지능정보연구 제17권 제4호 2011년 12월(pp.241~254)

재무예측을 위한 Support Vector Machine의 최적화

김경재
동국대학교_서울 경영정보학과
(kjkim@dongguk.edu)

안현철
국민대학교 경영정보학부
(hcahn@kookmin.ac.kr)

Support vector machines(SVM)은 비교적 최근에 등장한 데이터마이닝 기법이지만, 재무, CRM 등의 경영학 분야에서 많이 연구되고 있다. SVM은 인공신경망과 필적할 만큼의 예측 정확도를 보이는 사례가 많았지만, 암상자로 불리는 인공신경망 모형에 비해 구축된 예측모형의 구조를 이해하기 쉽고, 인공신경망에 비해 과도적합의 가능성이 적어서 적은 수의 데이터에서도 적용 가능하다는 장점을 가지고 있다. 하지만, 일반적인 SVM을 이용하려면, 인공신경망과 마찬가지로 여러 가지 설계요소들을 설계자가 선택하여야 하기 때문에 임의성이 높고, 국부 최적해에 수렴할 가능성도 크다. 또한, 많은 수의 데이터가 존재하는 경우에는 데이터를 분석하고 이용하는데 시간이 소요되고, 종종 잡음이 심한 데이터가 포함된 경우에는 기대하는 수준의 예측성과를 얻지 못할 가능성이 있다. 본 연구에서는 일반적인 SVM의 장점을 그대로 유지하면서, 전술한 두 가지 단점을 보완한 새로운 SVM 모형을 제안한다. 본 연구에서 제안하는 모형은 사례선택기법을 일반적인 SVM에 융합한 것으로 대용량의 데이터에서 예측에 불필요한 데이터를 선별적으로 제거하여 예측의 정확도와 속도를 제고할 수 있는 방법이다. 본 연구에서는 잡음이 많고 예측이 어려운 것으로 알려진 재무 데이터를 활용하여 제안 모형의 유용성을 확인하였다.

논문접수일 : 2011년 11월 18일 게재확정일 : 2011년 12월 19일
투고유형 : 학술대회우수논문 교신저자 : 안현철

1. 서론

주식시장예측은 많은 연구자들에 의해서 오랫동안 연구되어 온 주제이다. 그러나 분석에 고려하여야 할 데이터의 양도 매우 많고 잡음도 많은 특성을 가지고 있기에 예측이 매우 어려운 것으로 알려져 있다. 데이터마이닝은 원래 잡음이 많은 대용량의 데이터를 분석하기 위해 연구되어 오던 분야이다. 이를 위하여 컴퓨터공학이나 통계학에서 활용되어 온 많은 분석방법론을 이용하고 있으며, 최근에는 Support vector machines(SVM)과 같은 기법이 많은 연구자의 관심을 받고 있다.

SVM은 인공신경망과 필적할 만한 예측력을 가지고 있는 것으로 알려져 있지만, 인공신경망에 비해 도출된 모형에 대한 설명이 가능하고, 비교적 적은 데이터를 이용하는 경우에도 과도적합의 가능성이 적은 장점을 가지고 있다. 그러나 SVM과 같은 방법도 많은 양과 잡음이 많은 주식시장 데

* 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2009-327-B00212).

이터를 분석할 때에는 인공신경망과 같은 전통적인 방법과 마찬가지로 분석의 한계점을 보이게 된다. 따라서 데이터마이닝의 관점에서 볼 때 이러한 예측기법을 적용하기 이전에 분석할 데이터의 양을 적절하게 줄이거나 잡음을 제거하기 위한 노력을 할 필요가 있다.

전통적인 데이터마이닝에서는 이를 위하여 전처리과정을 거치게 된다. 즉, 분석에 사용할 데이터를 축약하기 위하여 전처리과정을 거치게 되는데, 이는 흔히 결측치, 이상치 제거와 정규화 등의 과정을 의미한다. 그러나 전통적인 전처리과정은 데이터 자체의 특성을 분석하여 분석을 보다 용이하게 할 수 있는 데이터를 추출하기 위하여 이상치나 결측치, 특정 범위를 벗어난 값을 가진 데이터 등을 제거하는 방법이다. 그러나 이러한 전처리과정은 제거되는 데이터를 포함할 때와 제거할 때의 예측정확도의 차이를 고려하지 않으므로, 예측정확도의 제고라는 관점에서는 얼마나 적합하게 데이터를 제거해 주는 지 판단할 수가 없다. 따라서 최근의 데이터마이닝 연구에서는 이러한 전처리과정도 예측정확도를 높일 수 있도록 예측정확도와 연동된 방법을 사용하는데, 이를 흔히 Wrapper model이라고 한다. Wrapper model은 예측 등의 데이터 분석에 있어서, 예측정확도를 고려하여 분석을 하는 방법을 의미한다. 반면에 예측정확도를 고려하지 않고 데이터 자체의 특성만을 이용하여 분석하는 방법을 Filter model이라고 한다(Liu and Motoda, 1998).

본 연구에서는 예측정확도 제고라는 목적을 가지고 있으므로 Wrapper model을 이용한 전처리를 통하여 양이 많고 잡음이 심한 주식시장 데이터에서의 예측을 위한 새로운 방법을 제안한다. 제안하는 방법은 데이터마이닝을 이용한 분석과정에서 예측정확도를 보다 제고할 수 있도록 분석에 사용될 데이터를 선택적으로 활용하는 방법이다. 데이터마이닝 연구에서는 이러한 방법을 “사례선택기법(instance selection technique)”이라고 한다. 본 연구에는 전통적인 전처리과정은 물론이고, 새롭게 제안하는 wrapper model에 의한 사례선택기법을 활용하여 전통적인 데이터마이닝 기법과 융합한 모형을 제안한다. 기존의 여러 연구들에서는 인공신경망, 사례기반선택 등에 사례선택기법이 결합되어 사용된 적이 있으나, 본 연구에서는 SVM에 사례선택기법을 결합한 새로운 분석모형을 제안한다.

본 연구는 아래와 같이 구성된다. 제 2장에서는 SVM에 대한 기본적인 이론 설명과 사례선택기법에 대해 설명한다. 제 3장에서는 연구 데이터에 대한 설명과 본 연구에서 제안하는 SVM과 사례선택기법이 결합된 새로운 분석모형을 설명한다. 제 제 4장에서는 실험결과를 정리하고, 제 5장에서는 연구의 결론과 함께, 연구의 한계점과 미래 연구방향에 대해 논의할 것이다.

2. 선행연구

SVM은 Vapnik에 의해 개발된 분류 또는 예측기법인데, 최근 몇 년간 인공신경망에 필적할 만큼 많은 SVM 응용연구가 진행되어 왔다. 초기 연구에서는 SVM을 문서분류, 영상인식, 문자인식 등에 적용하여 그 일반화 성능을 검증한 연구들이 많았다. 최근에는 경영학 분야에 SVM을 응용하고자 하는 연구들이 많이 진행되고 있는데 특히, 재무분야에 적용한 연구로는 주로 시계열 예측 및 분류에 관한 것이었다(Tay and Cao, 2001, Kim, 2003, 안현철 등, 2006, 안현철, 김경재, 2009). 이 연구들에서는 전형적인 SVM 모형의 재무예측에의 유용성을 평가하거나, 채권평가의 분야에서 일

반적인 이분류 SVM을 대체하는 다분류 SVM의 여러 모형들의 예측가능성을 검증한 것이었다. 그러나 선행연구에서는 대부분 일반적인 SVM 알고리즘에 대한 수정 없이 응용분야에 맞는 SVM의 모수 결정이나 기존 데이터마이닝 기법과의 성능 비교에 치중하였다. 아래에서는 일반적인 이분류 SVM의 기본적인 원리를 설명하고, 본 연구에서 사용하는 사례선택기법에 대해 설명한다.

2.1 SVM

SVM은 입력공간과 관련된 비선형문제를 고차원의 특징공간에서의 선형문제로 변환시켜 나타내기 때문에 수학적으로 분석하는 것이 수월하다는 점이 특징이다(Vapnik, 1995, 1998). 다른 데이터마이닝 기법에 비하여 SVM은 조정해야 할 모수의 수가 많지 않아 비교적 쉽게 학습에 영향을 미치는 요인들을 규명할 수 있고, 구조적위험을 최소화하여 과도적합문제에서 벗어날 가능성이 크며, 볼록함수를 최소화하는 학습을 진행하여 전역최적해를 찾을 가능성이 크다는 점이 장점이다.

일반적으로 이분류 SVM은 두 개의 분류집단을 갖는 종속변수를 가진 학습용 데이터의 입력벡터 x를 고차원의 특징공간(high-dimensional feature space)으로 매핑시킨 후, 두 분류집단 사이의 마진을 최대화시키는 분리 경계면를 찾는 것을 목표로 한다. 이러한 경계면을 최대폭 분리 경계면(maximum margin hyperplane)이라고 하고, 이는 두 분류집단 사이의 거리를 최대로 분리시키는 경계면의 역할을 한다. 이 때 최대마진 분리 경계면에 근접해 있는 데이터들을 서포트 벡터(support vector)라고 부른다(안현철 등, 2006).

이상의 과정을 간단한 수식으로 예를 들어 설명하면 다음과 같다. 일반적인 선형분리문제에서 3개의 독립변수를 가진 학습 데이터에 대한 분리 경계면은 식 (1)과 같이 표현할 수 있다.

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \tag{1}$$

일반적으로 상기의 수식에서 y는 출력값을, x_i는 변수값을, 4개의 w_i는 데이터마이닝 기법에 의해 학습된 가중치를 의미한다. 식 (1)에서는 가중치 w_i가 SVM에 의해 만들어 지는 분리 경계면을 결정하는 모수이다. 이 때 최대마진 분리 경계면은 서포트 벡터를 사용해서 식 (2)와 같이 나타낼 수 있다.

$$y = b + \sum \alpha_i y_i x(i) \cdot x \tag{2}$$

식 (2)에서, y_i는 학습용 데이터 $x(i)$의 분류값을, $\cdot$는 내적(dot product)을, 벡터 x는 검증용 데이터를, 벡터 $x(i)$는 서포트 벡터를 의미한다. 한편, 식 (2)에서, b와 α_i는 분리 경계면을 결정하는 모수이다. 서포트 벡터를 파악하고, 모수 b와 α_i를 결정하는 것은 선형적으로 제약된 이차계획문제(linearly constrained quadratic programming)를 해결하는 것과 같다(Chang and Lin, 2001).

SVM은 저차원의 입력변수들을 고차원의 특징공간으로 매핑시킴으로써 비선형 분류문제를 선형 분류기로 분류할 수 있도록 하는 방법이다. 아래 식 (3)은 비선형 분류문제에서 사용될 식 (2)의 고차원 형태를 표현한다.

$$y = b + \sum \alpha_i y_i K(x(i), x) \tag{3}$$

이 식에서 함수 $K(x(i), x)$는 커널함수(kernel function)이며, 이를 통하여 원래의 저차원 데이터

를 고차원 공간으로 매핑시킴으로써 특징공간 내에 선형적으로 분리가 가능한 입력 데이터군을 만들 수 있다. 일반적인 SVM에 사용되는 커널함수는 여러 가지가 있으며, 어떤 커널함수를 선택하는 것이 적합한가는 해결하여야 하는 문제에 따라 상이하다. 따라서 커널함수의 선택은 SVM을 적용하는 데 있어서 매우 중요한 요소 중의 하나라고 할 수 있다(안현철, 김경재, 2009). 선행연구에서 일반적으로 많이 사용된 커널함수는 선형함수(linear function)와 다항식 함수(polynomial function), 그리고 가우시안 RBF 함수(Gaussian radial basis function) 등이 있으며, 식 (4)~식 (6)은 각각을 수식으로 표현한 것이다.

선형 함수 : $K(x, y) = xy$ (4)

다항식 함수 : $K(x, y) = (xy+1)^d$ (5)

가우시안 RBF 함수 :

$$K(x, y) = \exp(-\frac{1}{\sigma^2}(x-y)^2) \quad (6)$$

상기 식들에서 d는 다항식 함수의 차수를, σ^2은 가우시안 RBF 함수의 대역폭을 의미한다. 대부분의 고차원 문제에서는 다항식 함수나 가우시안 RBF 함수를 이용한다.

2.2 사례선택기법

데이터마이닝은 대용량 자료에 내포되어 있는 유용한 정보나 지식을 찾기 위한 분석 과정이라고 할 수 있다. 그러나 아무리 뛰어난 데이터마이닝 기법이라고 하더라도 대용량의 잡음이 심한 데이터를 분석하는 일은 매우 어려운 일이 될 수 있다. 실생활의 데이터는 대부분 대용량인 경우가 많으며, 특히 재무 데이터의 경우에는 대용량의 잡음이 심한 고차원적 특성을 가진 경우가 많다. 따라서 성공적인 데이터마이닝 성과를 얻기 위해서는 대용량 원데이터를 분석이 용이하도록 적절하게 축약하여 사용하면 성과가 좋아지는 경우가 있으며, 이런 과정을 데이터마이닝에서는 “특징선택(feature selection)”과 “사례선택(instance selection)” 또는 “자료편집(data editing)” 등의 용어로 표현한다. 특징선택은 전통적인 데이터 분석과정에서의 독립변수 선정과정과 유사한 절차로 설명할 수 있다. 대부분의 연구에서 기본적으로 분석에 유용한 독립변수를 선정하기 위하여 단변량 또는 다변량 분석기법을 사용하여 특징을 추출하는 과정을 거친다. 특징선택은 이미 많은 연구에서 그 유용성이 확인되었으므로 본 연구에서는 이에 대해 더 이상 언급하지 않는다. 다음으로 사례선택기법 또는 자료편집기법은 분석에 사용할 모형 구축용 데이터에서 일부 데이터들만을 추출하여 분석에 사용하는 기법을 의미한다. 때로는 여러 데이터의 특성을 축약하여 적은 수의 데이터 집합을 만들어 사용하기도 한다.

그런데, 데이터의 축약과정은 정보손실이 발생할 가능성이 높으며, 이러한 경우에는 분석의 효율성은 높아질 수 있으나. 분석의 성과는 효과적이지 않을 수 있다. 따라서 정보의 손실을 최소화하면서 전체 데이터의 특성을 유지할 수 있는 대표성 있는 데이터를 선택하는 것이 매우 중요하다. 대표성 있는 데이터가 추출 되었는지는 원데이터를 활용하여 나온 성과와 사례선택기법을 통해 축약된 데이터를 활용하여 나온 분석결과를 비교하여 효율성과 효과성을 검토함으로써 확인이 가능하다.

많은 선행연구들이 여러 가지 사례선택기법들을 제안하였다. 특히, 데이터마이닝 기법 중에서 사례기반추론과 같은 사례기반의 기법들은 과다한 데이터 저장용량과 분석시간을 줄이고, 데이터 내의 잡음과 학습과정의 과도적합 가능성을 줄이

기 위해 사례를 선택해야만 하는 상황이 종종 발생하기에 사례선택기법에 대한 연구가 많이 이루어졌다(Wilson and Martinez, 2000).

Kuncheva(1993)는 '압축된 근접이웃규칙(condensed nearest neighbor rule)', '생성된 또는 수정된 프로토타입(generated or modified prototypes)', '이수준 분류기(two-level classifier)' 등의 세 가지 종류로 사례선택방법을 분류하였는데, 압축된 근접이웃규칙은 전체 데이터와 유사한 하위 데이터 집합을 생성하는 기법이며 이 기법을 활용한 연구로는 Hart(1968), Gates(1972), Wilson(1972), Ritter et al.(1975) 등이 있다. 생성된 또는 수정된 프포토타입 방법은 프로토타입이라 명명된 특징화된 데이터 집합을 새로 생성하거나 원 데이터 중에서 일부 데이터만을 수정하는 방법이다. 이수준 분류기라는 방법은 두 개 이상의 분류기에 데이터를 적절하게 분할하여 투입하고 이에 따라 분석을 수행하는 방법이다. 이수준 분류기를 활용한 연구는 Tetko and Villa(1997)의 연구가 있다.

사례선택기법은 일반적인 데이터마이닝 기법에서 유용하게 사용되지만, 특히 사례기반의 데이터마이닝 기법에서는 지식정제과정의 하나로 생각할 수 있다. 즉, 사례기반의 데이터마이닝 기법에서는 특정 사례에 기반하여 의사결정을 하게 되므로 다른 데이터마이닝 기법보다 사례선택의 효과가 직접적이고 그 정도도 크다. 선행 연구자들 중 일부는 이러한 관점에 주목하여 사례기반추론과 같은 사례기반 데이터마이닝 기법에서 사례선택기법을 활용하여 왔다. 대표적인 연구로는 Smyth(1998)가 사례기반추론의 사례 베이스에서 중복되거나 불필요하거나 분석결과에 해를 미칠 수 있는 사례를 삭제하여 남은 사례를 선택적으로 사용하는 방안에 대해 소개하였으며, McSherry(2000)가 사례기반추론에서 사례 베이스 내의 사례를 평가하고 선택하는 알고리즘을 개발하였다. 이러한 사례기반추론에 관한 연구들 이외에도 인공신경망과 같은 데이터마이닝 기법에서 사례선택에 대한 연구들이 진행되어 왔는데, Reeves and Taylor(1998)는 분류 문제에 신경망을 활용함에 있어서 유전자 알고리즘을 이용하여 학습용 데이터 집합을 선택적으로 이용하는 연구를 수행하였고, Reeves and Bush(2001)는 분류와 예측 문제에서 유전자 알고리즘을 이용하여 RBF 신경망에서의 학습용 데이터 집합을 선택하는 방안을 제시하였다. 한편, Kim(2006)은 유전자 알고리즘과 인공신경망의 결합 알고리즘을 개발하여 신경망에 사용될 학습용 데이터 집합을 유전자 알고리즘을 이용하여 유기적으로 선택하는 방법을 제시하였다.

Kim(2006)에서와 같이 유전자 알고리즘은 예측이나 분류문제 해결의 주과정인 데이터마이닝 기법의 목적함수를 최적화할 수 있도록 설계할 수 있으므로 데이터마이닝 기법과 유기적으로 결합된 과정을 통해 분류나 예측에 유용한 자료를 선택할 수 있다. 이는 데이터마이닝 과정과 사례선택의 과정이 동일한 목적함수(즉, 예측 또는 분류정확도 최대화)를 최적화하도록 설계함으로써 분류나 예측의 성과를 향상시킬 가능성이 높다.

본 연구에서는 사례기반추론이나 인공신경망에서의 사례선택에 성공적으로 이용된 유전자 알고리즘 기반의 사례선택기법을 활용하여 예측성능이 매우 우수한 것으로 알려져 최근 많은 연구에서 이용되고 있는 SVM에 적용하고자 한다. 이는 Reeves and Taylor(1998), Reeves and Bush(2001), Kim(2006)의 연구에서 제안된 방법을 SVM에 응용하는 것이라 할 수 있다. SVM에 관련된 선행연구는 대부분 SVM의 최적 파라미터 결정, 입력변수 최적화 등에 관한 연구만 진행되었을 뿐, 아직 SVM의 사례선택 최적화에 관한 연구는 진행된

바 없으므로, 본 연구가 SVM 응용연구에 기여하는 공헌도가 매우 높다는 의의가 있다.

3. 연구자료와 연구설계

제안하는 SVM에서의 사례선택 기법의 유용성을 확인하기 위해 예측이 매우 어려운 것으로 알려진 주식시장 예측에 관련된 데이터에 적용해 보기로 한다. 주식시장 예측과 관련된 데이터는 일반적으로 잡음과 이상치가 매우 많이 포함되어 있기에 예측능력이 매우 우수한 분석기법을 사용하지 않고서는 좋은 예측성과를 기대할 수 없는 것으로 알려져 있다. 따라서 본 연구에서 제안하는 모형의 성능을 평가하는 데에 유용한 데이터로 판단된다. 본 연구에서 사용하는 데이터는 Ahn and Kim (2008)에서 데이터마이닝 기법의 성능을 확인하기 위해 사용한 데이터이며, 1989년부터 1998년까지의 한국종합주가지수의 일별 종가자료이다. 실험에 사용된 총 표본의 크기는 2,218개이다.

본 연구에서 제안하는 모형은 SVM과 유전자 알고리즘(genetic algorithm, GA)을 이용한 최적화 과정을 통하여 구축되게 되므로 최적화에 따른 과도적합문제를 최소화하기 위해 모형구축을 위한 학습용 데이터셋, 최적화용 데이터셋, 검증용 데이터셋의 세 가지 데이터 집합으로 구성한다. 일반적으로 학습용 데이터셋은 모형구축을 위한 파라미터 추정에 사용되고, 최적화용 데이터셋은 최적화 모형에서의 과도적합문제를 통제하기 위한 목적으로 사용된다. 그리고 구축된 모형의 일반화 정도를 측정하기 위해 모형 구축과정에는 활용되지 않는 데이터 집합이 검증용 데이터셋이다. 본 연구에서는 모형에 따라 최적화 과정이 필요한 경우에는 최적화용 데이터셋을 따로 두고, 최적화를 필요로 하지 않는 경우에는 따로 두지 않았다. 각 데이터셋 별 표본의 숫자는 학습용 1,056개, 최적화용 581개, 검증용 581개로 구성되며, 본 연구에서 제안하는 모형에서는 초기에 1,056개의 학습용 데이터를 모두 이용하여 사례선택을 한 후 이 중 선택된 데이터만을 이용하여 SVM을 구축하게 된다.

본 연구에서 사용된 데이터는 선행연구에서 활용된 데이터로서, 일반적으로 주가지수 예측에 많이 이용되는 기술적 지표들을 입력변수로 선정하였다. 본 연구에서 이용된 기술적 지표는 주가지수 예측 관련 선행연구에서 사용된 지표 중에서 한국 주식시장의 특성을 반영하기 위하여 국내 투자전문기업에서 5인의 투자전문가들을 선정한 후, 이들의 검토를 거친 후 선정하였다. 이러한 방식을 사용한 이유는 기술적 지표가 대부분 과거 주가지수를 토대로 생성되기에 종속변수와의 관련성을 기반으로 통계적인 유의성을 통해 유용한 독립변수를 선정하는 것에 한계가 있기 때문이다. 본 연구에 사용된 독립변수의 목록은 <표 1>과 같다.

본 연구에서 제안하는 유전자 알고리즘에 기반한 사례선택 기법은 일반적인 사례선택 기법을 이용한 선행연구에서와 같이 대용량 데이터를 이용하면서도 정확한 의사결정지원을 할 수 있도록 설계되어야 한다. 이러한 목표는 분석에 사용되는 데이터마이닝 기법의 목적함수와 유전자 알고리즘의 적합도함수를 일치시키고 유전자 알고리즘이 적합도함수를 최적화하기 위해 탐색할 공간을 전체 원데이터로 설정함으로써 구현할 수 있다.

이상의 과정을 보다 구체적으로 살펴 보면 아래와 같다. 유전자 알고리즘을 진행하기 위해서는 탐색공간 내의 여러 변수 집합을 염색체(chromosome)라고 불리는 선형 스트링(string)에 매핑하는 과정이 필요하다. 매핑은 유전자 알고리즘이 탐색공간을 효율적인 탐색할 수 있도록 탐색 목적에 적합한 효과적인 매핑방법을 찾아야 한다.

<표 1> 독립변수 목록(Ahn and Kim(2008)에서 발췌)

변수명	계산식
%K	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$ (LL_t과 HH_t는 각각 최근 t일 동안의 저가 중 최저가와 일일 고가 중 최고가를 의미함)
%D	$\frac{\sum_{i=0}^{n-1} \% K_{t-i}}{n}$
Slow %D	$\frac{\sum_{i=0}^{n-1} \% D_{t-i}}{n}$
Momentum	$C_t - C_{t-4}$
ROC	$\frac{C_t}{C_{t-n}} \times 100$
Williams' %R	$\frac{H_n - C_t}{H_n - L_t} \times 100$
A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$
Disparity5	$\frac{C_t}{MA_5} \times 100$
Disparity10	$\frac{C_t}{MA_{10}} \times 100$
OSCP	$\frac{MA_5 - MA_{10}}{MA_5}$
CCI	$\frac{(M_t - SM_t)}{(0.015 \times D_t)}$ (M_t는 $\frac{(H_t + L_t + C_t)}{3}$, SM_t은 $\frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$, 그리고 D_t는 $\frac{\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert}{n}$를 의미함)
RSI	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} Up_{t-i}/n}{\sum_{i=0}^{n-1} Dw_{t-i}/n}}$ (Up_t는 t시점의 상향 가격 변동성을 Dw_t는 t시점의 하향 가격 변동성을 의미함)

주) 여기서 C_t는 t시점의 종가, L_t는 t시점의 종가, H_t는 t시점의 고가, MA_t는 t일 동안의 평균가격을 의미함.

본 연구에서의 유전자 알고리즘 수행 목적은 주식시장의 미래 방향을 예측하는 데에 유용하며 SVM 내에서 활용할 학습용 자료의 집합을 선택하는 것이다. 예측에 사용할 전체 데이터는 매우 많으며, 이들 데이터를 모두 SVM 분석과정에 사용한다면 소요되는 분석시간도 많이 필요하고, 불필요하거나 중복된 데이터 또는 잡음이 심한 데이터로 인해 효과적이지 못한 분석결과를 제시할 가능성이 높다.

본 연구에서는 전통적인 SVM을 이용하며, SVM의 목적함수를 유전자 알고리즘의 적합도 함수로 설정한 후, 유전자 알고리즘이 탐색할 공간을 전체 데이터 집합으로 정의한다. 본 연구에서 사용하는 SVM의 목적함수는 한국 주식시장의 익일 방향성에 대한 예측 정확성인데, 일반적으로 익일 방향성은 당일의 주가지수의 종가 대비 익일의 주가지수 종가의 상향 또는 하향 여부를 분류하는 것이므로, 실질적으로는 분류정확도(classification accuracy)를 의미한다.

본 연구에서는 SVM에 유전자 알고리즘을 적용하기 위하여 전체 실험자료에 대해 각각 분석에서의 사용여부를 나타내는 코드를 부여하고 이를 조합하여 모집단 염색체를 구성한 다음, 유전자 알고리즘 고유의 연산과정인 교배, 선택, 돌연변이 조작과 이에 따른 산출물인 적합도 평가를 하게 된다. 즉, 전체 원자료 중에서 여러 조합을 이용하여 SVM에서 사용될 실험 데이터를 선택적으로 재구성하고, 이 중에서 적합도 함수의 평가결과가 가장 우수한 자료의 집합을 최종적으로 선택하여 사례선택을 하는 방식이다.

본 연구에서는 SVM과 유전자 알고리즘 실험을 위해 Java platform 기반으로 개발된 유전자 알고리즘과 SVM 결합모형을 이용한다. 본 실험에서 SVM 실험은 SVM 공개 소프트웨어인 LIBSVM (Chang and Lin, 2001)을 사용하고, 유전자 알고리

즘을 SVM 코드에 결합하여 개발된다. 본 연구에서 사용될 유전자 알고리즘의 모수는 50개의 표본과 0.7의 교배비율, 0.1의 돌연변이비율을 기본으로 한다. 유전자 알고리즘 실험의 정지조건은 50세대로 한다.

본 연구에서는 제안하는 연구모형의 유용성을 확인하기 위하여 전통적인 방식의 SVM(본 연구에서는 'SVM'으로 표기), 유전자 알고리즘을 SVM의 모수 최적화에 사용한 SVM(SVM with optimized parameters; PSVM), 데이터마이닝 기법 중 예측 연구에 가장 많이 활용되어 온 오류역전파 인공신경망 모형(ANN)과 사례기반추론 모형 중 하나의 근접이웃만을 고려하는 기본모형(1-NN), 통계적인 기법 중에서 예측 연구에 가장 많이 사용되고 있는 로지스틱 회귀분석(Logit)의 성과를 비교한다.

먼저, 전통적인 방식의 SVM은 별도의 최적화 과정 없이 일정한 범위 내의 모수값을 적용하고 사례선택과정도 수행되지 않는 방식이다. 전통적인 SVM은 Polynomial 함수와 가우시안 RBF 함수를 이용하는 것이 일반적이다. 그러나 Kim(2003)에 의하면, polynomial 함수는 가우시안 RBF 함수에 비해 SVM 구축과정에서 시간이 많이 소요되고, 예측성과도 좋지 않았던 것으로 보고된 바 있다. 실제로 두 가지 함수를 모두 사용하여 여러 모수들을 적용하여 실험하는 일은 매우 많은 실험시간을 요한다. 따라서 본 연구에서는 전통적인 SVM 구축을 위해 가우시안 RBF 함수를 사용한다. 한편, 최적의 모수를 찾는 데에도 많은 실험을 요하는데, 본 연구에서는 Tay and Cao(2001)와 Kim(2003)의 연구에서 사용된 모수 범위 내에서 실험을 진행한다. 선행연구에서 제시된 적절한 모수 범위는 σ^2의 경우는 1에서 100사이의 값으로, C의 경우에는 10에서 100사이의 범위 내로 한정한다. 이러한 모수의 범위는 유전자 알고리즘을 사용하는 모형에도 동일하게 적용된다.

한편, ANN 모형은 가장 일반적인 오류역전파 신경망 모형을 이용하며, 실험도구는 NeuroShell이라는 상용 소프트웨어를 이용한다. 사례기반추론은 전통적인 근접이웃을 고려하는 사례기반추론 모형 중 가장 기본모형이라고 할 수 있는 현재 사례와 가장 가까운 하나의 사례를 이용하여 유클리드 거리에 의한 예측을 수행하는 1-NN을 사용한다. 로지스틱 회귀분석은 SPSS 18.0을 이용한다.

4. 실험결과

본 연구에서는 제안하는 모형의 유용성을 확인하기 위하여 전통적인 SVM(SVM), 모수 최적화 SVM(PSVM), 인공신경망(ANN), 사례기반추론(1-NN), 로지스틱 회귀분석(Logit)의 결과를 비교하였다. <표 2>는 각각의 여섯 모형에서 가장 좋은 예측성능을 나타내었을 때의 성과들을 정리한 것이다.

<표 2>에서 나타난 것과 같이 본 연구에서 제안한 사례선택 기법에 의한 SVM 모형(ISVM)이 비교 대상인 모든 모형들에 비해 예측성능이 가장 우수한 것으로 나타났다. 구체적으로는 1-NN에 비해 검증용 데이터에서 15.32%의 차이로 예측성능 차이가 가장 크게 나타났고, 인공신경망에 비해서는 6.89%, Logit과 전통적인 SVM에 비해서는 5.34%, 모수 최적화 SVM(PSVM)에 대해서는 4.82% 예측성능 개선이 있었다. 특히 ISVM은 전체 학습용 데이터 1,056개 중에서 사례선택 기법에 의해 556개의 데이터만을 선택적으로 사용함으로써 예측에 이용되는 사례의 수가 전통적인 모형들에 비해 거의 절반 수준으로 감소하였으며, 이는 향후 예측 등에 있어서 더 효율적인 분석이 가능

<표 2> 모형 별 예측정확도

모형	학습용	최적화용	검증용	모형 모수
ANN	58.62%	57.49%	56.28%	은닉노드 수 : 18
1-NN			47.85%	
Logit	56.38%		57.83%	전진선택방법
SVM	64.75%		57.83%	C = 78.0, σ^2 = 25.0
PSVM	58.90%	55.77%	58.35%	C = 6.45, σ^2 = 98.36
ISVM	67.63%	59.38%	63.17%	C = 12.47, σ^2 = 21.19

<표 3> 예측정확도 차이에 대한 통계적 유의성 검정 결과(p값)

	1-NN	Logit	SVM	PSVM	ISVM
ANN	0.0020	0.2969	0.2969	0.2383	0.0084
1-NN		0.0003	0.0003	0.0002	0.0000
Logit			0.5000	0.4292	0.0314
SVM				0.4292	0.0314
PSVM					0.0463

함을 의미한다.

<표 3>에서는 제안한 모형과 비교대상 모형들의 예측력 차이의 통계적 유의성을 검토하기 위하여 비율에 대한 이표본 검정(two-sample test for proportions)을 실시하였다. 이 검정을 통해, 두 개의 대응 표본에 대한 두 비율(본 연구에서는 예측정확도)의 차이가 유의한지를 확인할 수 있다(Harnett and Soni, 1991). 이 검정에서, i = 1, …, n이고 j = 2, …, m일 때, 귀무가설은 $H_0 : p_i - p_j = 0$이고, 대립 가설은 $H_a : p_i - p_j > 0$이다. 여기서, p_k는 k번째 모형의 예측정확도를 의미한다.

<표 3>에서 나타난 것과 같이 본 연구에서 제안한 모형은 모든 모형에 있어서 예측성과의 차이가 통계적으로 유의한 것으로 나타났다. 구체적으로는, 인공신경망, 1-NN에 대해 1% 유의수준에서 차이가 있는 것으로 나타났으며, 전통적인 SVM, 모수 최적화 SVM(PSVM), Logit에 대해서는 5% 유의수준에서 차이가 있는 것으로 나타나서 다른 모형들과의 예측성과 차이가 모두 통계적으로 유의한 것으로 나타났다.

5. 결론

본 연구에서는 유전자 알고리즘에 기반한 사례선택기법이라는 새로운 데이터마이닝 기법을 전통적인 데이터마이닝 기법인 SVM에 결합한 새로운 SVM 기법을 제안하였다. 본 연구에서 제안한 모형은 유전자 알고리즘을 활용하여 일반적인 SVM 모형의 예측성과를 제고할 수 있는 학습용 데이터를 선택적으로 활용함으로써 예측성과를 저하시키는 데이터들을 사전적으로 데이터 집합으로부터 제거할 수 있으며 이를 통하여 예측의 효율성과 효과성을 제고할 수 있었다. 특히 가장 잡음이 심하다고 알려진 금융시계열 데이터에서 유용한 데이터들을 선별적으로 분석 데이터 집합으로부터 제거함으로써 예측성과와 일반화 가능성을 높

일 수 있음을 연구결과에서 확인할 수 있다. 연구결론에서는 제안하는 모형의 우수성을 확인하기 위하여 일반적인 데이터마이닝 기법인 인공신경망, 사례기반추론, 로지스틱 회귀분석, 전통적 SVM 모형들과 비교하였으며, 그 결과 예측성과의 개선정도가 통계적으로 유의함으로 확인하였다.

본 연구에서 제안한 사례선택 기법에 기반한 SVM 모형은 상기와 같은 연구의의를 가지고 있지만 몇 가지 한계점도 가지고 있다. 첫째, 본 연구에서는 사례선택 기법에 의해서 학습용 데이터셋을 정제함으로써 좋은 예측성과를 나타냈다. 데이터마이닝의 선행연구들에서는 유의한 변수군을 사용할 때 전체 변수군을 사용하는 모형보다 우수한 분류성과를 가져 올 수 있다고 한다. 본 연구에서도 유의한 변수선정을 위한 변수군 선택과정이 본 모형에 추가된다면 더 우수한 예측성과를 가져올 수 있을 것으로 생각되며, 이는 향후 연구과제로 제안한다. 둘째, 본 연구에는 제안한 모형의 응용가능성을 재무분야의 주가지수 예측 문제에 응용하여 확인하였다. 향후 연구에서는 다른 특성을 가진 마케팅이나 생산 분야의 데이터에 응용하여 본 연구에서 제안하는 모형의 유용성을 확인하여 그 일반화 가능성을 높일 수 있을 것이다.

참고문헌

안현철, 김경재, "다양한 다분류 SVM을 적용한 기업채권평가", *Asia Pacific Journal of Information Systems*, 19권 2호(2009), 157~178.

안현철, 김경재, 한인구, "다분류 Support Vector Machine을 이용한 한국 기업의 지능형 기업채권평가모형", *경영학연구*, 35권 5호(2006), 1479~1496.

Ahn, H. and K. Kim, "Using genetic algorithms to optimize k-nearest neighbors for data mining", *Annals of Operations Research*, Vol.163, No.1(2008), 5~18.

Chang, C.-C. and C.-J Lin, *LIBSVM : a library for support vector machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

Gates, G. W., "The reduced nearest neighbor rule", *IEEE Transactions on Information Theory*, Vol.18, No.3(1972), 431~433.

Harnett, D. L., A. K. Soni, *Statistical methods for business and economics*, Addison-Wesley, MA, 1991.

Hart, P. E., "The condensed nearest neighbor rule", *IEEE Transactions on Information Theo*ry, Vol.14(1968), 515~516.

Kim, K., "Financial time series forecasting using support vector machines", *Neurocomputing*, Vol.55(2003), 307~319.

Kim, K., "Artificial neural networks with evolutionary instance selection for financial forecasting", *Expert Systems with Applications*, Vol.30, No.3(2006), 519~526.

Kuncheva, L. I., "'Change-glasses' approach in pattern recognition", *Pattern Recognition Letters*, Vol.14(1993), 619~623.

Liu, H. and H. Motoda, "Feature transformation and subset selection", *IEEE Intelligent Systems*, Vol.13, No.2(1998), 26~28.

McSherry, D., "Automating case selection in the construction of a case library", *Knowledge Based Systems*, Vol.13, No.2/3(2000), 133~140.

Reeves, C. R. and D. R. Bush, Using genetic algorithms for training data selection in RBF networks, In Liu, H. and H. Motoda, *Instance selection and construction for data mining*, Kluwer Academic Publishers, Mas-

sachusetts, (2001), 339~356.

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

Ritter, G. L., H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, "An algorithm for a selective nearest neighbor decision rule", *IEEE Transactions on Information Theory*, Vol.21, No.6(1975), 665~669.

Smyth, B., "Case-base maintenance", *Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, (1998), 507~516.

Tay, F. E. H. and L. Cao, "Application of support vector machines in financial time series forecasting", *Omega*, Vol.29(2001), 309~317.

Tetko, I. V. and A. E. P. Villa, "Efficient partition of learning data sets for neural network training", *Neural Networks*, Vol.10, No.8 (1997), 1361~1374.

Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

Vapnik, V. N., *Statistical Learning Theory*, Wiley, New York, 1998.

Wilson, D. L., "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.2, No.3(1972), 408~421.

Wilson, D. R. and T. R. Martinez, "Reduction techniques for instance-based learning algorithms", *Machine Learning*, Vol.38(2000), 257~286.

Abstract

Optimization of Support Vector Machines for Financial Forecasting

Kyoung-jae Kim* · Hyunchul Ahn**

Financial time-series forecasting is one of the most important issues because it is essential for the risk management of financial institutions. Therefore, researchers have tried to forecast financial time-series using various data mining techniques such as regression, artificial neural networks, decision trees, k-nearest neighbor etc. Recently, support vector machines (SVMs) are popularly applied to this research area because they have advantages that they don't require huge training data and have low possibility of overfitting.

However, a user must determine several design factors by heuristics in order to use SVM. For example, the selection of appropriate kernel function and its parameters and proper feature subset selection are major design factors of SVM. Other than these factors, the proper selection of instance subset may also improve the forecasting performance of SVM by eliminating irrelevant and distorting training instances. Nonetheless, there have been few studies that have applied instance selection to SVM, especially in the domain of stock market prediction.

Instance selection tries to choose proper instance subsets from original training data. It may be considered as a method of knowledge refinement and it maintains the instance-base. This study proposes the novel instance selection algorithm for SVMs. The proposed technique in this study uses genetic algorithm (GA) to optimize instance selection process with parameter optimization simultaneously. We call the model as ISVM (SVM with Instance selection) in this study.

Experiments on stock market data are implemented using ISVM. In this study, the GA searches for optimal or near-optimal values of kernel parameters and relevant instances for SVMs. This study needs two sets of parameters in chromosomes in GA setting : The codes for kernel parameters and for instance selection. For the controlling parameters of the GA search, the population size is set at 50 organisms and the value of the crossover rate is set at 0.7 while the mutation rate is 0.1. As the stopping condition, 50 generations are permitted. The application data used in this study consists of technical indicators and the direction of change in the daily Korea stock price index (KOSPI). The total number of samples is 2218 trading days. We separate the whole data into three subsets as training, test, hold-out data set. The number of data in each subset is 1056, 581, 581 respectively.

* Department of Management Information Systems, Dongguk University_Seoul
** School of Management Information Systems, Kookmin University

This study compares ISVM to several comparative models including logistic regression (logit), backpropagation neural networks (ANN), nearest neighbor (1-NN), conventional SVM (SVM) and SVM with the optimized parameters (PSVM). In especial, PSVM uses optimized kernel parameters by the genetic algorithm. The experimental results show that ISVM outperforms 1-NN by 15.32%, ANN by 6.89%, Logit and SVM by 5.34%, and PSVM by 4.82% for the holdout data. For ISVM, only 556 data from 1056 original training data are used to produce the result. In addition, the two-sample test for proportions is used to examine whether ISVM significantly outperforms other comparative models. The results indicate that ISVM outperforms ANN and 1-NN at the 1% statistical significance level. In addition, ISVM performs better than Logit, SVM and PSVM at the 5% statistical significance level.

Key Words : Instance Selection, Support Vector Machines, Hybrid Model, Financial Forecasting, Data Mining

저 자 소 개

김경재

현재 동국대학교 경영대학 경영정보학과 부교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, *경영학연구, 지능정보연구, Annals of Operations Research*, *Applied Intelligence, Applied Soft Computing*, *Asia Pacific Journal of Information Systems*, *Computers and Operations Research*, *Computers in Human Behavior*, *Expert Systems*, *Expert Systems with Applications*, *Intelligent Data Analysis, International Journal of Electronic Commerce*, *Intelligent Systems in Accounting, Finance and Management, Neural Computing and Applications*, *Neurocomputing* 등의 학술지에 논문을 게재하였다. 연구 관심분야는 데이터마이닝, 지능형 신용평가시스템, 지식경영, 고객관계관리 등이다.

안현철

현재 국민대학교 경영대학 경영정보학부 조교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 금융, 고객관계관리 및 인터넷 보안 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.