Information Systems Review
Vol. 16, No. 3 December 2014 http://dx.doi.org/10.14329/isr.2014.16.3.161

유전자 알고리즘을 이용한 다분류 SVM의 최적화: 기업신용등급 예측에의 응용

Optimization of Multiclass Support Vector Machine using Genetic Algorithm: Application to the Prediction of Corporate Credit Rating

안 현 철 (Hyunchul Ahn) 국민대학교 경영대학 경영정보학부, 부교수

요 약

기업신용등급은 금융시장의 신뢰를 구축하고 거래를 활성화하는데 있어 매우 중요한 요소로서, 오래 전부터 학계에서는 보다 정확한 기업신용등급 예측을 가능케 하는 다양한 모형들을 연구해 왔다. 구체적으로 다중판별분석(Multiple Discriminant Analysis, MDA)이나 다항 로지스틱 회귀분석(multinomial logistic regression analysis, MLOGIT)과 같은 통계기법을 비롯해, 인공신경망(Artificial Neural Networks, ANN), 사례기반추론(Case-based Reasoning, CBR), 그리고 다분류 문제해결을 위해 확장된 다분류 Support Vector Machines(Multiclass SVM)에 이르기까지 다양한 기법들이 학자들에 의해 적용되었는데, 최근의 연구결과들에 따르면 이 중에서도 다분류 SVM이 가장 우수한 예측성과를 보이고 있는 것으로 보고되고 있다. 본 연구에서는 이러한 다분류 SVM의 성능을 한 단계 더 개선하기 위한 대안으로 유전자 알고리즘(GA, Genetic Algorithm)을 활용한 최적화 모형을 제안한다. 구체적으로 본 연구의 제안모형은 유전자 알고리즘을 활용해 다분류 SVM에 적용되어야 할 최적의 커널 함수 파라미터값들과 최적의 입력변수 집합(feature subset)을 탐색하도록 설계되었다. 실제 데이터셋을 활용해 제안모형을 적용해 본 결과, MDA나 MLOGIT, CBR, ANN과 같은 기존 인공지능/데이터마이닝 기법들은 물론 지금까지 가장 우수한 예측성과를 보이는 것으로 알려져 있던 전통적인 다분류 SVM 보다도 제안모형이 더 우수한 예측성과를 보임을 확인할 수 있었다.

키워드 : 다분류 SVM, 유전자 알고리즘, 입력변수 집합 선택, 커널 파라미터, 기업신용등급

† 본 고는 지난 2013년 한국경영정보학회 추계학술대회에서 우수논문상을 수상하여 Fast Track으로 추천되었던 학술대회 발표 논문을 수정, 보완한 논문입니다.

이 논문(저서)은 2012년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 연구되었음(NRF-2012S1A5A8024199).

Ⅰ. 서 론

기업신용등급은 시장 참여자들에게 투자대상의 안정성과 관련한 정보를 제공하는 금융시장의 가장 중요한 정보원 중 하나이다. 때문에 오래 전부터 기업신용에 대한 정확한 등급평가는 개인의 투자위험관리는 물론, 금융시장 전체의 안정에 있어 필수적으로 요구되는 매우 중요한 금융분야의 의사결정문제로 인식되어 왔다(안현철 등, 2006; Kim and Ahn, 2012). 이러한 기업신용등급평가는 현재 스탠다드 앤 푸어스(Standard and Poor's), 무디스(Moody's), 한국신용평가, 한국기업평가 등 국내외 전문 신용평가회사에 의해 이루어지고 있는데, 이들 업체들이 신용등급 정보를 제공하고 청구하는 수수료는 상당히 높은 편인 것으로 알려져 있다. 아울러, 신용평가회사들이 공시하는 신용등급은 일정 주기를 갖고 갱신되는데, 간혹 평가대상 회사가 갖고 있는 부도위험을 제 때 반영하지 못하는 경우도 종종 발생한다. 이러한 이유로 오늘날 금융권 기업들을 필두로 하여, 많은 기업들이 신용평가회사들이 갱신된 신용등급 정보를 공표하기 전에 자체적으로 투자대상이나 거래처에 대한 신용등급을 예측할 수 있는 독자적인 기업신용등급예측 모형을 개발하여 운용하고 있다(안현철, 김경재, 2009; Cao *et al*., 2006).

이처럼 기업신용등급평가는 수요가 큰 분야이고, 보다 정교한 기업신용등급평가 모형개발에 대한 산업계의 요구가 꾸준히 있어 왔기 때문에, 학계 역시 이에 부응하고자 지금까지 많은 연구들을 수행해 왔다. 초기에는 다중판별분석(multiple discriminant analysis, MDA), 다항 로지스틱 회귀분석(multinomial logistic regression analysis, MLOGIT) 등 주로 통계적인 기법을 적용한 연구들이 주류를 이루었으나(Fisher, 1959; Pinches and Mingo, 1973), 최근에는 비선형적으로 복잡한 재무 관련 데이터의 특징을 보다 효과적으로 모형에 반영할 수 있는 것으로 알려진 인공지능 기법을 이용한 연구들이 보다 활발하게 진행되고 있다(안현철 등, 2006; Cao *et al*., 2006; Huang *et al*., 2004; Kim and Ahn, 2012; Lee, 2007; Shin and Han, 1999 참고). 특히, 인공지능기법을 이용한 연구들 중에는 인공신경망(artificial neural networks, ANN)의 우수한 예측능력을 활용한 연구가 활발하게 진행되어 왔다. 그러나, 인공신경망은 학습을 위해 다량의 데이터가 필요하고, 과도적합(over-fitting) 문제로 인해 일반화의 어려움이 발생한다는 문제가 있다. 또한, 지역 최소값(local minima)을 피하기 위한 초기화 작업이 경험에 의존해야 하고, 기본적으로 '암상자 모형(black-box model)'이라서 각 변수의 중요도 등 모형을 해석하기 어렵다는 점 등도 인공신경망 기법의 한계점으로 지적될 수 있다. 특히, 기업신용등급평가와 같이 다분류(multiclass classification) 문제의 경우에는 각 등급별 데이터가 희소한 경우가 많아, 인공신경망처럼 다량의 학습데이터를 필요로 하는 모형은 많은 경우 부적합할 수 있다는 문제도 존재한다.

이러한 인공신경망의 문제를 해결할 수 있는 대안으로 최근 SVM(support vector machine)이 부상하고 있다. 기존 인공신경망 모형은 경험적 위험 최소화(empirical risk minimization) 원칙에 기반하고 있어 지역 최적화된 해로 수렴할 위험이 높지만, SVM은 구조적 위험 최소화(structural risk minimization) 원칙에 의해 학습을 수행하므로 이론적으로 전역 최적해를 얻을 수 있다(김진화 등, 2008; 안현철, 김경재, 2009). 또한 인공신경망과 비교해 최적화가 요구되는 설계 파라미터의 수도 적고, 소위 서포트 벡터(support vector)라는 이름의 경계면 주변의 데이터만 사용해 학습이 이루어지기 때문에, 적은 수의 데이터만으로도 학습이 이루어질 수 있다는 장점이 있다. 다만, 분류를 위한 SVM(support vector classification, SVC)은 본래 이분류(binary classification) 문제를 해결하기 위한 방법으로 설계된 알고리즘이기 때문에, 신용등급 분류와 같은 다분류 문제의 해결에

는 직접적으로 적용될 수 없다는 한계가 있다(Vapnik, 1995). 따라서, 수많은 학자들이 전통적인 SVM을 다분류 기법으로 확장/변형시키기 위한 다양한 접근법을 연구해 왔으며, 이렇게 개발된 기법들은 기업신용등급예측 연구에도 활발하게 적용되어 왔다. 지난 2004년 이후 수행된 경영학 분야의 기업신용등급예측 연구들을 종합해 보면, 대부분의 연구에서 다분류 SVM(multiclass SVM, 이하 MSVM) 기법들이 기존 방법론에 비해 더 우수한 예측성과를 보이는 것으로 보고되고 있다(안현철, 김경재, 2009; 안현철 등, 2006; Cao *et al*., 2006; Chen and Shih, 2006; Huang *et al*., 2004; Kim and Ahn, 2012; Lee, 2007; Zhong *et al*., 2014).

이렇게 이미 기업신용등급평가 분야에서 우수한 성과를 보이고 있는 MSVM 기법이지만, 여전히 그 성능을 보다 개선시킬 수 있는 여지는 충분히 남아있다. 우선 인공신경망에 비해 추정해야 할 파라미터의 종류가 크게 줄었다는 장점을 갖고 있기는 하지만, MSVM에서도 여전히 모형 설계자의 직관에 의해 설정되어야 할 여러 파라미터들이 존재한다. 아울러, MSVM 역시 다른 인공지능/데이터마이닝 기법과 마찬가지로 적절한 특징변수만을 활용해 모형을 학습할 경우, 그렇지 않을 경우와 비교해 예측성과를 개선할 수 있는 가능성을 가지고 있다(홍태호, 박지영, 2011; Chatterjee, 2013; Shieh and Yang, 2008). 때문에 모형에 포함될 파라미터들을 최적화하고, 학습과정에서 활용할 특징변수를 최적화한다면 보다 안정적이면서도 예측오차가 크게 줄어든 MSVM 모형을 도출하는 것이 가능하다.

이러한 배경에서 본 연구는 기업신용등급평가 모형의 예측 정확도 제고를 목표로 하여, 최근 활발하게 연구되고 있는 MSVM 기법을 보다 개선시킬 수 있는 '유전자 알고리즘(genetic algorithms, GA) 기반의 최적화 모형'을 제안한다. 본 연구의 제안모형은 OAO(One-Against-One) 방식의 MSVM에 있어, 최적 커널함수의 파라미터값과 입력변수 집합(feature subset)을 유전자 알고리즘을 활용해 탐색하도록 설계되었다. 전통적인 이분류 SVM에 대해 최적 커널함수의 파라미터 값과 입력변수 집합(feature subset)의 선택이 예측 정확도 개선에 기여할 수 있다는 연구결과들(Huang *et al*., 2007; Korkmaz *et al*., 2014; Maldonado *et al*., 2014; Min *et al*., 2006; Miranda *et al*., 2014; Zhang *et al*., 2015)이 최근까지도 계속해서 발표되고 있음을 고려할 때, MSVM에서도 동일하게 예측 정확도가 개선될 수 있을 것으로 기대된다. 본 연구에서는 이러한 제안모형의 성능을 검증하고, 실제 경영 분야 문제해결에 적용될 수 있는 가능성이 얼마나 큰 지 확인해 보기 위해, 실존하는 국내 한 대형 기업신용평가업체로부터 수집된 데이터에 제안모형을 적용해 보고 그 예측력을 다른 비교모형들과 비교해 봄으로서, 그 성능을 살펴보고자 하였다.

본 논문의 뒷부분은 다음과 같이 구성된다. 우선 제 Ⅱ장에서는 본 연구의 이론적 배경을 간략히 살펴보고, 제 Ⅲ장에서는 본 연구의 제안모형인 GAMSVM의 커널 파라미터와 입력변수 집합의 최적화모형의 과정을 소개한다. 제 Ⅳ장에서는 제안모형의 유용성을 검증하기 위한 실험 데이터 및 설계 내용을 설명하고, 최종 산출된 실험결과를 종합적으로 정리해 제시한다. 끝으로 마지막 장에서는 결언과 함께 연구의 의의와 한계점이 함께 제시된다.

Ⅱ. 이론적 배경

본 연구에서 제안하는 모형은 기본적으로 OAO 기반의 MSVM과 GA가 결합된 형태로 구성되어 있다. 이에 기존 문헌을 검토하게 될 본 절에서는 우선 MSVM의 기본적인 개념과 원리에 대해 먼저 살펴보고, 이어 GA에 대한 기본적인 소개와 함께, MSVM과 GA를 결합하고자 시도했던 기존 연구들을 살펴본다.

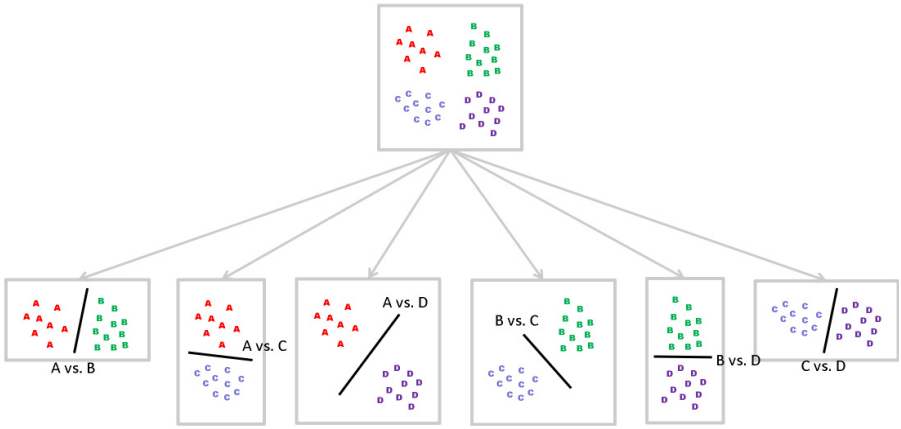
2.1 다분류 SVM

SVM은 통계학자인 Vapnik에 의해 개발된 분류기법으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간에서의 선형문제로 대응시켜 적은 수의 표본만으로도 상대적으로 우수한 예측성과를 기대할 수 있는 최근 가장 주목받고 있는 분류기법이다(김진화 등, 2008; 안현철 등, 2006; Vapnik, 1995). SVM은 구조적 위험을 최소화하는 원리를 가지고 있어, 과도적합의 위험이 높은 인공신경망과 비교해 이론적으로 더 우월한 기계학습기법으로 인식되고 있다.

하지만, 전통적인 SVM은 이분류 문제만 해결할 수 있도록 설계되어 있기 때문에 이를 다분류 문제에 적용하기 위해서는 SVM을 변형 혹은 확장시킬 필요가 있다. 지금까지 기존 SVM의 변형 혹은 확장을 통해 다분류 문제를 해결하기 위한 다양한 접근법들이 연구되어 왔는데, 이러한 접근법들은 크게 두 부류로 구분된다(Lorena and de Carvalho, 2008; Wu *et al*., 2008). 첫 번째 부류는 이분류 SVM 모형을 여러 개 만들어 결합하는 방식인데, 여기에는 우선 분류할 클래스 개수만큼 이분류 SVM 모형을 구축해, 각 등급에 해당하는 그룹과 해당하지 않는 그룹을 판별하는 형태의 One-Against-All(OAA) 방법과 분류할 전체 등급에 대해 구성될 수 있는 모든 쌍(pair)별로 독립된 SVM 모형을 구축하는 One-Against-One(OAO) 방법이 포함된다. 이 중, 일반적으로 OAO가 OAA에 비해 효율성은 떨어지나 예측정확도는 상대적으로 더 높게 나타나는 것으로 보고되고 있다(안현철, 김경재, 2009; 안현철 등, 2006). 그 외에 OAO 방법과 마찬가지로 모든 클래스 쌍 별로 독립된 SVM 모형을 구축하지만, 이후 방향성 비순환 그래프(Directed Acyclic Graph, DAG)를 이용해 보다 효율적으로 최종 클래스를 판단하도록 설계된 DAGSVM 기법도 같은 부류로 분류되는데, 여러 기존 연구들에서 OAO와 더불어 DAGSVM이 MSVM 기법들 중 가장 성능이 뛰어난 것으로 보고되고 있다(안현철, 김경재, 2009; Kim and Ahn, 2012; Lorena and de Carvalho, 2004; Wu *et al*., 2008).

MSVM을 구현하기 위한 두 번째 부류는 모든 클래스를 한 번에 고려하여 하나의 최적화 문제로 해결하는 방법이다. 이러한 부류에 속하는 방법으로 Weston and Watkins의 방법(WW)과 Crammer and Singer의 방법(CS)이 있다(안현철, 김경재, 2009; 안현철 등, 2006). 이 두 기법은 서로 매우 유사하지만, Crammer and Singer의 방법의 경우 분류평면 도출을 위해 사용되는 최적화 모형의 제약식에 있어 상대적으로 더 적은 수의 여유 변수(slack variable)를 요구한다는 점에서 차이가 있다(Crammer and Singer, 2000). 이렇게 모든 클래스를 한 번에 고려해 하나의 최적화 문제로 해결하는 접근법들은 분류평면을 찾는 과정과 방법이 수리적으로 상당히 복잡하여, 컴퓨팅 자원이 많이 요구된다(Hsu and Lin, 2002; Lorena and de Carvalho, 2008). 이러한 문제로 인해 대부분의 MSVM 연구들은 분할 정복(divide and conquer)의 원리에 기반한 OAA나 OAO, DAGSVM과 같은 기법들을 사용하고 있다(안현철, 김경재, 2009; El-Bendary *et al*., 2015; Lorena and de Carvalho, 2008).

본 연구에서는 이상의 접근법들 중에서, 보편적으로 가장 많이 활용되면서 상대적으로 높은 예측정확도를 보이는 것으로 알려져 있는 OAO 방식을 활용한다(El-Bendary *et al*., 2015). 다음의 <그림 1>은 이러한 OAO 방식의 원리를 도식으로 나타나고 있다. 이 그림에서 볼 수 있듯이, A~D까지 총 4개의 클래스로 구성된 다분류 문제에 OAO를 적용하게 되면, 총 6개(= $_3C_2$)의 이분류 SVM 모형을 구축하고, 이들의 예측결과를 투표(voting)로 종합하여 최종 클래스를 예측하게 된다.



〈그림 1〉 OAO 방식의 다분류 SVM 개념도

2.2 GA와 다분류 SVM의 결합

최근 각종 기계학습 알고리즘의 최적화와 관련한 연구들에서 최적화를 위한 수단으로 가장 많이 활용되고 있는 알고리즘 중 하나가 바로 GA다. GA는 찰스 다윈(Charles Darwin)의 적자생존의 원리(survival of the fittest)로 대표되는 생물학의 진화이론에 그 근간을 두고 있다. GA는 방대하고 복잡한 공간을 탐색하면서, 최적 혹은 유사최적(near-optimal) 결과를 찾아주는 확률적 검색방법을 이용하는데, 이러한 특징 때문에 다양한 제약식을 포함한 상황에서 목적 함수(objective function)를 최적화 하는 파라미터(parameter) 추정에 널리 적용되고 있다(김경재 등, 2006; Shin and Han, 1999).

GA는 지금까지 전통적인 이분류 SVM의 최적화를 위한 도구로 다양한 연구에서 적용되어 왔다. 예를 들어, Pai and Hong(2005), Wu *et al.*(2007), Chen and Hsiao(2008), Gu *et al.*(2011) 등은 GA를 이용해 고정된 SVM 커널함수의 파라미터들을 최적화하는 모형을 제안하였으며, Howley and Madden(2005)는 한걸음 더 나아가 GA를 이용해 커널함수의 파라미터는 물론 커널함수의 종류까지 동시에 최적화하는 모형을 제안하였다. Lee and Byun(2003), Li *et al.*(2004), Sun *et al.*(2004), Yu and Cho(2006), 그리고 Yu *et al.*(2005) 등은 GA로 SVM의 입력변수 집합을 최적화하는 모형을 제안하였다. 이들은 제안 모형을 화상 식별, 암 예측, 개인화 마케팅, 주식시장 예측 등에 적용하여 해당 분야의 예측력 개선에 제안 모형이 기여할 수 있음을 실증적으로 확인하였다. 한편 Min *et al.*(2006), Huang *et al.*(2007) 등은 상기 두 가지 접근법을 결합하여, GA를 활용해 SVM의 커널 파라미터와 입력변수 집합을 동시에 최적화하는 모형을 제시하였다. 두 연구 모두 제안모형을 기업부도예측에 적용하였는데, 적용 결과 제안 모형이 보다 적은 입력변수만 활용하면서도 더 나은 예측정확도를 산출함을 확인하였다.

이처럼 지금까지 GA를 활용해 전통적인 이분류 SVM을 최적화하려는 시도는 상당히 많이 이루어져 왔다. 하지만, GA나 다른 최적화 기법을 활용해 MSVM의 최적화를 시도하는 연구는 상대적으로 많이 찾아보기 어렵다. MSVM의 최적화를 시도한 대표적인 기존 논문으로 Lorena and de Carvalho(2008)가 있다. Lorena and de Carvalho (2008)의 경우, GA를 이용해 MSVM의 커널 파라미터를 최적화하는 모형을 제시하였다. 하지만 이 연구는 (1) 단순히 파라미터 최적화만 시도했다는 점, (2) 다분류 문제의 대상이 기업신용등급평가과 같은 경영분야의 문제가 아닌 홍채인식과 같은 공학분야의 문제였다는 점에서 본 연구와 차이를 보인다.

Shieh and Yang(2008)은 MSVM 모형에서 입

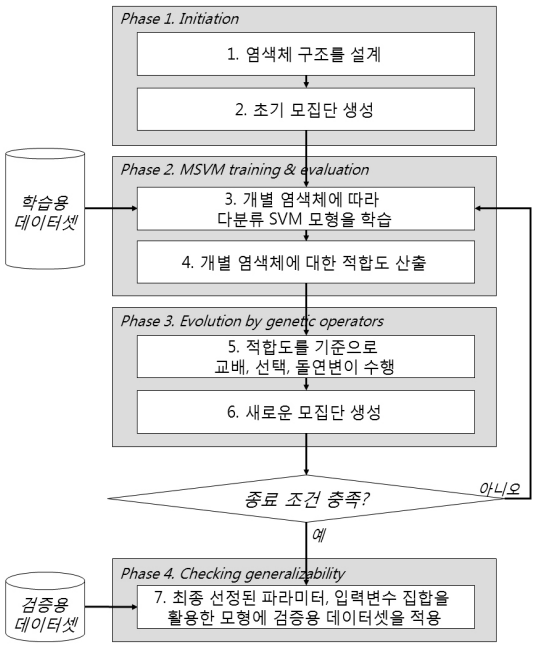
력변수 집합의 최적화를 시도했다는 점에서 본 연구와 유사한 측면이 있다. 하지만, 이 역시 (1) 파라미터의 최적화는 시도되지 않고, 단순히 입력변수 선택만 최적화가 시도되었다는 점, (2) 최적화를 위한 수단으로 GA가 아닌 RFE(recursive feature elimination) 기법이 적용되었다는 점, (3) 적용분야가 기업신용등급평가가 아닌 휴대폰 유형 분류라는 점에서 본 연구와 확연히 차별화되는 특징을 갖는다.

홍태호, 박지영(2011)은 MSVM의 입력변수 집합을 최적화하는 연구를 수행하였다. 이 연구의 경우, 제안모형을 S&P 500 기업들의 신용등급 예측에 응용하였다는 점에서 본 연구와 상당히 유사하나, (1) 앞의 Shieh and Yang(2008)과 마찬가지로 파라미터의 동시 최적화가 시도되지 않았다는 점, (2) 입력변수 집합을 선택하는 수단으로 GA가 아닌 의사결정나무의 불순도 지표(impurity measure)를 활용했다는 점에서 본 연구와 차이를 갖는다.

최근에 발표된 Chatterjee(2013)에서는 본 연구와 동일하게 MSVM 모형의 입력변수 집합과 파라미터 최적화를 GA를 통해 시도하였다. 이 연구에서 저자는 사진으로부터 추출된 정보를 활용해 석회암의 유형(6종)을 판별하는데 해당 모형을 적용하였는데, 예측 정확도와 민감도, 그리고 특이도의 관점에서 인공신경망과 같은 분류기법이나 PCA(Principal Component Analysis)와 같은 차원 축소 기법을 사용할 때보다 제안모형을 사용했을 때 더 우수한 성과를 보임을 확인하였다. 이러한 기존 연구결과로 미루어 볼 때, 비슷한 수의 클래스(4~5종)로 구분되는 기업신용등급 예측에서도 GA를 통한 MSVM 모형의 입력변수 집합 및 파라미터 최적화가 성과 개선에 효과적일 수 있음을 예상해 볼 수 있다.

Ⅲ. GA를 활용한 다분류 SVM의 커널 파라미터와 입력변수 집합의 최적화 모형

앞서 살펴본 기존 문헌들에 대한 분석결과를 토대로 본 연구에서는 기업신용등급 예측 개선을 위해 MSVM의 커널 파라미터와 입력변수 집합을 동시에 최적화하는 모형을 새롭게 제안한다. 특히 본 연구에서는 Chatterjee(2013)에서 그 효과가 입증된 GA를 활용해 최적화가 수행되는 모형을 제안한다. 편의상 본 연구에서는 제안모형을 GAMSVM(Genetic Algorithm-based Multiclass Support Vector Machine)으로 칭하기로 한다. 다음의 <그림 2>는 본 연구의 제안모형인 GAMSVM이 작동하는 과정을 시각화하여 제시하고 있다. 그림에서 볼 수 있듯이, GAMSVM은 크게 4단계 절차에 의해 구현된다.



〈그림 2〉 GAMSVM의 작동과정

1단계: 1단계는 초기화 단계이다. 이 단계에서는 최적화 탐색 대상이 되는 요인들, 즉 커널 파라미터 값들과 특징변수 선택을 적절하게 염색체(chromosome) 구조 형태로 반영하고, 이렇게 설계된 염색체들을 기반으로 적절한 난수발생을 통해 초기 모집단(population)을 생성하는 작업이 이루어지게 된다.

유전자 알고리즘을 진행시키기 위해서는 탐색공간 내의 여러 변수 집합을 염색체라 불리는 이진 선형 스트링(binary string)에 매핑(mapping)해야 한다. 이 때, 유전자 알고리즘이 공간을 효율적으로 탐색할 수 있도록 탐색 목적에 적합한 효과적인 매핑방법을 찾는 것이 중요한데, 이를 종합적으로 반영하여 본 연구에서는 하단의 <그림 3>과 같이 염색체 구조를 설계한다.

기본적으로 본 연구에서 제안하고 있는 GAMSVM은 보편적으로 가장 많이 사용되는 커널함수인 Gaussian RBF에 기반하고 있다. 다음의 식 (1)은 Gaussian RBF 커널함수의 산식을 나타낸다.

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

SVM에서 Gaussian RBF를 커널 함수로 사용할 경우 과대적합을 조절하는 파라미터인 C와 커널함수 내에 포함된 조절변수인 σ^2의 값을 적절하게 설정하여야 한다. GAMSVM에서는 이 두 값을 최적화하기 위하여, 각각 14비트씩을 할당함으로서 상당히 정밀하게 해당 값들을 탐색할 수 있게끔 하였다. 최적 입력변수 선택의 경우에는 '1'을 선택, '0'을 비선택으로 해석하여 활용할 수 있기 때문에, 이진 스트링으로 표현하는 것이 상대적으로 훨씬 용이하다. 그리하여 총 m개의 입력변수를 가지고 있는 기업신용등급평가 예측문제의 경우, GAMSVM을 적용하기 위해 요구되는 염색체의 길이는 m+28(비트)이 된다. 염색체의 설계가 끝나고 나면, 임의의 난수를 발생시켜 초기 모집단을 생성하는 초기화 작업이 이루어지게 된다.

2단계: 1단계에서 초기화 작업이 끝나고 나면, 그 다음 단계에서는 생성된 초기 모집단에 수록된 파라미터 값들을 적용하여 실제 MSVM 모형을 학습시켜보고, 해당 염색체값이 얼마나 목적에 부합하는지를 평가하는 적합도 함수값을 산출하는 과정을 거치게 된다. 본 연구에서 제안하는 GAMSVM 모형은 대용량 데이터를 이용하면서도 정확한 의사결정을 지원할 수 있도록 설계되어야 한다. 이러한 목표는 분석에 사용되는 데이터마이닝 기법의 목적함수와 GA의 적합도함수를 일치시키고 GA가 적합도함수를 최적화하기 위해 탐색할 공간을 전체 원 데이터로 설정함으로써 구현할 수 있다. 이러한 배경에서, 유전자 알고리즘의 적합도 함수는 학습용 데이터 셋에 대한 분류 정확도(classification accuracy)로 설정하였으며, 학습용 데이터 셋을 대상으로 GA가 최적의 커널 파라미터와 입력변수 선택을 탐색할 수 있도록 하였다.

3단계: 이렇게 하여 모집단에 소속된 모든 염색체들에 대한 평가가 마무리되면, 이 중 우성인 염색체를 따로 선택(selection)하고, 이들을 적절하게 교배(crossover)하며, 필요에 따라 돌연변이(mutation)도 일으키면서 새로운 모집단을 생성시키는 3단계를 수행하게 된다. 이러한 3번째 단계는 모형 설계자가 설계한 종료 조건에 도달하기 전까지 계속 반복되도록 함으로서, 수십 세대에 걸친 진화를 통해 우리가 찾고자 하는 파라미터 값들이 최적값에 수렴해 가도록 한다.

모집단	SVM 파라미터										선택						
	C					σ^2					입력 변수						
	c_1	c_2	...	c_{13}	c_{14}	g_1	g_2	...	g_{13}	g_{14}	F_1	F_2	F_3	...	F_{m-2}	F_{m-1}	F_m
염색체 1	1	1	…	0	0	1	0	…	0	1	1	0	1	…	0	0	1
염색체 2	0	1	…	0	0	1	0	…	0	1	1	0	1	…	1	0	1
염색체 3	1	0	…	1	1	1	0	…	0	1	0	1	1	…	0	0	0
⋮						⋮								⋮			
염색체 *n*	0	1	…	1	0	0	0	…	0	1	1	0	0	…	1	0	0

〈그림 3〉 GAMSVM의 염색체 구조

4단계: 종료 조건에 다다를때까지 3단계에서 진화가 충분하게 이루어지고 나면, 마지막 4단계에서는 최종 선정된 파라미터 값들을 기반으로 한 MSVM 모형을 검증용 데이터 셋에 적용해 보고, 예측성과를 최종 점검하는 작업이 이루어지게 된다. 이 과정을 통해, 제안모형이 과연 새로운 데이터(unknown data)의 예측에 있어서도 우수한 성과를 보이는지 확인하여, 제안모형의 일반화 가능성을 검증하게 된다.

Ⅳ. 실증 분석

4.1 적용 사례 및 실험 설계

본 연구에서는 GAMSVM을 검증하기 위해, 실제 현장에서 사용되는 국내 기업들의 채권등급 평가 관련 데이터에 해당 모형을 적용해 보았다. 대상이 된 데이터는 제조업으로 분류되는 KOSPI에 상장되어 있거나 KOSDAQ에 등록되어 있는 1,295개의 기업 데이터인데, 여기에는 이들 기업들의 39개 재무관련 변수들과 당해연도 회사채 신용등급 결과를 포함되어 있다. 이들 기업의 당해연도 신용등급은 국내 N 신용정보기관의 공시자료를 이용하였고, 이들의 재무제표 자료는 한국상장회사협의회에서 제공하는 데이터베이스에서 추출된 것을 사용하였다.

일반적으로 기업의 신용등급은 크게 A1, A2, A3, B, C의 5등급으로 나뉘는데, 본 연구에서는 A1을 1로, A2를 2로, A3를 3으로, 그리고 B와 C를 4로 표기하였다. B와 C를 하나의 등급으로 취급한 이유는 C등급에 속하는 사례들의 빈도가 상대적으로 너무 부족하고, 보통 신용평가회사들이 기업들에 대한 신용등급의 하한을 B등급으로 부여하는 관행을 갖고 있어 B등급 이하는 그 자체로 투자부적격채권(junk bond)의 의미로 해석할 수 있기 때문이다(안현철, 김경재, 2009; 안현철 등, 2006).

본 연구에서는 학습용 데이터로 각 등급별로 80%에 해당하는 데이터(총 1,037건)를 층화추출하여 사용하였고, 나머지 20%(총 258건)를 검증용 데이터로 사용하였다. 하지만, 제안모형의 성과를 정밀하게 검증하기에 데이터의 양이 다소 부족하다고 판단되어, 5겹 교차검증(five-fold cross validation)을 적용하였다. 초기 입력변수 후보군으로는 일원배치 분산분석(ANOVA)과 순차적 다중판별분석(Stepwise MDA)를 통해 최종 선택된 총 14개의 변수들을 사용하였다. 다음의 <표 1>은 본 연구에서 최종 선택된 14개의 독립변수를 나타내고 있다.

〈표 1〉 최종 선택된 독립변수 현황

연번	변수코드	변수명	구 분
1	X4	자기자본	규모 지표
2	X5	매출액	
3	X7	총부채	
4	X13	업력	
5	X12	주당순이익	수익성 지표
6	X15	유보액대총자산비율	
7	X16	금융비용부담율	
8	X18	금융비용대총비용비율	
9	X20	고정자산구성비율	안정성 지표
10	X23	재고자산대유동자산비율	
11	X24	단기차입금대총차입금비율	
12	X27	현금흐름대총자본비율	현금흐름 지표
13	X39	(영업활동으로인한현금흐름-현금배당)/(고정자산+운전자본)	
14	X9	1인당매출액	생산성 지표

GA 탐색을 위한 제어 파라미터로 개체군의 규모를 200개체(organisms)로 설정하였으며, 교배 및 돌연변이 비율에 대해서는 각각 50%, 10%로 설정하였다. 아울러 중지 조건으로는 10,000회 반복, 즉 50세대만큼 탐색을 반복하도록 설정하였다. 커널 파라미터의 탐색 범위와 관련해서는 기존 문헌(Tay and Cao, 2001)을 참고하여 C의 경우, $10 \leq C \leq 100$, σ^2은 $1 \leq \sigma^2 \leq 100$사이의 값을 탐색하도록 설계하였다.

본 연구의 제안모형이 상대적으로 얼마나 큰 성과개선을 도모하는지 확인하기 위해, 다수의 비교모형에 대한 실험을 추가로 수행하여 그 성과를 비교해 보았다. 비교모형으로는 기업신용등급 예측에 전통적으로 많이 적용되어 온 각종 통계 및 인공지능 기법들을 모두 적용하였다. 구체적으로 통계 모형인 (1) 다중판별분석(MDA), (2) 다항 로지스틱 회귀분석(MLOGIT), 인공지능 모형인 (3) 사례기반추론(CBR), (4) 인공신경망(ANN), 그리고 (5) 전통적인 MSVM을 실험하였는데, 전통적인 MSVM의 경우에는 대표적인 5가지 기법-즉 OAO, OAA, DAGSVM, WW, CS를 모두 적용해 보고, 이들과 비교해 제안모형인 GAMSVM이 의미 있는 성과개선을 가져오는지 확인해 보고자 하였다.

비교모형의 실험설계와 관련하여, 통계모형인 MDA와 MLOGIT은 SPSS for Windows 13.0을 활용해 단계별 변수선택법에 기반해 실험을 수행하였다. ANN은 상용 프로그램인 Neuroshell2 R4.0을 이용해 실험하였고, 입력변수와 종속변수 개수의 합을 n이라 할 때, 은닉층 노드의 수를 $n/2$, n, $3n/2$, 그리고 $2n$의 4가지 경우로 나누어 실험해 보고 이 중 가장 우수한 성과를 보이는 결과를 채택하였다. ANN 학습시 학습률과 모멘텀률은 모두 10%로 설정하였으며, 학습용 데이터셋의 25%로 구성된 테스트용 데이터셋을 기준으로 최소 오류 도달 후 38,950회가 지나도록 학습을 계속해도 더 이상 개선이 이루어지지 않을 경우에 학습을 중단하도록 하였다. CBR의 경우, 직접 개발한 Microsoft Excel VBA 프로그램을 이용해 실험하였으며, 1-NN(Nearest Neighbor)를 적용해 성과를 측정하였다. CBR에서 사례간 유사도를 측정하는 기준으로는 가장 널리 사용되는 유클리드 거리(Euclidean distance)를 사용하였다. MSVM과 GAMSVM 실험을 위한 시스템은 공개 소프트웨어인 LIBSVM v2.8(Chang and Lin, 2011)과 상용 프로그램인 Evolver 5.5를 결합시키는 Microsoft Excel VBA를 활용해 직접 개발하여 적용하였다. 단, MSVM 중 WW와 CS 모형의 경우는 BSVM v2.05를 활용해 실험하였다. 전통적인 MSVM 모형들의 경우, 최적 커널함수와 커널 파라미터 값은 그리드 탐색(grid search)을 통해 찾도록 하였다.

4.2 실험 결과

다음의 <표 2>는 5개의 데이터 셋에 GAMSVM을 적용한 결과, 최종적으로 얻어진 각 독립변수별 최적 선정 결과 및 최적 커널 파라미터 값을 나타내고 있다. 이 표에서 볼 수 있듯이, 모든 데이터 셋에서 12개의 변수만을 사용하는 것이 최적으로 나타났는데, 최적으로 선정된 변수 내역은 데이터 셋에 따라 조금씩 다르게 나타났다. 하지만, X7(총부채), X9(1인당 매출액), X13(업력), X15(유보액대총자산비율), X39((영업활동으로인

〈표 2〉 GAMSVM의 최적 변수선정 및 커널 파라미터 산출 결과

데이터셋	최적 변수 선정														C	σ^2
	X4	X5	X7	X9	X12	X13	X15	X16	X18	X20	X23	X24	X27	X39		
#1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	100	1
#2	1	1	1	1	1	1	1	0	1	1	0	1	1	1	64.81	1
#3	0	1	1	1	0	1	1	1	1	1	1	1	1	1	100	1
#4	1	0	1	1	1	1	1	1	1	1	1	0	1	1	97.28	1
#5	1	1	1	1	1	1	1	1	0	0	1	1	1	1	100	1

〈표 3〉 제안모형과 전통적인 비교모형들의 예측 성과

데이터 셋	LOGIT		MLOGIT		CBR	ANN			GAMSVM(제안모형)	
	학습	검증	학습	검증	검증	학습	테스트	검증	학습	검증
#1	65.86%	59.30%	67.70%	63.95%	48.45%	70.22%	64.73%	64.73%	80.14%	68.22%
#2	64.90%	62.02%	67.98%	63.57%	47.67%	71.76%	65.50%	65.12%	78.79%	69.38%
#3	64.51%	69.38%	65.96%	68.99%	54.26%	69.45%	63.18%	66.67%	83.51%	69.77%
#4	64.80%	65.50%	67.79%	67.44%	56.98%	69.19%	64.73%	67.44%	80.23%	72.87%
#5	65.09%	59.30%	66.15%	63.18%	49.61%	67.14%	68.60%	64.34%	79.65%	68.60%
평균	65.03%	63.10%	67.12%	65.43%	51.40%	69.55%	65.35%	65.66%	80.46%	69.77%

〈표 4〉 MSVM 비교모형들의 예측 성과

데이터 셋	OAO		OAA		DAGSVM		WW		CS	
	학습	검증	학습	검증	학습	검증	학습	검증	학습	검증
#1	75.80%	65.12%	73.10%	65.50%	65.19%	64.73%	67.12%	65.89%	81.58%	65.12%
#2	84.09%	65.50%	76.37%	63.18%	69.62%	65.12%	75.99%	64.73%	63.74%	62.79%
#3	67.60%	69.38%	63.36%	64.34%	67.60%	69.77%	66.44%	69.38%	63.26%	64.73%
#4	76.52%	68.22%	80.04%	65.12%	76.18%	68.60%	81.20%	68.99%	86.40%	66.67%
#5	75.31%	68.22%	64.80%	63.57%	75.60%	68.22%	67.31%	66.28%	64.51%	65.12%
평균	75.86%	67.29%	71.53%	64.34%	70.84%	67.29%	71.61%	67.05%	71.90%	64.89%

한현금흐름-현금배당)/(고정자산+운전자본))의 5개 변수는 데이터 셋에 관계없이 항상 선택되는 것으로 나타나, 신용등급을 설명하는데 가장 영향력이 큰 변수들임을 알 수 있었다.

커널 파라미터의 경우, C는 5개 중 4개의 데이터 셋에서 100에 가까운 큰 값이 최적값인 것으로 나타났고, σ^2은 데이터 셋에 관계없이 항상 1이 최적인 것으로 나타나 데이터 셋 간의 최적값 차이가 거의 없는 것으로 나타났다.

본 연구에서는 제안모형과 비교모형 간의 성과를 비교하기 위해, 예측 정확도(hit ratio)를 그 기준으로 사용하였다. 상단의 <표 3>에 제안모형인 GAMSVM과 비교모형인 MDA, MLOGIT, CBR, ANN의 성과가, <표 4>에 역시 비교모형인 MSVM의 5가지 모형들-OAO, OAA, DAGSVM, WW, CS-의 성과가 제시되어 있다. 전체적인 평균으로 볼 때, 제안모형의 예측정확도는 69.77%로 전통적인 비교모형들은 물론 MSVM 중에서 가장 성과가 우수한 OAO나 DAGSVM 보다도 더 우수한 예측성과를 나타내고 있음을 확인할 수 있다. 특히 모든 데이터셋에 대해 단 한 번의 예외없이 제안모형인 GAMSVM의 성과가 가장 높게 나타나고 있음을 고려할 때, 제안모형은 확실히 MSVM의 예측력을 유의미하게 개선시킬 수 있는 좋은 대안이 될 수 있을 것으로 예상된다.

<표 3>과 <표 4>에 제시된 예측 정확도의 차이가 통계적으로 유의한 지를 검증하기 위해 비모수 통계기법인 멕네마(McNemar) 검정을 수행하였다. 멕네마 검정은 4가지 전통적인 비교모형과 가장 우수한 성과를 보인 MSVM 모형이었던 OAO와 DAGSVM, 그리고 제안모형인 GAMSVM에 적용되었다. 다음의 <표 5>는 이러한 멕네마 검정의 수행 결과를 나타내고 있다. 이 표를 통해 GAMSVM이 OAO와 DAGSVM을 제외한 모든 비교모형에 대해 99% 신뢰수준 하에서, 그리고 OAO와 DAGSVM에 대해서는 95% 신뢰수준 하에서 유의한 성과의 차이를 보이고 있음을 알 수 있다.

〈표 5〉 멕네마 검정 결과

	MLOGIT	CBR	ANN	OAO	DAGSVM	GAMSVM
MDA	4.163**	49.020***	5.044**	10.859***	11.903***	23.924***
MLOGIT		73.469***	0.028	2.766*	4.069**	12.100***
CBR			76.111***	98.117***	96.111***	134.207***
ANN				1.861	2.073	11.506***
OAO					0.000	5.476**
DAGSVM						4.620**

*90% 신뢰수준 하에서 유의, **95% 신뢰수준 하에서 유의, ***99% 신뢰수준 하에서 유의.

V. 결 론

본 논문에서는 GA를 활용해 커널함수의 파라미터와 입력변수 선택을 최적화 하는 새로운 MSVM 모형(GAMSVM)을 제안하고, 이를 기업신용등급 예측 분야에 적용하여 예측성과가 가시적으로 개선됨을 확인하였다. 특히 GAMSVM의 경우, 전통적으로 사용되어 온 MDA, MLOGIT, CBR, ANN은 물론, 많은 기존 문헌에서 예측성과가 가장 우수한 기법으로 소개되어 온 MSVM과 비교해도 월등하게, 그리고 일관성 있게 우수한 성능을 나타내고 있어 향후 기업신용등급을 보다 정확하게 예측하고자 하는 산업계에서 유용하게 응용될 수 있을 것으로 기대된다.

본 연구가 갖는 의의를 살펴보면, 크게 3가지 정도를 들 수 있다. 첫째, 본 연구는 기업신용평가 분야에 MSVM 최적화 모형을 처음으로 적용해 보려고 시도하였다는 점에서 학술적인 의의를 갖는다. 앞서 문헌 고찰을 통해 살펴보았듯이, 이분류 SVM과 달리 MSVM의 경우에는 최적화에 관한 연구가 아직 걸음마 단계에 있는 실정이다. 이러한 상황에서 새로운 MSVM의 최적화 모형을 제시하고, 이를 경영분야의 다분류 문제 해결에 선도적으로 적용했다는 점에서 본 연구는 기존의 연구들과 차별화되는 특징을 갖는다고 평가할 수 있다.

둘째, 본 연구는 실증분석을 통해 국내 기업의 신용등급예측에 유의미한 지식을 발견하였다는 점에서 실무적인 의의를 갖는다. 물론 연구결과 요약에서 설명한 것처럼 예측 정확도를 크게 개선하였다는 점도 주목할 성과 중 하나겠으나, 제안모형을 통해 발견된 국내 신용등급예측의 핵심 영향 변수들을 식별할 수 있었다는 점 역시 본 연구의 주요한 발견 중 하나라 할 수 있다. 특히 핵심 영향 변수로 선정된 5개의 변수들이 규모, 생산성, 수익성, 현금흐름 등 기업을 균형잡힌 시각에서 종합적인 상태를 점검할 수 있는 지표들로 선택되었다는 점에서, 해당 변수들의 실무적 활용 가치는 상당히 높을 것으로 기대된다.

셋째, 본 연구에서 제안한 GAMSVM 모형은 모든 다분류 문제에 적용될 수 있는 범용성을 갖고 있다는 점에서 학술적, 실무적 의의를 갖는다. 비록 본 연구에서는 제안모형을 기업신용등급 예측에 적용하였지만, GAMSVM은 이론상 경영분야의 모든 다분류 문제에 적용 가능하다. 때문에, 앞으로 학계나 산업계에서 본 연구의 제안기법을 활용한 후속 연구나 사업 응용이 이루어질 수 있을 것으로 기대된다.

하지만 이러한 여러 학술적, 실무적 의의에도 불구하고, 본 연구는 다음과 같은 한계점을 갖는다. 첫째, 본 연구의 제안모형은 상당히 제한적인 영역에 한하여 최적화를 시도하고 있다는 문제가 있다. 연구에서 제안된 GAMSVM은 커널함수를 Gaussian RBF로 고정한 상태에서 커널 파라미터만 최적화하고 있는데, 기존 연구에 따르면 커널 파라미터 뿐 아니라 커널 함수도 함께 최적화할 때 더 우수한 예측성과를 도모하는

것이 가능하다(Howley and Madden, 2005). 때문에 GA의 염색체 설계를 보다 확장된 형태로 재설계하여, 커널 함수도 포함하여 최적화할 수 있는 후속 연구에 대한 필요성이 제기된다.

한 단계 더 나아가, 본 연구의 제안모형은 현재 입력변수 최적화만 시도하고 있는데, 적절한 학습표본의 선정(instance selection) 역시 이분류 SVM의 예측정확도 개선을 가져올 수 있다는 기존 연구가 이미 발표된 바 있다(안현철 등, 2006). 때문에 향후 연구에서 입력변수 집합 외에 학습표본 집합도 함께 최적화되는 MSVM 모형에 대해 실험해 보고, 과연 해당 모형이 만족할만한 예측성과의 개선을 가져오는지 확인할 필요가 있겠다.

마지막으로 앞의 연구의의에서 설명했듯이 본 연구의 제안모형은 모든 다분류 문제에 적용 가능한 범용성을 가지고 있다. 때문에 제안모형을 생산관리, 마케팅 등 다른 경영 분야의 다분류 문제 해결에 적용해 보고, 제안모형의 일반화 가능성을 점검하는 추후 후속 연구가 이루어질 필요가 있다.

참 고 문 헌

김경재, 안현철, 한인구, “유전자 알고리즘을 이용한 사례기반추론 시스템의 최적화 모형 개발: 주식시장에의 응용”, Asia Pacific Journal of Information Systems, 제16권, 제1호, 2006, pp. 71-84.

김진화, 남기찬, 이상종, “Support Vector Machine 기법을 이용한 고객의 구매의도 예측”, Information Systems Review, 제10권, 제2호, 2008, pp. 137-158.

안현철, 김경재, “Corporate Bond Rating using Various Multiclass Support Vector Machines”, Asia Pacific Journal of Information Systems, 제19권, 제2호, 2009, pp. 157-178.

안현철, 김경재, 한인구, “다분류 Support Vector Machine을 이용한 한국 기업의 지능형 기업채권평가모형”, 경영학연구, 제35권, 제5호, 2006, pp. 1479-1496.

홍태호, 박지영, “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.

Cao, L., L. K. Guan, and Z. Jingqing, “Bond rating using support vector machine”, *Intelligent Data Analysis*, Vol.10, No.3, 2006, pp. 285-296.

Chang, C.-C. and C.-J. Lin, “LIBSVM: a library for support vector machines”, *ACM Transactions on Intelligent Systems and Technology*, Vol.2, No.3, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

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 H. D. Hsiao, “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.

Chen, W.-H. and J.-Y. Shih, “A study of Taiwan’s issuer credit rating systems using support vector machines”, *Expert Systems with Applications*, Vol.30, No.3, 2006, pp. 427-435.

Crammer, K. and Y. Singer, “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.

El-Bendary, N., E. El Hariri, A. Ella Hassanien, and A. Badr, “Using machine learning techniques for evaluating tomato ripeness”, *Expert Systems with Applications*, Vol.42, No.4, 2015, pp. 1892-1905.

Fisher, L., "Determinants of risk premiums on corporate bonds", *Journal of Political Economy*, Vol. 67, 1959, pp. 217-237.

Gu, J., M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine", *Expert Systems with Applications*, Vol.38, No.4, 2011, pp. 3383-3386.

Howley, T. and M. G. Madden, "The Genetic Kernel Support Vector Machine: Description and Evaluation", *Artificial Intelligence Review*, Vol.24, No. 3-4, 2005, pp. 379-395.

Huang, Z., H. Chen, C.-J. Hsu, W-H. Chen, and S. Wu, "Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study", *Decision Support Systems*, Vol.37, No.4, 2004, pp. 543-558.

Huang, C. L., M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, Vol.33, No.4, 2007, pp. 847-856.

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

Kim, K.-J. and H. Ahn, "A Corporate Credit Rating Model using Multi-class Support Vector Machines with an Ordinal Pairwise Partitioning Approach", *Computers and Operations Research*, Vol.39, No. 8, 2012, pp. 1800-1811.

Korkmaz, S., G. Zararsiz, and D. Goksuluk, "Drug/nondrug classification using Support Vector Machines with various feature selection strategies", *Computer Methods and Programs in Biomedicine*, Vol.117, No.2, 2014, pp. 51-60.

Lee, Y.-C. "Application of support vector machines to corporate credit rating prediction", *Expert Systems with Applications*, Vol.33, No.1, 2007, pp. 67-74.

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

Lorena, A. C, and A. P. L. F. de Carvalho, "Evolutionary tuning of SVM parameter values in multiclass problems", *Neurocomputing*, Vol.71, No. 16~18, 2008, pp. 3326-3334.

Maldonado, S., J. Perez, R. Weber, and M. Labbe, "Feature selection for Support Vector Machines via Mixed Integer Linear Programming", *Information Sciences*, Vol.279, 2014, pp. 163-175.

Min, S.-H., J. Lee, and I. Han, "Hybrid genetic algorithms and support vector machines for bankruptcy prediction", *Expert Systems with Applications*, Vol.31, No.3, 2006, pp. 652-660.

Miranda, P. B. C., R. B. C. Prudencio, A. P. L. F. de Carvalho, and C. Soares, "A hybrid meta-learning architecture for multi-objective optimization of SVM parameters", *Neurocomputing*, Vol.143, 2014, pp. 27-43.

Pai, P.-F. and W.-C. Hong, "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.

Pinches, G. E. and K. A. Mingo, "A multivariate analysis of industrial bond ratings", *Journal of Finance*, Vol.28, No.1, 1973, pp. 1-18.

Shieh, M.-D. and C.-C. Yang, "Multiclass SVM-RFE for product from feature selection", *Expert Systems with Applications*, Vol.35, No. 1-2, 2008, pp. 531-541.

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

Tay, F. E. H. and L. J. Cao, "Application of support vector machines in financial time series forecasting", *Omega*, Vol.29, No.4, 2001, pp. 309-317.

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

Wu, C. H., G. H. Tzeng, Y. J. Goo, and W. C. Fang, "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.

Zhang, X., D. Qiu, and F. Chen, "Support vector machine with parameter optimization by a novel hybrid method and its application to fault diagnosis", *Neurocomputing*, Vol.149, Pt.B, 2015, pp. 641-651.

Zhong, H., C. Miao, Z. Shen, and Y. Feng, "Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings", Neurocomputing, Vol.128, 2014, pp. 285-295.

Information Systems Review
Volume 16 Number 3
December 2014

Optimization of Multiclass Support Vector Machine using Genetic Algorithm: Application to the Prediction of Corporate Credit Rating

Hyunchul Ahn*

Abstract

Corporate credit rating assessment consists of complicated processes in which various factors describing a company are taken into consideration. Such assessment is known to be very expensive since domain experts should be employed to assess the ratings. As a result, the data-driven corporate credit rating prediction using statistical and artificial intelligence (AI) techniques has received considerable attention from researchers and practitioners. In particular, statistical methods such as multiple discriminant analysis (MDA) and multinomial logistic regression analysis (MLOGIT), and AI methods including case-based reasoning (CBR), artificial neural network (ANN), and multiclass support vector machine (MSVM) have been applied to corporate credit rating.

Among them, MSVM has recently become popular because of its robustness and high prediction accuracy. In this study, we propose a novel optimized MSVM model, and appy it to corporate credit rating prediction in order to enhance the accuracy. Our model, named 'GAMSVM (Genetic Algorithm-optimized Multiclass Support Vector Machine),' is designed to simultaneously optimize the kernel parameters and the feature subset selection. Prior studies like Lorena and de Carvalho (2008), and Chatterjee (2013) show that proper kernel parameters may improve the performance of MSVMs. Also, the results from the studies such as Shieh and Yang (2008) and Chatterjee (2013) imply that appropriate feature selection may lead to higher prediction accuracy. Based on these prior studies, we propose to apply GAMSVM to corporate credit rating prediction.

* School of Management Information Systems, Kookmin University

As a tool for optimizing the kernel parameters and the feature subset selection, we suggest genetic algorithm (GA). GA is known as an efficient and effective search method that attempts to simulate the biological evolution phenomenon. By applying genetic operations such as selection, crossover, and mutation, it is designed to gradually improve the search results. Especially, mutation operator prevents GA from falling into the local optima, thus we can find the globally optimal or near-optimal solution using it. GA has popularly been applied to search optimal parameters or feature subset selections of AI techniques including MSVM. With these reasons, we also adopt GA as an optimization tool.

To empirically validate the usefulness of GAMSVM, we applied it to a real-world case of credit rating in Korea. Our application is in bond rating, which is the most frequently studied area of credit rating for specific debt issues or other financial obligations. The experimental dataset was collected from a large credit rating company in South Korea. It contained 39 financial ratios of 1,295 companies in the manufacturing industry, and their credit ratings. Using various statistical methods including the one-way ANOVA and the stepwise MDA, we selected 14 financial ratios as the candidate independent variables. The dependent variable, i.e. credit rating, was labeled as four classes: 1(A1); 2(A2); 3(A3); 4(B and C). 80 percent of total data for each class was used for training, and remaining 20 percent was used for validation. And, to overcome small sample size, we applied five-fold cross validation to our dataset.

In order to examine the competitiveness of the proposed model, we also experimented several comparative models including MDA, MLOGIT, CBR, ANN and MSVM. In case of MSVM, we adopted One-Against-One (OAO) and DAGSVM (Directed Acyclic Graph SVM) approaches because they are known to be the most accurate approaches among various MSVM approaches. GAMSVM was implemented using LIBSVM-an open-source software, and Evolver 5.5-a commercial software enables GA. Other comparative models were experimented using various statistical and AI packages such as SPSS for Windows, Neuroshell, and Microsoft Excel VBA (Visual Basic for Applications).

Experimental results showed that the proposed model-GAMSVM-outperformed all the competitive models. In addition, the model was found to use less independent variables, but to show higher accuracy. In our experiments, five variables such as X7 (total debt), X9 (sales per employee), X13 (years after founded), X15 (accumulated earning to total asset), and X39 (the index related to the cash flows from operating activity) were found to be the most important factors in predicting the corporate credit ratings. However, the values of the finally selected kernel parameters were found to be almost same among the data subsets. To examine whether the predictive performance of GAMSVM was significantly greater than those of other models, we used the McNemar test. As a result, we found that GAMSVM was better than MDA, MLOGIT, CBR, and ANN at the 1% significance level, and better than OAO and DAGSVM at the 5% significance level.

Keywords: Multiclass SVM, Genetic Algorithm, Feature Subset Selection, Kernel Parameter, Corporate Credit Rating

◐ 저 자 소 개 ◑

안 현 철 (hcahn@kookmin.ac.kr)

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

논문접수일 : 2014년 10월 07일 **게재확정일 : 2014년 12월 16일**
1차 수정일 : 2014년 12월 02일