는 사례기반추론의 최적화 요소로서 특징변수 선택과 참조사례 선택의 2가지만을 고려하고 있다. 하지만, 기존 연구에 따르면 특징변수에 대한 가중치 부여(feature weighting)나 여러 요소들을 동시에 최적화(simultaneous optimization)이 더 나은 사례기반추론의 성과개선을 도모할 수도 있다. 때문에 최적화 대상을 한층 더 확장하는 후속연구가 추후 이루어져야 할 것으로 판단된다.

참고문헌

김선웅, 안현철. “Support Vector Machines 와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발.” 지능정보연구 16권 1호, 2010, pp.71-92.

박윤주, “통계적 분석 기법을 기반으로 한 사례기반추론에 대한 연구,” 박사학위논문, 경영공학전공, 한국과학기술원, 2006.

안현철, 김경재, 한인구, “효과적인 고객관계관리를 위한 사례기반추론 동시 최적화 모형.” 지능정보연구 11권 2호, 2005a, pp.175-195.

안현철, 이형용. “투자 의사결정 지원을 위한 유전자 알고리즘 기반의 다중 인공지능 기법 결합 모형: KOSPI 에의 응용.” e-비즈니스연구 10권 1호, 2009, pp.215-236.

안현철. “사례기반추론의 유사 임계치 및 커버리지 최적화.” 정보처리학회논문지. 소프트웨어 및 데이터 공학 2권 8호, 2013, pp.535-542.

옥중경, 김경재. “유전자 알고리즘 기반의 기업부실예측 통합모형.” 지능정보연구 15권 4호, 2009, pp.99-120.

이우식. “딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측.” 한국데이터정보과학회지 28권 2호, 2017, pp.287-295.

이재식, 송영균, 허성회. “인공신경망 앙상블을 이용한 옵션 투자예측 시스템,” 한국지능정보시스템학회 학술대회논문집, 2000, pp.489-497.

이형용. “한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형.” Entrue Journal of Information Technology Vol.7, No.2, 2008, pp.33-43.

이훈영, 박기남. “사례기반예측시스템의 정확한 예측을 위한 최적 결합 사례개수결정방법에 관한 연구.” 경영학연구 27권 5호, 1999, pp.1239-1252.

송성환, 황선호, 이용희, 이현경, 한경석, 김종배, “트레이딩을 위한 소셜 빅데이터 분석 모델”, *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.6 No.3, 2016, pp.91-100.

송유정, 이종우, “텐서플로우를 이용한 주가 변동 예측 딥러닝 모델 설계 및 개발.” 한국정보과학회 학술발표논문집, 2017,

pp.799-801.

홍승현, 신경식, “유전자 알고리즘을 활용한 인공신경망 모형 최적입력변수의 선정: 부도예측 모형을 중심으로.” 한국지능정보시스템학회 9권 1호, 2003, pp.227-249.

Aamodt, A., and Plaza, E.. “Case-based reasoning: Foundational issues, methodological variations, system approaches.” AI communications, Vol.7, No.1, 1994, pp.39-59.

Achelis, S. B., Technical Analysis from A to Z. New York: McGraw Hill, 2001.

Ahn, H., and Kim, K.-j., “Using genetic algorithms to optimize nearest neighbors for data mining,” Annals of Operations Research, Vol. 163, No.1, 2008, pp. 5-18.

Ahn, H., and Kim, K. J., “Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach.” Applied Soft Computing 9.2 (2009): 599-607.

Ahn, H., Kim, K J.., and Han, I., “Global optimization of feature weights and the number of neighbors that combine in a case based reasoning system.” Expert Systems, Vol.23, No.5, 2006a, pp.290-301.

Ahn, H., Kim, K. J., and Han, I., “A case-based reasoning system with the two-dimensional reduction technique for customer classification.” Expert Systems with Applications, Vol.32 No.4, 2007, pp.1011-1019.

Ahn, H., Kim, K. J., and Han, I., “Hybrid genetic algorithms and case based reasoning systems for customer classification.” Expert Systems, Vol.23, No.3, 2006b, pp.127-144.

Alexander, S. S. “Price Movements in Speculative Markets: Trends or Random Walks, Number 2.” IMR; Industrial Management Review (pre-1986), Vol.5, No.2, 1964, 25.

Atsalakis, G. S., and Valavanis, K. P., “Forecasting stock market short-term trends using a neuro-fuzzy based methodology.” Expert Systems with Applications, Vol.36, No.7, 2009a, pp.10696-10707.

Atsalakis, G. S., and Valavanis, K. P., “Surveying stock market forecasting techniques – Part II: Soft computing methods.” Expert Systems with Applications, Vol.36, No.3, 2009b, pp.5932-5941.

Babu, T. R. and M. N. Murty, “Comparison of genetic algorithm based prototype selection schemes”, Pattern Recognition, Vol.34, No.2, 2001, pp.523-525.

Bao, D., and Yang. Z., “Intelligent stock trading system by turning point confirming and probabilistic reasoning.” Expert Systems with Applications, Vol.34, No.1 (2008, pp.620-627.

Brock, W., Lakonishok, J., and LeBaron, B, "Simple technical trading rules and the stochastic properties of stock returns." The Journal of Finance, Vol.47, No.5, 1992, pp.1731-1764.

Caporale, G. M., and Pittis, N., "Cointegration and predictability of asset prices." Journal of International Money and Finance, Vol.17, No.3, 1998, pp.441-453.

Cardie, C., "Using decision trees to improve case-based learning", Proceedings of the Tenth International Conference on Machine Learning, San Francisco, CA, 1993, pp.25-32.

Chang, C. C., and Lin, C. J.,. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST), Vol.2, No.3, 2011, 27.

Chavarnakul, T., and Enke, D., "A hybrid stock trading system for intelligent technical analysis-based equivolume charting." Neurocomputing, Vol.72, No.16, 2009, pp.3517-3528.

Chiu, C., "A case-based customer classification approach for direct marketing." Expert Systems with Applications, Vol.22, No.2, 2002, pp.163-168.

Chiu, C., Chang, P. C., and Chiu, N. H., "A case-based expert support system for due-date assignment in a wafer fabrication factory." Journal of Intelligent Manufacturing, Vol.14, No.3, 2003, pp.287-296.

Choi, J. "Technical indicators." Seoul: Jinritamgu Publishing, 1995.

Chun, S. H., and Park, Y. J., "Dynamic adaptive ensemble case-based reasoning: application to stock market prediction." Expert Systems with Applications, Vol.28, No.3, 2005, pp.435-443.

Dao, T., and Ahn, H., "An Optimized Combination of π-fuzzy Logic and Support Vector Machine for Stock Market Prediction," Journal of Intelligence and Information Systems, Vol.20, No.4, 2014, pp.43-58.

Donaldson, R. G., and Kamstra, M., "An artificial neural network-GARCH model for international stock return volatility." Journal of Empirical Finance, Vol.4, No.1, 1997, pp.17-46.

Elton, E. J. and M. J. Gruber, Modern Portfolio Theory and Investment Analysis, Wiley, 1984.

Fama, E. F. "The behavior of stock-market prices." The journal of Business, Vol.38, No.1, 1965, pp.34-105.

Fama, E. F., "Efficient capital markets: A review of theory and empirical work." The journal of Finance, Vol.25, No.2, 1970, pp.383-417.

Fuller, R. J., & Kling, J. L., "Is the stock market predictable?." The Journal of Portfolio Management, Vol.16, No.4, 1990,

pp.28-36.

Granger, C. W, “Some properties of time series data and their use in econometric model specification.” Journal of econometrics, Vol.16, No.1, 1981, pp.121-130.

Hart, P.E., “The condensed nearest neighbor rule”, IEEE Transactions on Information Theory, Vol.14, No.3, 1968, pp.515-516.

Huang, Y. S., C.C. Chiang, J. W. Shieh and E. Grimson, “Prototype optimization for nearest-neighbor classification”, Pattern Recognition, Vol. 35, No.6, 2002, pp.1237-1245.

i Guiu, J. G., i Ribé, E. G., i Mansilla, E. B., and i Fàbrega, X. L., “Automatic diagnosis with genetic algorithms and case-based reasoning.” Artificial Intelligence in Engineering, Vol.13, No.4, 1999, pp.367-372.

Jarmulak, J., Craw, S., and Rowe, R., “Self-optimising CBR retrieval.” Tools with Artificial Intelligence, 2000. ICTAI 2000. Proceedings. 12th IEEE International Conference on. IEEE, 2000.

Kim, K. J., “Toward global optimization of case-based reasoning systems for financial forecasting.” Applied intelligence, Vol.21, No.3, 2004, pp.239-249.

Kim, K.-j., and Ahn, H., “Simultaneous optimization of artificial neural networks for financial forecasting,” Applied Intelligence, Vol.36, No.4, 2012, pp.887-898.

Kim, K. J., and Han, I., “Application of a hybrid genetic algorithm and neural network approach in activity-based costing.” Expert Systems with Applications, Vol.24, No.1, 2003, pp.73-77.

Kim, K. J., and Han, I., “Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index.” Expert systems with Applications, Vol.19, No.2, 2000, pp.125-132.

Kim, K. J., and Han, I., “Maintaining case-based reasoning systems using a genetic algorithms approach.” Expert Systems with Applications, Vol.21, No.3, 2001, pp.139-145.

Kim, K. J., and Lee, W. B., “Stock market prediction using artificial neural networks with optimal feature transformation.” Neural computing & applications, Vol.13, No.3, 2004, pp.255-260.

Kuncheva, L. I., and Jain, L. C., “Nearest neighbor classifier: Simultaneous editing and feature selection.” Pattern recognition letters, Vol.20, No.11, 1999, pp.1149-1156.

Lipowezky, U., “Selection of the optimal prototype subset for 1-NN classification”, Pattern Recognition Letters, Vol.19,

No.10, 1998, pp.907-918.

Lo, A. W., and MacKinlay, A. C., "Stock market prices do not follow random walks: Evidence from a simple specification test." Review of financial studies, Vol.1, No.1, 1988, pp.41-66.

Malkiel, B. G., A random walk down Wall Street: including a life-cycle guide to personal investing. WW Norton & Company, 1999.

McMillan, D. G., :"Non-linear forecasting of stock returns: Does volume help?." International Journal of forecasting, Vol.23, No.1, 2007, pp.115-126.

Nunez-Letamendia, L., "Fitting the control parameters of a genetic algorithm: An application to technical trading systems design." European journal of operational research, Vol.179, No.3, 2007, pp.847-868.

Poon, S. H., and Taylor, S. J., "Stock returns and volatility: an empirical study of the UK stock market." Journal of banking & finance, Vol.16, No.1, 1992, pp.37-59.

Sanchez, J. S., F. Pla and F. J. Ferri, "Prototype selection for the nearest neighbour rule through proximity graphs", Pattern Recognition Letters, Vol.18, No.6, 1997, pp.507-513.

Schulmeister, S., "Profitability of technical stock trading: Has it moved from daily to intraday data?." Review of Financial Economics, Vol.18, No.4, 2009, pp.190-201.

Shin, K. S., and Han, I., "Case-based reasoning supported by genetic algorithms for corporate bond rating." Expert Systems with Applications, Vol.16, No.2, 1999, pp.85-95.

Siedlecki, W., and Sklansky, J., "A note on genetic algorithms for large-scale feature selection." Pattern recognition letters, Vol.10, No.5, 1989, pp.335-347.

Silvapulle, P., and Choi, J. S., "Testing for linear and nonlinear Granger causality in the stock price-volume relation: Korean evidence." The Quarterly Review of Economics and Finance, Vol.39, No.1, 1999, pp.59-76.

Skalak, D.B., "Prototype and feature selection by sampling and random mutation hill climbing algorithms", Proceedings of the Eleventh International Conference on Machine Learning, New Jersey, NJ, 293-301, 1994.

Sun, J., and Hui, X. F., "Financial distress prediction based on similarity weighted voting CBR." International Conference on Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2006.

Vince, R., Portfolio management formulas: mathematical trading methods for the futures, options, stock markets, Vol. 1. John Wiley & Sons, 1990.

Wang, Y., and Ishii, N., “A method of similarity metrics for structured representations.” Expert Systems with Applications, Vol.12, No.1, 1997, pp.89-100.

Wilder Jr, J. W., “The Relative Strength Index,” Journal of Technical Analysis of Stocks and Commodities, Vol.4, 1986, pp.343-346.

Williams, L., “The Ultimate Oscillator.” Technical Analysis of Stocks and Commodities, Vol.3, No.4, 1985, pp.140-141.

Wilson, D.L., “Asymptotic properties of nearest neighbor rules using edited data”, IEEE Transactions on Systems, Man, and Cybernetics, Vol.2, No.3., 1972, pp.408-421.

Zhang, Y., & Wu, L., “Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network.” Expert systems with applications, Vol.36, No.5, 2009, pp.8849-8854.

한 현 웅 (Han, Hyun-Woong)

현재 국민대학교 비즈니스IT전문대학원에서 박사과정에 재학 중이다. 세종대학교 산업대학원 디지털정보산업학석사 학위를 취득하였으며, 주요 관심분야는 기업리스크, 기업신용평가 시스템이다.

안 현 철 (Ahn, Hyun-Chul)

현재 국민대학교 비즈니스IT전문대학원 부교수로 재직중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학 석사와 박사를 취득하였다. 주요 관심분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.

<Abstract>

System Trading using Case-based Reasoning based on Absolute Similarity Threshold and Genetic Algorithm

Han, Hyun-Woong · Ahn, Hyun-Chul

Purpose

This study proposes a novel system trading model using case-based reasoning (CBR) based on absolute similarity threshold. The proposed model is designed to optimize the absolute similarity threshold, feature selection, and instance selection of CBR by using genetic algorithm (GA). With these mechanisms, it enables us to yield higher returns from stock market trading.

Design/Methodology/Approach

The proposed CBR model uses the absolute similarity threshold varying from 0 to 1, which serves as a criterion for selecting appropriate neighbors in the nearest neighbor (NN) algorithm. Since it determines the nearest neighbors on an absolute basis, it fails to select the appropriate neighbors from time to time. In system trading, it is interpreted as the signal of 'hold'. That is, the system trading model proposed in this study makes trading decisions such as 'buy' or 'sell' only if the model produces a clear signal for stock market prediction. Also, in order to improve the prediction accuracy and the rate of return, the proposed model adopts optimal feature selection and instance selection, which are known to be very effective in enhancing the performance of CBR. To validate the usefulness of the proposed model, we applied it to the index trading of KOSPI200 from 2009 to 2016.

Findings

Experimental results showed that the proposed model with optimal feature or instance selection could yield higher returns compared to the benchmark as well as the various comparison models (including logistic regression, multiple discriminant analysis, artificial neural network, support vector machine, and traditional CBR). In particular, the proposed model with optimal instance selection showed the best rate of return among all the models. This implies that the application

of CBR with the absolute similarity threshold as well as the optimal instance selection may be effective in system trading from the perspective of returns.

Keyword: Case-based reasoning, Absolute similarity threshold, Feature selection, Instance selection, System trading, Genetic algorithm

* 이 논문은 2017년 6월 14일 접수, 2017년 8월 22일 1차 심사, 2017년 9월 28일 게재 확정되었습니다.