ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 December: 23(4): 147~168 http://dx.doi.org/10.13088/jiis.2017.23.4.147

입력변수 및 학습사례 선정을 동시에 최적화하는 GA-MSVM 기반 주가지수 추세 예측 모형에 관한 연구

이종식
한화자산운용㈜
(*jslee1@hanwha.com*)

안현철
국민대학교 비즈니스IT전문대학원
(*hcahn@kookmin.ac.kr*)

오래 전부터 학계에서는 정확한 주식 시장의 예측에 대한 많은 연구가 진행되어 왔고 현재에도 다양한 기법을 응용한 예측모형들이 연구되고 있다. 특히 최근에는 딥러닝(Deep-Learning)을 포함한 다양한 기계학습기법(Machine Learning Methods)을 이용해 주가지수를 예측하려는 많은 시도들이 진행되고 있다. 전통적인 주식투자 거래의 분석기법으로는 기본적 분석과 기술적 분석방법이 사용되지만 보다 단기적인 거래예측이나 통계학적, 수리적 기법을 응용하기에는 기술적 분석 방법이 보다 유용한 측면이 있다. 이러한 기술적 지표들을 이용하여 진행된 대부분의 연구는 미래시장의 (보통은 다음 거래일) 주가 등락을 이진분류-상승 또는 하락-하여 주가를 예측하는 모형을 연구한 것이다. 하지만 이러한 이진분류로는 추세를 예측하여 매매시그널을 파악하거나, 포트폴리오 리밸런싱(Portfolio Rebalancing)의 신호로 삼기에는 적합치 않은 측면이 많은 것 또한 사실이다. 이에 본 연구에서는 기존의 주가지수 예측방법인 이진 분류 (binary classification) 방법에서 주가지수 추세를 (상승추세, 박스권, 하락추세) 다분류 (multiple classification) 체계로 확장하여 주가지수 추세를 예측하고자 한다. 이러한 다분류 문제 해결을 위해 기존에 사용하던 통계적 방법인 다항로지스틱 회귀분석(Multinomial Logistic Regression Analysis, MLOGIT)이나 다중판별분석(Multiple Discriminant Analysis, MDA) 또는 인공신경망(Artificial Neural Networks, ANN)과 같은 기법보다는 예측성과의 우수성이 입증된 다분류 Support Vector Machines(Multiclass SVM, MSVM)을 사용하고, 이 모델의 성능을 향상시키기 위한 래퍼(wrapper)로서 유전자 알고리즘(Genetic Algorithm)을 이용한 최적화 모델을 제안한다. 특히 GA-MSVM으로 명명된 본 연구의 제안 모형에서는 MSVM의 커널함수 매개변수, 그리고 최적의 입력변수 선택(feature selection) 뿐만이 아니라 학습사례 선택(instance selection)까지 최적화하여 모델의 성능을 극대화 하도록 설계하였다. 제안 모형의 성능을 검증하기 위해 국내 주식시장의 실제 데이터를 적용해본 결과 ANN이나 CBR, MLOGIT, MDA와 같은 기존 데이터마이닝 기법들이나 인공지능 알고리즘은 물론 현재까지 가장 우수한 예측 성과를 나타내는 것으로 알려져 있던 전통적인 다분류 SVM 보다도 제안 모형이 보다 우수한 예측성과를 보임을 확인할 수 있었다. 특히 주가지수 추세 예측에 있어서 학습사례의 선택이 매우 중요한 역할을 하는 것으로 확인 되었으며, 모델의 성능의 개선효과에 다른 요인보다 중요한 요소임을 확인할 수 있었다.

주제어 : 다분류 SVM, 유전자 알고리즘, 입력변수 선택, 학습사례 선택, 주가지수 추세 예측

논문접수일 : 2017년 10월 31일 논문수정일 : 2017년 11월 19일 게재확정일 : 2017년 12월 2일
원고유형 : 일반논문 교신저자 : 안현철

1. 서론

주식시장에 있어서 예측의 대상은 거래의 목적에 따라 다양하다고 볼 수 있다. 주식 가치 상승에 대한 투자라면 일반적으로 주가지수(또는 개별주식)에 대한 상승 또는 하락에 대한 예측이

주요 목표가 될 수 있고, 또한 옵션투자와 같이 변동성에 대한 투자라면 변동성 지수의 변화폭에 대한 예측이 주요 연구목표가 될 수 있다. 이와 유사한 내용으로 근래 들어 주식시장의 추세를 예측하는 것이 중요한 실무적인 이슈로 떠오르고 있다. 이는 특히 로보어드바이저(robo-advisor)와 같이 고객의 자산을 자동으로 배분하고, 조정하는 알고리즘에 트리거(trigger) 역할을 하는 중요한 이벤트 생성자이기 때문이다. 다시 말하면 추세의 전환이 예측될 경우 고객의 포트폴리오는 어떤 식으로든 조정이 필요하고, 심지어 추세가 지속되거나 예측이 힘든 경우라도 고객의 포트폴리오는 조정이 필요한데, 그러한 포트폴리오의 조정 알고리즘을 수행시켜주는 역할이 바로 추세 전환이라는 이벤트라고 할 수 있는 것이다. 따라서 이러한 추세를 예측하는 것은 실무적으로 매우 중요한 작업이며, 투자수익률을 결정하는 주요 요인의 하나로 볼 수 있는 것이다.

주가 예측 모델은 다양한 통계 및 인공 지능 기법을 사용한다. 이러한 기법에는 판별 분석(Discriminant Analysis), 로지스틱 회귀(Logistic Regression), 의사 결정 트리(Decision Tree), k-nearest neighbor(KNN) 및 딥뉴럴네트워크(Deep-Neural Network, DNN) 알고리즘 등이 포함된다. 그 중 최근에 딥뉴럴네트워크는 이미지 분류에 있어 탁월한 예측 정확도를 보여주고 있는데 이를 응용해 주가 예측 연구에도 사용되고 있다. 그러나 많은 금융 회사는 여전히 DNN 사용에 어려움을 겪고 있다. DNN은 대용량 데이터 샘플이 필요하고, 과적합(overfitting) 가능성이 높으며, 타겟변수의 값이 많아질수록 정확도가 떨어지고, 결과에 대한 설명력이 부족한 (일종의 블랙박스 알고리즘) 것과 같은 딥뉴럴네트워크의 고유의 한계를 갖고 있기 때문이다.

SVM (Support Vector Machine)은 이러한 딥뉴럴네트워크의 한계를 해결할 수 있는 대안이 될 수 있다. 일반적인 딥뉴럴네트워크 모델은 오분류 에러 또는 학습 데이터의 정확한 결과값과의 편차를 최소화하기 위한 경험적 위험 최소화 원칙을 구현한다. 그러나 SVM은 일반화 오류의 상한선을 최소화하는 구조적 위험 최소화를 원칙으로 구현한다. 또한, SVM의 솔루션은 글로벌 최적일 수 있지만 딥뉴럴네트워크 알고리즘은 로컬 최적해로 빠질 가능성이 있다. 결과적으로 주가 예측에 대한 최근의 여러 연구에서 SVM을 분류자로 사용하여 시스템 트레이딩 모델이나 주가지수의 변동성 예측에도 사용될 수 있음을 보여주었다(Kim and Ahn, 2010; Ra et al., 2016).

SVM 솔루션에는 예측 성능에 영향을 주는 몇 가지 요인이 있다. 이러한 요인은 일반적으로 경험적 방법에 의해 설정된다. 특히 SVM에서 적절한 커널함수 및 매개 변수 (예: C, d, σ^2) 그리고 적절한 입력변수 선택(feature selection)에 대한 문제는 분류를 기반으로 하는 예측문제에 대해서는 대중적인 연구 주제였다. 이러한 요인 외에도 적절한 학습사례 선택(instance selection)은, 관련성이 없고 왜곡된 학습사례를 제거하여 SVM의 분류 정확도를 향상시킬 수 있는 주요 요인중의 하나이다. 특히 제대로 된 학습사례를 선택하게 되면 수행시간을 급격하게 단축시킬 수 있어 컴퓨터 연산자원(computing resource)의 효율적인 사용 측면에서도 매우 중요한 요인으로 볼 수 있다. 그럼에도 불구하고 SVM에 학습사례 선택을 적용한 연구는 거의 없었으며, 특히 주가 예측 영역에서는 더욱 찾아보기가 쉽지 않은 것이 현실이다.

Kim and Ahn (2011)의 연구에서는 기업의 신

용등급 분류에 적용하는 모델에 GA-MSVM 모형을 적용하였는데, 여기서는 커널함수 매개변수와 입력변수 선택을 동시에 최적화하도록 모델 설계를 제안하였다. 이의 근거는 Chatterjee (2013)의 연구에서 적절한 커널함수 매개변수가 MSVM의 성능을 향상시킬 수 있음을 보여주었고, Hong and Park (2011), Chatterjee (2013) 등의 연구 결과에서도 적절한 입력변수 선택이 예측 정확도를 높일 수 있음을 보여주었다. Kim and Ahn (2011)의 연구에서는 주식시장의 예측모형에 사례선택기법을 SVM기법과 결합하여 예측하는 모델을 제안하였다. 학습사례 선택은 전체 학습용 데이터에서 적절히 축약된 하위 집합을 선택하고 이러한 하위 집합만을 학습에 사용하는 방법이다. 이는 서포트벡터로 잡음이 많은 학습사례를 선택할 가능성을 줄임으로써 SVM의 왜곡된 훈련을 방지할 수 있게 되고, 이로 인해 SVM의 분류 정확성을 보다 향상시킬 수 있게 하는 것이다. 이러한 효과적인 장점으로 인해 신경망 (Reeves and Taylor, 1998)과 사례 기반 추론 (Babu and Murty, 2001)을 포함한 다양한 분류 기법에 적용되어 왔다(Ahn et al., 2006).

본 연구에서는 이러한 사전연구를 토대로 하여 커널함수 매개변수와 입력변수 선택, 그리고 학습사례 선택까지 동시에 최적화하여 이를 주식시장 추세 예측에 적용할 것을 제안한다. 주식시장을 대변하는 주요 지표로서 KOSPI200지수를 사용하고 이 지수의 가격정보 및 이를 이용한 파생변수, 그리고 금리 및 환율, 미국의 주가지수 등의 정보를 사용하여 비교적 단기 추세를 예측하는 모형을 제안한다. 추세에 대한 기준은 25일 MA(Moving Average)를 사용하여 주간 추세(weekly trend)를 계산한다. 추세의 분류는 추세상승, 추세하락, 박스권(boxed)의 세 가지로 구분되며(Dash and Dash, 2016), 당일의 데이터로 익일의 추세를 예측하는 것이 실험의 주요 목표가 된다.

본 연구는 다음과 같이 구성된다. 2장에서는 SVM을 이용한 전통적인 이분류 기법으로 모델을 제시한 연구 및 MSVM을 적용한 연구들에 대해 살펴보고 MSVM과 GA알고리즘에 대해 설명한다. 3장에서는 유전자알고리즘(GA)과 MSVM을 결합한 모형의 프로세스에 대해 설명하고, 4장 실증분석에서는 실험설계에 대한 설명 및 학습과 검증에 사용한 데이터의 소개와 GA-MSVM모형의 비교모델들 대비 우위점에 대하여 기술할 것이다. 마지막 5장에서는 연구의 결론과 함께, 본 연구의 한계점에 대해 논의하고 향후의 방향성에 대해서도 약술할 것이다.

2. 관련 연구

본 연구에서 제안하는 모형은 OAO(One-Against-One) 기반의 다분류 SVM과 유전자 알고리즘이 결합된 형태의 모델이다. 이에 기존 문헌의 검토에서는 우선 MSVM의 기본적인 원리와 개념에 대해 먼저 설명한 이후, 유전자 알고리즘에 대한 기본적인 소개를 하고 이와 더불어 MSVM과 GA를 결합하여 성과를 향상시키고자 한 기존 연구들에 대해 살펴보도록 한다.

2.1 다분류 SVM(Multi-classification SVM, MSVM)

서포트 벡터 머신(support vector machine, SVM)은 기계 학습의 알고리즘중 하나로 패턴 인식, 자료 분석을 위한 지도학습(supervised training)

모델이며, 주로 분류와 회귀 분석을 위해 사용한다(Vapnik, 1995). 러시아의 통계수리학자인 Vapnik에 의해 개발된 분류알고리즘으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간에서 선형문제로 대응시켜 적은 수의 샘플만으로도 상대적으로 우수한 분류성과를 도출할 수 있는 가장 효율적인 분류알고리즘중의 하나이다(Ahn et al., 2006; Vapnik, 1995). 두 개의 카테고리 중 어느 한 카테고리에 속한 데이터 집합이 주어졌을 때, SVM은 주어진 데이터 집합을 기반으로 하여 새로운 데이터가 어느 카테고리에 속하는지를 판단하는 비확률적 이진선형분류 모델을 만들어 준다. 이렇게 만들어진 모델은 데이터가 사상된 공간에서 하나의 경계로 표현되는데 SVM 알고리즘은 그 경계 중 가장 큰 폭(margin)을 갖는 경계를 찾는 알고리즘이다. 이러한 SVM은 구조적 위험을 최소하는 원리를 가지고 있기 때문에, 과적합(overfitting)의 위험성이 높은 인공신경망과 비교해 이론적으로는 더 우월한 기계학습 알고리즘으로 인식되고 있다. 하지만, 전통적인 SVM은 이분류(Binary-Classification) 문제만 해결할 수 있도록 설계되어 있기 때문에 이를 다분류 문제에 적용하기 위해서는 SVM을 변형하거나 혹은 확장시킬 필요가 있다(Lorena and de Carvalho, 2008; Wu et al., 2007). 이진분류법을 확장해서 멀티클래스로 분류 하는 대표적인 방법으로 OAO(One-Against-One), OAA(One-Against-All), WW(Weston & Watkins), CS(Crammer & Singer) 방법 등이 있다.

2.1.1 OAA, OAO 방법

일반적으로 OAA(One-Against-All)방법은 이분류 SVM모형을 다수개 만들어 결합하는 방식으로, 여기에는 우선 분류할 클래스 개수만큼의 SVM 모형을 구축하여 각 클래스에 해당하는 그룹과 해당하지 않는 그룹을 판단하는 방식으로 수행하는 것이 OAA이다.

OAO(One-Against-One) 방법은 분류할 전체 클래스에 대해 구성 가능한 모든 쌍(pair)별로 독립된 SVM 모델을 구축하는 방법이다. 이러한 두 가지 방법가운데 일반적으로 OAO가 OAA에 비해 수행의 효율성은 낮지만 예측성과의 측면에서는 조금 더 우수한 방법인 것으로 연구 되었다. (Ahn and Kim, 2009; Ahn et al., 2006)

이 밖에도 OAO와 유사하게 모든 클래스 페어(class pair)별로 독립된 SVM 모형을 구축하지만 이후 DAG(Directed Acyclic Graph)를 이용하여 클래스 구분을 예측하도록 설계된 DAGSVM기법은 여러 연구들에서 OAO기법만큼 뛰어난 성능을 보이는 것으로 평가 받고 있다(Ahn and Kim, 2009; Kim and Ahn, 2012; Lorena and de Carvalho, 2004; Wu et al., 2008).

2.1.2 WW, CS 방법

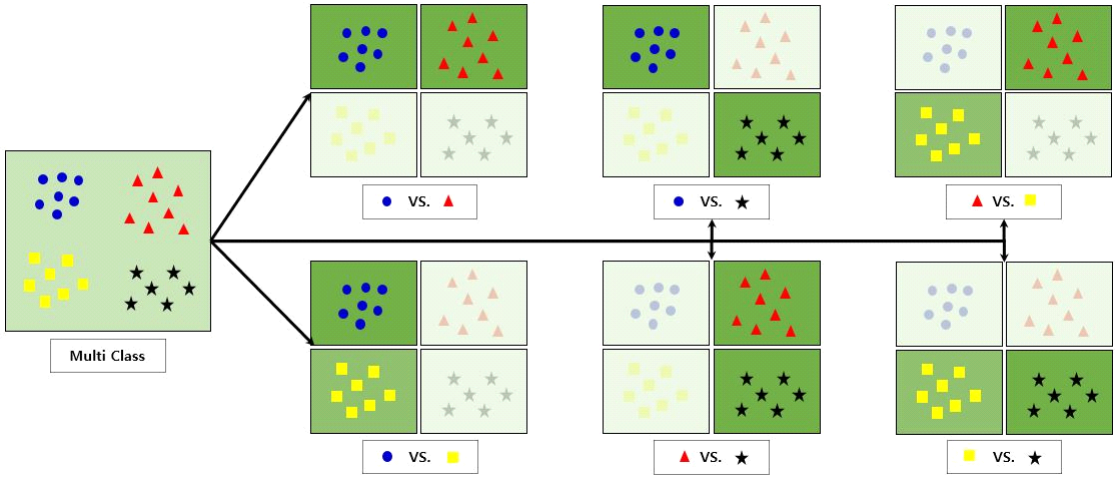
WW(Weston & Watkins), CS(Crammer & Singer)방법은 SVM을 클래스별로 고려하는 것이 아니라 모든 클래스를 한 번에 일괄적으로 반영하여 하나의 최적화 문제로 해결하는 방법을 말한다. 이 두 가지 방법은 유사해 보이지만, CS방법의 경우 분류평면 도출을 위해 사용되는 최적화 모형의 제약식에 있어 상대적으로 더 적은 개수의 여유 변수(slack variable)를 요구한다는 점에서 차이가 있다(Crammer and Singer, 2000). 이렇게 모든 클래스를 한 번에 모두 고려하여 하나의 최적화 문제로 해결하는 방법들은 분류평면을 찾는 수행과정이 수학적으로 상당

히 복잡할 뿐만 아니라 이런 복잡성의 이유로 인해 컴퓨터 연산자원이 굉장히 다량으로 소요되는 것이 단점이다(Hsu and Lin, 2002; Lorena and de Carvalho, 2008). 이러한 컴퓨터 연산자원의 문제로 인해 대부분의 MSVM 연구들은 분할정복(divide and conquer) 방식의 OAA나 OAO, DAGSVM과 같은 기법들을 사용하고 있다(Ahn and Kim, 2009; Lorena and de Carvalho, 2008).

본 연구에서는 상기의 접근법들 중에서, 활용의 효율성 측면과 상대적인 예측의 정확도 등을 감안하여 일반적으로 가장 많이 활용되고 있는 OAO 방법을 사용한다. 다음의 <Figure 1>은 이러한 OAO 방법의 원리를 그림으로 나타내고 있다. 아래 그림을 보면, 모두 4종류의 클래스로 구성된 다분류 문제에 OAO방식을 적용한 예시인데, 모두 6개의 이분류 SVM 모형이 분할/구축되고, 분할된 모델들의 분류결과를 투표(voting)로 합산하여 최종 클래스를 예측하게 하는 방식이다.

2.2 유전자 알고리즘(GA)과 MSVM 결합모형

최근 각종 기계학습 기법의 최적화 문제와 관련한 연구들에서 최적화를 위한 방법으로 가장 많이 시용되고 있는 알고리즘 중 하나가 유전자 알고리즘이다. 유전자 알고리즘(Genetic Algorithm, GA)은 자연계의 진화과정을 모사한 계산 알고리즘으로서 1975년에 존 홀랜드(John Holland)에 의해서 개발된 전역 최적화 기법이다. 이러한 GA기법은 생명체의 진화를 모방한 진화 연산의 가장 대표적인 기법으로, 실제 생물학적 진화의 과정에서 많은 부분을 차용하였으며, 선택, 돌연변이, 교배 연산 등이 존재한다. GA는 자연계의 생물 유전학에 기반한 이론을 바탕으로, 병렬적이고 전역적인 탐색 알고리즘으로서의 특징을 갖는다. GA는 해결하고자 하는 문제에 대한 가능한 모든 해들을 정해진 구조의 자료형태로 표현한 다음, 이 자료들을 순차적으로 변형함으로써 점점 보다 더 좋은 해들을 만들



〈Figure 1〉 Multiclass SVM conceptual diagram of OAO scheme

어 낸다. 여기서 해를 나타내는 자료구조는 유전자(chromosome)이며, 이들을 변형시킴으로써 점점 더 좋은 해에 다가가는 과정을 진화라는 단어로 표현할 수 있는 것이다. 즉, GA는 어떤 함수 Y = f(x)를 최적화하는 해 x를 찾기 위해, 진화의 형태를 모방한(simulated evolution) 탐색 알고리즘이라고 할 수 있다. GA는 복잡하고 방대한 공간을 탐색하면서, 최적 혹은 근사최적 결과를 찾아주는 확률적 방법을 이용하는 특징으로 인해 다양한 제약조건을 포함한 상황에서 목적 함수(objective function)를 최적화 하는 매개변수 추정에 널리 적용되고 있다(Kim et al., 2006; Shin and Han, 1999).

이러한 GA는 지금까지 전통적인 이분류 SVM의 최적화를 위한 도구로 다양한 연구에서 적용되어 왔다. Pai and Hong (2005)은 GA를 사용하여 SVM의 커널함수에 사용되는 자유매개변수(free parameters)를 최적화했다. 이 연구의 SVM 모델은 Gaussian RBF를 커널 함수로 사용하였고 GA를 사용하여 C, σ^2, ε 매개변수를 최적화하기 위한 시스템을 제안했다. Howley and Madden (2005)은 한 단계 더 나아가 커널함수 매개변수(예 : C, σ^2, d, ε) 뿐만 아니라 커널함수 자체까지 최적화 하는 모델을 제안했다. 결과적으로, 이들의 모델은 전역적으로 최적화된 커널함수와 매개변수를 제공할 수 있는 모델이라 할 수 있다. 이러한 연구의 패턴에서 조금 더 발전한 모델이 매개변수의 최적화와 동시에 입력변수 선택의 최적화를 시도한 연구들이다. 상기의 연구에서 보았듯이 매개변수와 입력변수 선택은 모두 SVM의 분류 성능에 크게 영향을 미치는 요인들이다. 따라서 이러한 요인들을 동시에 최적화 한다면 개별 최적화 보다 효과적이고 우수한 실험결과를 얻을 수 있을 것이다. 그러나 이 두 가지 요인을 동시에 최적화 하려는 관련 연구는 아직은 많이 찾아보기가 힘들다. Jack and Nandi (2002)가 기계고장 검출을 위해 이 기법을 적용하였고, Kim et al. (2005)이 네트워크 침입 탐지 연구에 사용된 예가 있다. Zhao et al. (2005)는 SVM을 GA로 최적화 하려는 다른 시도였는데 입력변수 선택의 최적화를 시도한 연구이다. 입력변수 선택은 목표종속변수(target)와 관련이 있는 입력변수 중 효과적인 일부 입력변수만을 사용하는 방법이다. 대부분의 분류 문제에서 적절한 입력변수 선택은 각각의 샘플을 보다 정확하게 특성화함으로써 분류 성능을 향상시키고 계산 시간을 단축시키기 때문에 중요한 요인으로 인식되고 있다. 따라서 많은 연구자들이 GA를 사용하여 SVM의 입력변수 선택을 최적화하려고 시도하였다. 그러한 연구사례로는 Li et al. (2004)이 암(cancer) 검진에 사용한 예가 있고 Lee and Byun (2003)과 Sun et al. (2004)은 이미지 식별을 위해 이 기법을 사용하기도 하였다.

Thi et al. (2009)는 입력변수 선택 최적화기법을 시계열데이터인 주가지수와 개별 주식 예측에 적용해 본 결과 기존 SVM의 성능개선에 많은 효과가 있음을 입증하였다. Chen and Hsiao (2007)는 재무비율 데이터만을 사용하여 기업의 위기상황을 알려주는 일종의 조기경보 모델을 제안하였다. 이 연구에서는 위의 연구들과 마찬가지로 GA와 SVM을 결합하여 커널함수 매개변수 최적화와 동시에 기업의 위기를 알려주는 재무데이터의 입력변수를 6개로 변별하여 95% 이상의 예측 정확도로 보여줄 수 있었다.

GA를 사용하여 SVM의 다양한 요인을 최적화하려는 많은 연구가 이전에 있었지만 또 다른 최적화가 필요한 요인이 바로 학습사례 선택(instance selection)의 문제이다. 학습사례 선택은

학습 데이터를 적절한 방법으로 축소하고 이렇게 축소된 데이터를 이용하여 모델을 학습시키는 방법이다. 이는 서포트벡터로 잡음이 많은 학습용 데이터를 선택할 가능성을 줄임으로써 SVM의 왜곡된 훈련을 방지하게 할 수 있으며 이로 인해 SVM의 분류성능을 향상시킬 수 있다. 학습사례 선택의 이러한 장점으로 인해 인공신경망 기법이나 사례기반추론 기법을 포함한 다양한 분류 기법에 적용되어 왔다.

Ahn and Kim (2011)은 특히 데이터량이 많고 심한 잡음 특성을 보이는 주식시장의 데이터를 이용하여 SVM과 GA 기반의 사례선택기법(instance selection technique)을 결합한 분석모델을 제안하였다. 이 연구에서 제안된 결합 모형을 전통적인 SVM, 매개변수 최적화 SVM, 인공신경망, 사례기반추론(1-NN), 로지스틱 회귀분석 모형과 비교한 결과, 보다 정확한 분류성능을 나타냈으며 특히 샘플 데이터의 수가 50% 가까이 줄어듦으로써 보다 효율적인 분석이 가능함을 보여주었다. 위의 연구들에서 한 단계 더 나아간 연구사례를 보면 Ahn et al. (2006)은 기업부도예측모델을 구축하는데 있어 GA를 사용하여 SVM에 대한 커널함수의 매개변수뿐만이 아니라 입력변수 선택과 학습사례 선택을 동시에 최적화하는 전역최적화 모델을 제안하였다. 이 연구에서 제안된 모형은 일반적으로 커널함수 매개변수와 입력변수 선택만을 최적화한 모델에 비해 예측의 정확성이 높은 것으로 나왔으며, 특히 입력변수 선택 보다도 학습사례 선택이라는 요인이 SVM의 성능개선에 더 큰 영향을 준다는 것을 보여주었다.

이처럼 현재까지 GA를 활용해 일반적인 이분류 SVM을 최적화하려는 시도는 상당히 많이 이루어진 편이다. 하지만, GA또는 다른 최적화 기법을 활용하여 MSVM의 최적화를 시도한 연구는 상대적으로 찾아보기가 쉽지 않다. MSVM의 최적화를 시도한 대표적인 기존 연구로는 Lorena and de Carvalho (2008)가 있다. 이 연구는 적절한 커널함수의 매개변수가 MSVM의 성능을 향상시킬 수 있음을 보여주었다. 일반적으로 SVM 분류기의 성능은 매개변수의 값 선택에 따라 달라진다. 이 연구는 일반적인 멀티 클래스 분류에서 이진분류(binary-classification) SVM의 매개변수를 조정(tuning)하기 위해 GA를 활용한다. GA는 모든 이진 분류기에 공통인 매개변수 값 지정 또는 각 이진 분류기에 대한 차별화된 값을 찾아낼 수 있다.

Shieh and Yang (2008)은 모바일폰의 12가지 입력변수(연속적인 특성 네 가지와 이산적 특성 요인 8가지)를 사용하여 스타일을 5개의 클래스로 구분하는 SVM-RFE (SVM recursive feature elimination)모델을 제안하였다. 연구 결과 총 12가지의 특성요인 중 7가지의 특징요인으로만 오류율 1% 미만의 매우 높은 예측 정확도를 갖는 분류모델을 구축 할 수 있었다. Chatterjee (2013)의 암석 유형을 분류하는 연구에서는SVM 모델의 하이퍼 매개변수와 입력변수 개수는 GA에 의해 선택된다는 것을 보여주었다. 그 결과 SVM 모델은 189개의 추출된 입력변수 중에서 40개의 입력변수가 선택되었을 때 가장 예측정확도가 높음을 보여주었다. 결과는 암석 분류에 대해 제안된 기법의 전체 정확도가 96.2%라는 비교적 높은 정확도를 보여주었으며 비교 모형으로 제시된 신경망(ANN) 모델보다 제안된 SVM 모델이 우수하게 수행되었음을 보여주었다. Ahn (2014)은 기업의 신용등급을 예측하는데 있어 Multiclass SVM(다분류SVM, MSVM)과 GA를 활용한 최적화 모델을 제안하였다. 이 연구의 제

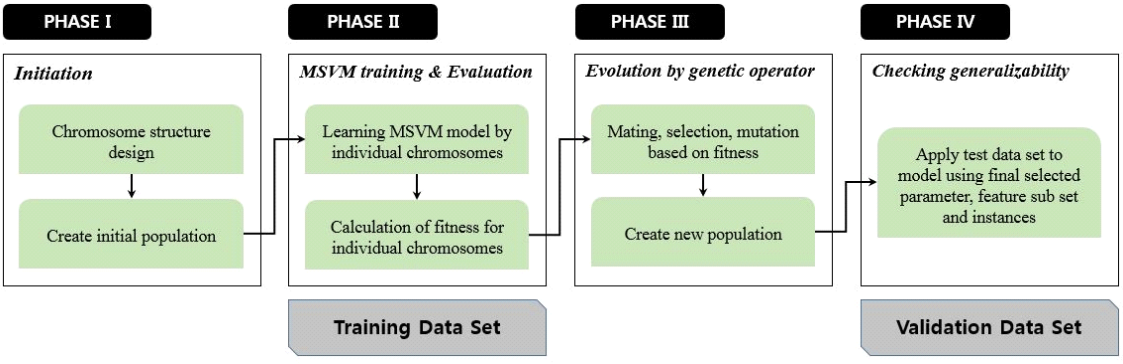
안 모형은 GA를 활용하여 Multiclass SVM에 적용되는 커널함수의 매개변수 및 최적의 입력변수들을 선택하도록 설계되었다. 이러한 모델에 실제 데이터를 적용하여 실험한 결과, 다항 로지스틱 회귀분석 (MLOGIT)이나 다중판별분석 (MDA), 사례추론기반기법 (CBR), 인공신경망기법 (ANN) 같은 기존의 알고리즘 보다 우수한 성과를 보이는 것으로 나타났다.

GA로 MSVM의 매개변수 및 학습사례 선택을 최적화하려고 시도한 연구도 있다. Chen (2013)은 많은 수의 클래스와 학습사례가 포함된 거대한 데이터 세트에 SVM를 효과적으로 적용하는 방법을 연구하였다. 특히 이 연구에서는 다중 클래스 문제에 대한 학습사례 선택 방법을 제안하였다. 학습사례 선택의 효율성을 높이기 위하여 학습사례를 군집화(clustering) 하는 방법을 제안하였고, 이는 대부분의 데이터 세트에서 가장 높은 분류 정확도를 유지한다는 것을 보여주었다. 즉, 다중 분류뿐만이 아닌 이진 분류에서도 우수한 성능을 발휘하였다는 것을 입증한 것이다.

이러한 기존 연구결과들로 미루어 볼 때, 유사한 수의 클래스(3개)로 구분되는 주가지수 추세 예측에서도 GA를 통한 MSVM 모형의 커널함수 매개변수 및 입력변수 선택 그리고 학습사례 선택 최적화가 성능 개선에 효과적일 수 있음을 유추해 볼 수 있다.

3. GA기반 MSVM 최적화 모형

앞장에서 살펴본 기존 연구들에 대한 분석결과를 토대로 본 연구에서는 주가지수 추세 예측을 위해 MSVM의 커널함수 매개변수와 입력변수 선택 그리고 학습사례 선택을 동시에 최적화하는 모형을 새롭게 제안한다. 특히 본 연구에서는 Chatterjee (2013)에서 그 효과가 입증된 GA를 활용해 최적화가 수행되는 모형을 제안한다. 본 연구에서는 편의상 제안 모형을 GA-MSVM (Genetic Algorithm-based Multiclass Support Vector Machine)으로 명명하고자 한다. 다음의 <Figure 2>는 본 연구의 제안모형인 GA-MSVM이 동작하는 단계를 프로세스 흐름 형식으로 보여주고 있다. 그림에서 보듯이, GA-MSVM은 다음과 같이 크게 4개 단계에 의해 구현되도록 설계되었다.



〈Figure 2〉 GA-MSVM Process

Phase 1. 단계 1은 초기화(Initiation) 단계이다. 이 단계에서는 최적화의 탐색 대상 요인들인 커널함수 매개변수 값들과 입력변수 선택, 그리고 학습사례 선택의 3가지 요인들을 염색체(chromosome) 구조 형태로 반영하고, 이러한 구조로 설계된 염색체들을 기본으로 적절한 난수를 발생시켜 초기 모집단(population)을 생성하는 작업을 수행한다.

GA를 수행하기 위해서는 탐색공간 내부의 여러 가지 변수들의 집합을 염색체라고 하는 이진 선형 스트링에 매핑(mapping)하는 과정이 필요하다. GA가 공간내부를 효율적으로 탐색할 수 있도록 그 목적에 적합한 효과적 매핑 방법을 찾는 것이 매우 중요하며, 이러한 점들을 포괄적으로 반영하여 <Figure 3>과 같이 염색체의 구조를 설계하여야 한다.

본 연구에서 제안하고 있는 GA-MSVM은 일반적으로 가장 널리 알려진 커널함수인 Gaussian RBF를 사용하고 있다. 다음의 식 (1)은 Gaussian RBF 커널함수의 산식이다.

$$\mathbf{K}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{e}^{-\frac{\mathbf{1}}{\boldsymbol{\sigma^2}}(\boldsymbol{x}-\boldsymbol{y})^2} \qquad (1)$$

따라서, 커널 함수의 매개변수 σ^2과 C를 최적화 해야 하는데, GA-MSVM에서는 이 두 가지 값에 각 14비트를 할당하여, 상당히 정밀하게 값들을 탐색할 수 있도록 설계되었다. 최적 입력변수 선택인 경우에는 '1'을 선택으로 하고, '0'을 비선택으로 치환하여 활용할 수 있기 때문에, 이러한 값들은 이진 스트링으로 표현해 주는 것이 훨씬 용이하다. 따라서 총 15개의 입력변수를 가지고 있는 주가지수추세 예측문제라고 할 경우, GA-MSVM을 적용하기 위해 필요한 염색체의 길이는 15+28=43 (비트) 이 된다. 염색체의 설계가 끝나면, 초기화 작업을 수행하게 되는데 이 작업은 난수를 사용하여 초기 모집단을 생성하는 작업이다.

Phase 2. 단계 2에서는 단계 1의 초기화 작업이 끝나고 난 이후 생성된 초기 모집단에 소속된 요인값들을 적용하여 실제 GA-MSVM모형을 학습시켜보고, 해당 염색체값이 목적에 어느 정도 부합하는지를 평가하는 적합도 함수값(fitness function)을 계산하는 과정을 수행하게 된다. 본 연구에서 제안하는 GA-MSVM 모형은 대량의 데이터를 사용하면서도 적확한 의사결정을 지원

Population	SVM Parameters														Selections													
	C							σ^2							Feature set							Instances							
	C_1	C_2	C_3	...	C_{13}	C_{14}	C_{15}	g_1	g_2	g_3	...	g_{13}	g_{14}	g_{15}	F_1	F_2	F_3	...	F_{m-2}	F_{m-1}	F_m	I_1	I_1	I_1	...	I_{n-2}	I_{n-1}	In
chromosome 1	0	1	0	...	0	0	1	1	1	1	...	0	1	0	0	0	1	...	1	1	1	0	1	0	...	0	1	0
chromosome 2	1	1	0	...	0	1	1	1	1	0	...	1	1	0	0	1	1	...	1	1	0	1	1	0	...	1	1	0
chromosome 3	0	1	1	...	1	0	0	0	1	1	...	0	1	1	1	0	0	...	0	1	1	0	1	1	...	0	1	1
chromosome 4	1	0	0	...	0	0	1	1	0	0	...	1	0	0	0	0	1	...	1	0	0	1	0	0	...	1	0	0
...				.....							.....							.....							.....			
chromosom n-1	1	0	1	...	1	1	1	0	0	1	...	0	0	0	1	0	1	...	1	1	1	0	0	1	...	1	1	1
chromosom n	1	1	1	...	1	1	0	1	1	0	...	1	0	1	1	1	1	...	1	1	0	1	1	0	...	1	1	0

〈Figure 3〉 Chromosome structure of GA-MSVM

할 수 있는 모델이 될 수 있도록 설계되어야 한다. 이러한 목적을 달성하기 위해서 분석에 사용되는 데이터마이닝 알고리즘의 목적함수와 GA의 적합도 함수를 일치시키고 GA가 적합도 함수를 최적화 시키기 위해서 탐색할 내부공간을 전체 원래 데이터로 설정해 구현해야 한다. 이러한 이유에서, GA의 적합도 함수 지표는 학습용 데이터에 대한 분류 정확도(classification accuracy)로 설정하였으며, 학습용 데이터를 대상으로 GA가 최적의 커널함수 매개변수와 입력변수 선택 및 학습사례 선택을 탐색할 수 있도록 하였다.

Phase 3. 단계 2에서 모집단에 소속된 모든 염색체들에 대한 적합도 평가가 끝나게 되면 새로운 모집단을 생성시키는 단계 3을 수행하게 된다. 단계 3은 적합도 평가가 마무리 된 염색체들 중에서 우성 염색체들을 별도로 선택(selection)하고, 이러한 우성 염색체들을 적절하게 교배(crossover)하며, 때에 따라 돌연변이(mutation)도 발생시키게 된다. 이러한 단계 3은 모델 내에 설계한 종료 조건이 충족되기 전까지 지속적으로 반복되게 함으로써, 수십세대에 걸친 진화과정을 거치며 각 요인들의 최적화값에 수렴해 가도록 한다.

Phase 4. 종료 조건이 충족될 때까지 단계 3에서 진화가 충분하게 이루어지면, 마지막 단계 4에서는 최종 선택된 요인값들을 기반으로 하여 MSVM 모형을 검증용 데이터에 적용해 보고, 예측성과를 최종적으로 점검하는 작업을 수행하게 된다. 이러한 과정을 통하여, 제안모형이 새로운 데이터(validation data)의 예측에 있어서도 탁월한 성과를 보여주는지 확인하여, 제안모형의 일반화 가능성을 검증한다.

4. 실증 분석

4.1 실험 설계

본 연구에서는 GA-MSVM을 검증하기 위해 2004년 4월 2일부터 2017년 2월10일까지의 KOSPI200 데이터 및 이와 관련된 국내외 주식시장 일간(daily) 가격/기술 지표28개와 학습사례 3,192개로 구성하였다. 실험을 위한 학습용 데이터셋(training data set)은 전체 데이터의 70%인 2004년 4월 2일부터 2013년 3월22일까지의 데이터로 2,234개의 데이터를 사용하였고, 검증용 데이터셋(validation data set)은 2013년 3월 25일부터 2017년 2월 10일까지 958개의 데이터를 사용하였다.

종속변수로는 익일의 추세를 적용하였다. 이 때, 추세의 분류는 Upward trend, Boxed, Downward trend의 세 가지 등급으로 하였으며 이에 대한 규칙은 Dash and Dash(2016)의 연구를 참조하여 국내 주식시장 환경에 맞게 조정하였다.

본 연구에서의 추세 분류 규칙은 다음과 같다.

[규칙1: Upward Trend] 당일 종가가 당일 MA_{25}(25일이동평균) 보다 크고, 최근5일 상승 추세

[규칙2: Downward Trend] 당일 종가가 당일 MA_{25}(25일이동평균) 보다 작고, 최근5일 하락 추세

[규칙3: Boxed] 규칙1과 규칙2에 해당되지 않는 경우 (추세라고 보기 어려운 경우)

단순이동평균(Simple Moving Average, MA)은 다음과 같은 산식으로 산출되며, t는 25를 사용한다.

$$MA = \frac{1}{t}\sum_{i=1}^{t} cp(i) \quad (2)$$

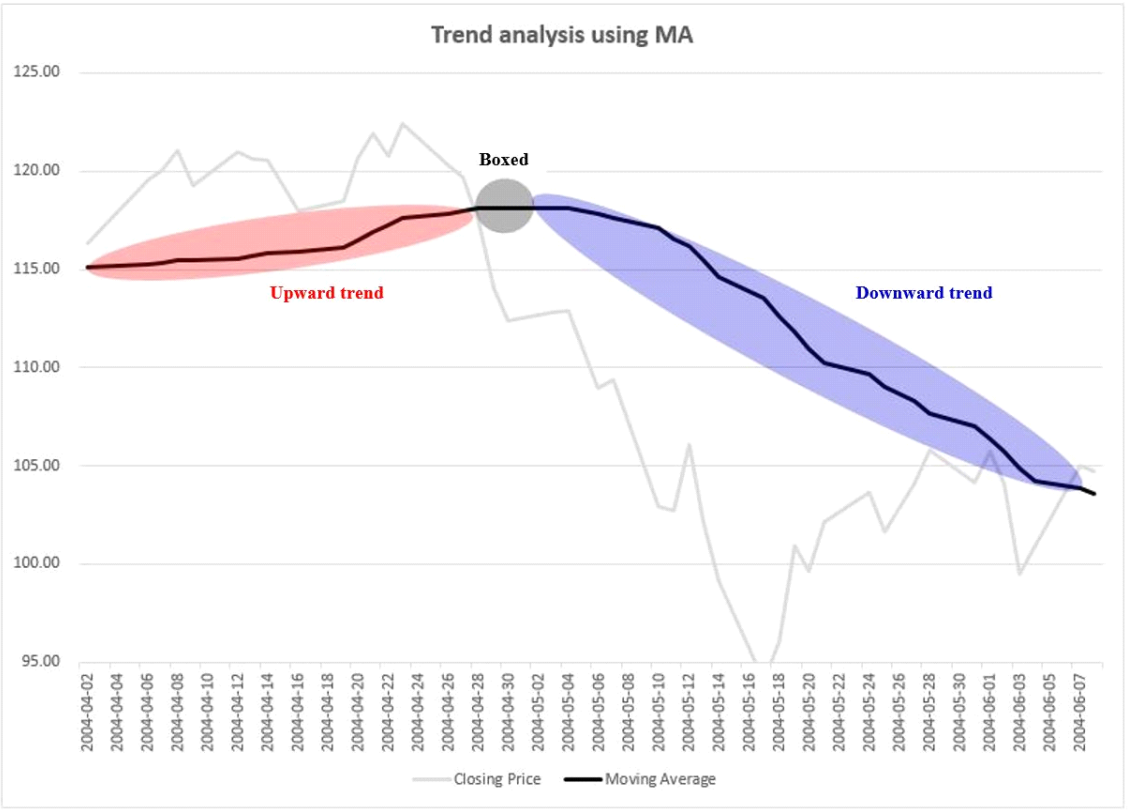
(단, $cp(i)$는 i일의 종가)

〈Table 1〉 Entire data set

	Year	Up	box	Down	Sum of year	Ratio
Training Data Set	2004	84	33	71	188	70%
	2005	172	37	40	249	
	2006	142	57	48	247	
	2007	132	61	53	246	
	2008	55	65	128	248	
	2009	151	55	47	253	
	2010	158	49	44	251	
	2011	101	45	102	248	
	2012	115	61	72	248	
	2013	23	19	14	56	
Sum of training data		1,133	482	619	2,234	
Validation Data Set	2013	69	50	72	191	30%
	2014	89	59	97	245	
	2015	103	41	104	248	
	2016	156	48	70	274	
Sum of Validation data		417	198	343	958	
Sum of total		**1,550**	**680**	**962**	**3,192**	

〈Table 2〉 Example of trend analysis on sample data set.

Day	Closing Price	Moving Average(t = 25)	Trend Type
2004-04-02	116.35	115.10	Up
…	…	…	…
2004-04-27	119.70	117.96	Up
2004-04-28	117.48	118.13	Boxed
…	…	…	…
2004-05-03	112.84	118.14	Boxed
2004-05-04	112.92	118.11	Down
…	…	…	…
2004-06-04	100.95	104.24	Down
2004-06-07	105.02	103.88	Boxed
2004-06-08	104.73	103.57	Boxed



〈Figure 4〉 Example of trend analysis on sample data set.

실험을 위해 수집된 데이터는 총 28개 설명변수로 구성되어 있다. KOSPI200종가, KOS200종가 수익률, 거래량, 거래대금, 종가의 25일 이동평균 및 변동성 지표인 PR(Parkinson Range), GR(Garman & Klass Range), RR(Roger & Satchell Range), TR(True Range) 그리고 기술적 지표들인 slow %K, slow %D, Price ROC_종가, MACD Oscillator_12, OSCP, AD/Oscillator, Momentum, Disparity 5, Disparity 10, William %R, CCI, 그리고 외부 영향요인으로 파악되는 CAC40지수(프랑스) 수익률, DAX30지수(독일) 수익률, 미국 NASDAQ지수 수익률, DOW지수 수익률, 원달러 변동률, 단기국채 수익률, 장기국채 수익률, CMO(Collateralized Mortgage Obligations) 수익률 등이다. 기술적 지표들의 선정은 Lee (2008)의 연구에서 사용한 지표들을 참고하였다. 추가적으로 추세에 영향을 줄 수 있는 모든 가능성을 파악하기 위해 변동성 지표로서 가격범위지표를 도입하였다. 이러한 가격범위지표는 하루의 주가 자료 중 시가(open), 고가(high), 저가(low), 종가(close)자료를 이용하며, 주로 변동성 추정에 활용된다. Kim (2011)은 2009년 새롭게 개발된 VKOSPI지수와 가격범위지표의 비교실증 분석결과 두 지표 사이의 실현

변동성에 대한 설명력에는 유의한 차이가 없는 것으로 나타나 실제 실무업무에서 간단하게 계산하여 업무에 적용 가능한 가격범위지표의 유용성을 보여주었다. 본 연구에서 1차적으로 후보로 선정된 가격범위지표의 종류는 PR, GR, RR, TR의 네 가지이며 각각의 산출식은 <Table 3>과 같다. (단, t시점의 시가는 O_t, 고가는 H_t, 저가는 L_t, 종가는 C_t)

이렇게 수집된 28개의 변수를 1차적으로 유의한 변수들을 선정하기 위해 one-way ANOVA와 MDA의 stepwise selection을 이용해 분석한 결과 최종적으로 <Table 5>와 같이 유의한 변수 15개를 선정하였다.

<Table 6>은 MDA분석에 의하여 추출된 변수들이다.

〈Table 3〉 Candidate technical indicators

X1	X2	X3	X4	X5
Closing Price	Return of KOSPI200	Transaction Volume	Transaction Amount	MA25
X6	X7	X8	X9	X10
PR	GR	RR	TR	Slow %K_5,3
X11	X12	X13	X14	X15
Slow %D_3	Price ROC ,14	MACD Oscillator_12,26,9	OSCP	AD/Oscillator
X16	X17	X18	X19	X20
Momentum	Disparity 5	Disparity 10	William %R	CCI
X21	X22	X23	X24	X25
Return of CAC40	Return of DAX30	Return of Nasdaq	Return of DOW	US dollar exchange rate
X26	X27	X28		
Reverse short-term_12	R long-term_24	CMO_10		

〈Table 4〉 Price range indicators

Variable Name	Output formula
PR	$(H_t - L_t)^2 / 4\ln 2$
GR	$0.511(H_t - L_t)^2 - 0.019[(C_t - O_t)(H_t + L_t - 2O_t) - (H_t - O_t)(L_t - O_t)] - 0.0383(C_t - O_t)^2$
RR	$(H_t - O_t)(H_t - C_t) + (L_t - O_t)(L_t - C_t)$
TR	$\mathrm{Max}[(H_t - L_t)^2, (H_t - C_{t-1})^2, (L_t - C_t)^2]$

〈Table 5〉 Selected technical indicators

No	Variable	Output formula
X2	Return of KOSPI200	$LN(C_t / C_{t-1})$
X3	Transaction Volume	None
X6	PR	(Ht − Lt)2/ 4ln2
X7	GR	$0.511(H_t - L_t)^2 - 0.019[(C_t - O_t)(H_t + L_t - 2O_t) - (H_t - O_t)(L_t - O_t)] - 0.0383(C_t - O_t)^2$
X10	Slow %K_5,3	{(current price − low price during the recent N period) / (high price during the recent N period - low price during the recent N period)} * 100
X11	Slow %D_3	M period of moving average for {(current price − low price during the recent N period) / (high price during the recent N period - low price during the recent N period)} * 100
X12	Price ROC ,14	It is an indicator to judge whether the current price is in an upward trend or a downward trend by comparing a price at a certain point in time with a current price.
X13	MACD Oscillator_12,26,9	It is an indicator of the gap between the MACD and the signal.
X15	AD/Oscillator	(Short-term n-day moving average of A / D line - m of exponential moving average of A / D line)
X18	Disparity 10	(Stock price / 10-day moving average)
X19	William %R	An indicator of where the closing price of the day is located between high and low market prices over a period of time, similar to Stochastics.
X20	CCI	It is an indicator of how recent prices are offset from moving average of average prices. It is an indicator of the intensity and direction of the trend.
X26	Reverse short-term_12	Short-term government bond yield
X27	R long-term_24	Long-term government bond yield
X28	CMO_10	Collateralized Mortgage Obligations yield

〈Table 6〉 Input / Removed Variables [a] [b] [c]

Step	Input	Remove	Wilks's Lambda							
			statistic	Df1	Df2	Df3	Exact F			
							statistic	Df1	Df2	Significance Level
1	X27		.252	1	2	2231.0	3306.128	2	2231.0	0.000
2	X26		.226	2	2	2231.0	1229.492	4	4460.0	0.000
3	X12		.219	3	2	2231.0	844.078	6	4458.0	0.000
4	X20		.216	4	2	2231.0	642.302	8	4456.0	0.000
5	X17		.210	5	2	2231.0	526.741	10	4454.0	0.000
6	X3		.207	6	2	2231.0	444.337	12	4452.0	0.000
7	X13		.205	7	2	2231.0	384.562	14	4450.0	0.000
8	X18		.204	8	2	2231.0	338.078	16	4448.0	0.000
9	X7		.203	9	2	2231.0	301.710	18	4446.0	0.000
10	X10		.202	10	2	2231.0	272.737	20	4444.0	0.000
11	X19		.199	11	2	2231.0	250.349	22	4442.0	0.000
12	X2		.198	12	2	2231.0	230.855	24	4440.0	0.000
13	X11		.197	13	2	2231.0	213.890	26	4438.0	0.000
14		X17	.197	12	2	2231.0	231.632	24	4440.0	0.000
15	X28		.196	13	2	2231.0	215.019	26	4438.0	0.000
16	X15		.195	14	2	2231.0	200.239	28	4436.0	0.000
17	X6		.194	15	2	2231.0	187.420	30	4434.0	0.000

[a] The maximum number of steps is 50.
[b] The minimum F to enter is 3.84.
[c] The maximum fraction F to be removed is 2.71

4.2 실험 결과

제안모형의 성능을 검증하기 위해 본 연구에서는 다음의 <Table 7>에서와 같이 5개의 전통적인 모형과 2개의 GA-MSVM 모형을 사용하였고, 비교모형과 제안모형의 성과비교를 위해 지표는 Hit Ratio를 사용하였다. 비교모형은 MDA, CBR, CART, ANN, MSVM 등의 전통적 모형과 FS-MSVM, IS-MSVM 모형 등이다. 우선 제안모형의 hit ratio를 보면 92.17%로 전통적인 모형들의 전체 평균 정확도인 84.59%보다 7.5%8P 높은 성과를 보여주었고 이중 가장 높은 예측정확도를 보인 MSVM의 88.52%보다도 3.65%P 더 나은 성과를 나타냈다.

이전 Ahn (2014)의 연구에서 논의되었던 바와 같이 GA-FSMSVM 모형과 GA-ISMSVM 모형도 전통적인 모형보다 예측의 성과가 우수한 것으로 나타났다. 그리고, 예측정확도 비교측면에서도 GA-ISMSVM 이 GA-FSMSVM을 약간 앞서는 것으로 나타난 점도 선행연구들의 결과와 궤를 같이한다고 볼 수 있을 것이다. 본 연구의 목적인 제안모형 FIS-MSVM은 IS-MSVM모형보다도 1.67%P높은 결과를 보여주었다. 이러한 본 연구결과와 선행연구들을 입증하는 결과를 종합해 보았을 때 제안모형 FIS-MSVM은 MSVM의 예측성능을 개선 시킬 수 있는 하나의 좋은 대안이 될 수 있을 것으로 기대된다.

비교모형과 제안모형간의 예측 정확도의 차이가 통계적으로 유의한 지를 검증하기 위해 멕네마르(McNemar) 검정을 수행하였다. 멕네마르 검정은 5개의 전통적인 모형과 GA- MSVM 두 가지, 그리고 제안모형에 적용되었다. <Table 8>은 멕네마르 검정의 수행 결과를 나타낸 것이다. 여기서 FIS-MSVM은 전통적인 비교모형들과는 99% 신뢰수준 하에서 유의한 성과 차이를 나타냄을 알 수 있었다. 다만 다른 비교모형인 FS-MSVM과는 99% 신뢰수준에서 유의한 차이

〈Table 7〉 Hit ratio of proposed model and comparative model

Model		Option	Hit Ratio		
			Train	Test	Valid
MDA		Stepwise Selection	85.09%		79.85%
CBR		k=6*	86.17%		84.66%
CART		Gini, Max tree depth=5, Maximum risk gap of standard error for pruning = 1	86.93%		82.15%
ANN		h=18**	90.51%	87.05%	87.79%
MSVM		RBF, C=10, σ^2=25	92.79%		88.52%
GA-MSVM	FS-MSVM	RBF, C=64.00, σ^2=19.88, FS=12***	94.45%		89.04%
	IS-MSVM	RBF, C=73.65, σ^2=3.18, IS=1115****	100.0%		90.50%
	FIS-MSVM	RBF, C=10, σ^2=1, FS=9, IS=1114	100.0%		92.17%

* k : the number of the nearest neighbors
** h : the number of the nodes in a hidden layer
*** FS : the number of the selected features
**** IS : the number of the selected instances

〈Table 8〉 Results from McNemar Test

	CBR	CART	ANN	MSVM	FS-MSVM	IS-MSVM	FIS-MSVM
MDA	10.55***	3.68*	53.07***	45.13***	45.60***	60.01***	**72.81*** **
CBR		3.73*	7.93***	10.71***	16.16***	27.01***	**42.01*** **
CART			29.88***	28.35***	30.18***	41.61***	**57.85*** **
ANN				0.46	1.38	6.01**	**17.51*** **
MSVM					0.28	3.9**	**11.45*** **
FS-MSVM						2.01	**9.14*** **
IS-MSVM							**2.68***

* statistical significant at 10%, ** statistical significant at 5%, *** statistical significant at 1%

를 보여주었지만, IS-MSVM모형과는 90% 수준에서 유의한 차이를 보여주었다.

5. 결론

본 연구에서는 유전자 알고리즘(GA)을 래퍼(wrapper)로 사용하여 커널함수 매개변수와 입력변수 선택, 그리고 학습사례 선택 등 3 가지를 동시에 최적화 하는 새로운 MSVM 모형(GA-MSVM)을 제안하고, 이를 주가지수 추세예측 분야에 적용하여 기존의 연구에서 제시한 비교모형들보다 예측성과가 우수하게 개선될 수 있다는 것을 확인하였다. 특히 제안모형의 경우, 전통적으로 연구되어 왔던 MDA, MLOGIT, CBR, ANN과 같은 알고리즘과 많은 기존 연구에서 성과가 우수한 기법으로 소개되어 왔던 Multiclass SVM과 비교하여도 비교적 우수한 성능을 나타내고 있어 향후 주식시장 또는 자본시장에서 가격이나 리스크를 예측하고자 하는 실무자들과 학계에 유용하게 사용될 수 있을 것으로 기대된다.

다음으로 본 연구가 갖는 성과와 의의를 다음과 같이 두 가지로 살펴 볼 수 있을 것이다. 첫째, 학술적인 의의 측면에서 본 연구는 이전에 시도된 적이 거의 없는 커널함수 매개변수, 입력변수 선택, 학습사례 선택을 동시에 GA로 최적화 하는 새로운 GA-MSVM 모형을 제안하였다는 점이다. 이 제안모형은 비록 IS-MSVM모형과 90% 유의수준에서 성과 의 차이를 보이긴 했지만 3가지 동시 최적화가 성능개선에 효과있음을 보여 준 것이 하나의 성과라 할 수 있을 것이다.

두 번째 의의는 이러한 다분류 모형의 응용분야를 주식시장이라는 주류 자본시장의 데이터를 적용하여 MSVM의 응용분야를 한 걸음 더 넓힐 수 있다는 가능성을 보여준 측면이다. 그 동안 다분류 모형의 연구 분야는 주로 응용과학이나 공학분야에서 주로 사용되어 왔고 최근에서야 경제학 분야로 그 범위를 넓히고 있는 실정이다. 이전 Ahn(2014)의 연구에서 기업신용평가 분야에서 사용된 경우가 있지만 이후 더 이상의 후속 연구는 거의 없는 실정이었다. 이러한 점을 감안할 때 본 연구는 향후 주식시장의 각종 지표들,

예를 들면 주식가격, 위험, 변동성 등 가치가 있는 지표 등을 예측할 수 있는 하나의 방법론으로서 사용될 가능성을 높여 실무적으로 사용범위가 한층 넓어지는 것을 기대할 수 있을 것이다.

그러나 본 연구는 이러한 학술적, 실무적 의의에도 불구하고, 다음과 같은 몇 가지 한계점 또한 가지고 있다. 첫째, 학습 데이터 선택기간에 대해서 추가적인 보완점이 필요하다는 부분이다. 우리나라 주식시장의 경우 2008년 이후와 이전의 패턴이 많은 부분 달라졌다는 것이 현재 실무 전문가들의 공통된 의견임을 감안할 때 이에 대한 고려도 충분히 반영할 필요가 있을 것이다. 둘째, 모형의 최적화 요인으로 제안 모형은 비록 이전 연구에 없었던 3가지 요인을 동시에 최적화 하긴 하였지만 커널함수를 RBF로 하나로 고정해서 사용했다는 점과 MSVM을 구현하는 여러 가지 방법 중에서 OAO기법만을 사용했다는 점도 향후 추가연구가 필요하다고 생각되는 점들이다. 셋째, 주가지수 추세를 결정하는 변수들을 보다 다양하게 추출하여 테스트를 해볼 필요가 있을 것이다. 본 연구에서 사용된 설명변수들은 비가격지표들이 배제되어 있는 데이터를 사용하였는데, 향후 연구에서는 비가격 지표 또는 파생상품시장 지표 등 다양한 변수들을 도입하여 모형을 구축하는 노력이 필요할 것으로 보인다. 넷째, 기계학습의 특성상 앙상블 모델을 적용할 경우 보다 성능 좋은 결과를 얻을 경우가 많다. 본 연구에서 성과를 보인 다분류 알고리즘들을 딥러닝 모델들과 결합하여 적용하는 방안을 고려해 보는 것도 후속연구에서 보완되어야 할 점으로 생각된다. 마지막으로 모형의 성능에 대한 검증뿐 아니라 실제 예측된 결과의 데이터로 수익률을 개선 할 수 있는지에 대한 연구가 실무적으로 보다 중요할 것으로 보인다.

참고문헌(References)

Ahn, H., Kim, K.-j., and Han, I., "Intelligent Credit Rating Model for Korean Companies using Multiclass Support Vector Machines," Korean Management Review, Vol. 35, No. 5 (2006), pp. 1479-1496.

Ahn, H., and Kim, K.-j. "Corporate Bond Rating Using Various Multiclass Support Vector Machines", Asia Pacific Journal of Information Systems, Vol.19, No.2(2009), pp. 157-178.

Ahn, H., Lee, K. and Kim, K.-j. "Global Optimization of Support Vector Machines Using Genetic Algorithms for Bankruptcy Prediction," Lecture Notes in Computer Science, Vol. 4234 (2006), pp. 420-429.

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

Chatterjee, S. "Vision-based rock-type classification of limestone using multi-class support vector machine," Neurocomputing, Vol.39, No.1 (2013), pp. 14-27.

Chen, L.H., and Hsiao, H.D. "Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study," Expert Systems with Applications, Vol.35, No.3 (2008), pp. 1145-1155.

Crammer, K. and Singer, Y. "On the Learnability and Design of Output Codes for Multiclass Problems," Proceedings of the 13th Annual Conference on Computational Learning Theory, Palo Alto, California, USA (2000), pp. 35-46.

Dash, R., and Dash, P. K. “A hybrid stock trading framework integrating technical analysis with machine learning techniques,” The Journal of Finance and Data Science, Vol.2 (2016) 42-57.

Hong, T., and Park, J. “Feature Selection for Multi-Class Support Vector Machines Using an Impurity Measure of Classification Trees:An Application to the Credit Rating of S&P 500 Companies,” Asia Pacific Journal of Information Systems, Vol. 21, No. 2(2011), pp. 43-58.

Howley, T., and Madden, M.G. “The Genetic Kernel Support Vector Machine: Description and Evaluation,” Artificial Intelligence Review, Vol. 24, Nos. 3-4 (2005), pp. 379-395.

Hsu, C.W., and Lin, C.J. “A Comparison of Methods for Multiclass Support Vector Machines,” IEEE Transactions on Neural Networks, Vol. 13, No. 2 (2002), pp. 415-425.

Jack, L.B. and Nandi, A.K., “Fault Detection Using Support Vector Machines and Artificial Neural Networks, Augmented by Genetic Algorithms,” Mechanical Systems and Signal Processing, Vol.16 (2002), pp. 373-390.

Kim, K.-j., “Artificial neural networks with evolutionary instance selection for financial forecasting,” Expert Systems with Applications, Vol.30, No.3 (2006), pp. 519-526.

Kim, K.-j. and Ahn, H. “Optimization of Support Vector Machines for Financial Forecasting,” Journal of Intelligence and Information Systems, Vol.17, No.4 (2011), pp. 223-236.

Kim, S.W., “Comparison of Predictability of Stock Price Volatility: Focusing on Price Range and VKOSPI,” Journal of Korean Data Analysis Society, Vol.13, No.2 (2011), pp. 915-925.

Kim, S. W. and H. C. Ahn, “Development of an Intelligent Trading System using Support Vector Machines and Genetic Algorithms,” Journal of Intelligence and Information Systems, Vol.16, No.1(2010), 71~92.

Lee, H., “A Combination Model of Multiple Artificial Intelligence Techniques Based on Genetic Algorithms for the Prediction of Korean Stock Price Index(KOSPI),” Entrue Journal of Information Technology, Vol.7, No.2 (2008), pp. 33-43.

Lee, K., and Byun, H., “A New Face Authentication System for Memory-Constrained Devices,” IEEE Transactions on Consumer Electronics, Vol.49, No.4 (2003), pp. 1214-1222.

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., Clark, R.A., “Data mining techniques for cancer detection using serum proteomic profiling,” Artificial Intelligence in Medicine, Vol.32, No.2 (2004), pp. 71-83.

Lorena, A.C., and de Carvalho, A.C.P.L,F. “Comparing Techniques for Multiclass Classification Using Binary SVM Predictors,” Lecture Notes in Artificial Intelligence, Vol.2972 (2004), pp. 272-281.

Lorena, A.C., and de Carvalho, A.C.P.L,F. “Evolutionary tuning of SVM parameter values in multiclass problems,” Neurocomputing, Vol.71, Nos.16-18 (2008), pp. 3326-3334.

Pai, P.-F., and Hong, W.-C. “Forecasting regional electricity load based on recurrent support

vector machines with genetic algorithms," Electric Power Systems Research, Vol.74, No.3 (2005), pp. 417-425.

Ra, Y.S., H. S. Choi, and S.W. Kim, " VKOSPI Forecasting and Option Trading Application Using SVM," Journal of Intelligence and Information Systems, Vol.22, No.4(2016), 177~192.

Reeves, C.R., Taylor, S.J., "Selection of training sets for neural networks by a genetic algorithm. In Eiden," A.E., Back, T., Schoenauer M., Schwefel, H.-P., "Parallel problem-solving from nature-PPSN V.," Springer. Berlin (1998)

Shieh, M.-D., and Yang, C.-C. "Multiclass SVM-RFE for product from feature selection," Expert Systems with Applications, Vol.35, Nos.1-2 (2008), pp. 531-541.

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.

Sun, Z., Bebis, G., Miller, R., "Object detection using feature subset selection," Pattern Recognition, Vol.37, No.11 (2004), pp. 2165-2176.

Thi N., Lee G.-B., Peter W., and Jim P., "GA-SVM Based Framework for Time Series Forecasting," Proceedings of the Fifth International Conference on Natural Computation (2009).

Vapnik, V. The Nature of Statistical Learning Theory. New York, NY: Springer-Verlag, 1995.

Wu, C.H., Tzeng, G.H., Goo, Y.J., and Fang, W.C. "A real-valued genetic algorithm to optimize the parameters of support vector machine for prediction bankruptcy," Expert Systems with Applications, Vol. 32, No. 2 (2007), pp. 397-408.

Zhao, X.-M., Cheung, Y.-M., Huang, D.-S., "A novel approach to extracting features from motif content and protein composition for protein sequence classification," Neural Networks, Vol.18, No.8 (2005), pp. 1019-1028.

Abstract

A Study on the Prediction Model of Stock Price Index Trend based on GA-MSVM that Simultaneously Optimizes Feature and Instance Selection

Jong-sik Lee* · Hyunchul Ahn**

There have been many studies on accurate stock market forecasting in academia for a long time, and now there are also various forecasting models using various techniques. Recently, many attempts have been made to predict the stock index using various machine learning methods including Deep Learning. Although the fundamental analysis and the technical analysis method are used for the analysis of the traditional stock investment transaction, the technical analysis method is more useful for the application of the short-term transaction prediction or statistical and mathematical techniques. Most of the studies that have been conducted using these technical indicators have studied the model of predicting stock prices by binary classification - rising or falling - of stock market fluctuations in the future market (usually next trading day). However, it is also true that this binary classification has many unfavorable aspects in predicting trends, identifying trading signals, or signaling portfolio rebalancing.

In this study, we try to predict the stock index by expanding the stock index trend (upward trend, boxed, downward trend) to the multiple classification system in the existing binary index method. In order to solve this multi-classification problem, a technique such as Multinomial Logistic Regression Analysis (MLOGIT), Multiple Discriminant Analysis (MDA) or Artificial Neural Networks (ANN) we propose an optimization model using Genetic Algorithm as a wrapper for improving the performance of this model using Multi-classification Support Vector Machines (MSVM), which has proved to be superior in prediction performance. In particular, the proposed model named GA-MSVM is designed to maximize model performance by optimizing not only the kernel function parameters of MSVM, but also the optimal selection of input variables (feature selection) as well as instance selection. In order to verify the performance of the proposed model, we applied the proposed method to the real data. The results show

* Hanwha Asset Management
** Corresponding Author: Hyunchul Ahn
Associate Professor, Graduate School of Business IT Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Republic of Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

that the proposed method is more effective than the conventional multivariate SVM, which has been known to show the best prediction performance up to now, as well as existing artificial intelligence / data mining techniques such as MDA, MLOGIT, CBR, and it is confirmed that the prediction performance is better than this. Especially, it has been confirmed that the 'instance selection' plays a very important role in predicting the stock index trend, and it is confirmed that the improvement effect of the model is more important than other factors.

To verify the usefulness of GA-MSVM, we applied it to Korea's real KOSPI200 stock index trend forecast. Our research is primarily aimed at predicting trend segments to capture signal acquisition or short-term trend transition points. The experimental data set includes technical indicators such as the price and volatility index (2004 ~ 2017) and macroeconomic data (interest rate, exchange rate, S&P 500, etc.) of KOSPI200 stock index in Korea. Using a variety of statistical methods including one-way ANOVA and stepwise MDA, 15 indicators were selected as candidate independent variables. The dependent variable, trend classification, was classified into three states: 1 (upward trend), 0 (boxed), and -1 (downward trend). 70% of the total data for each class was used for training and the remaining 30% was used for verifying. To verify the performance of the proposed model, several comparative model experiments such as MDA, MLOGIT, CBR, ANN and MSVM were conducted. MSVM has adopted the One-Against-One (OAO) approach, which is known as the most accurate approach among the various MSVM approaches. Although there are some limitations, the final experimental results demonstrate that the proposed model, GA-MSVM, performs at a significantly higher level than all comparative models.

Key Words : Multiclass SVM, Genetic Algorithm, Feature Selection, Instance Selection, Stock Market Index Trend Prediction

Received : October 31, 2017 Revised : November 19, 2017 Accepted : December 2, 2017
Publication Type : Regular Paper Corresponding Author : Hyunchul Ahn

저 자 소 개

이 종 식

현재 한화자산운용㈜ 정보시스템팀장으로 재직 중이다. 국민대학교 정보관리학과를 졸업하고, 수원대학교 금융대학공학원에서 금융공학석사 학위를 취득하였다. 국민대학교 비즈니스IT전문대학원 박사과정을 수료하였고, 주요 관심분야는 금융 및 머신러닝 응용과 관련된 데이터분석이다.

안 현 철

현재 국민대학교 비즈니스IT전문대학원 부교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심 분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.