http://dx.doi.org/10.14400/JDC.2015.13.3.155

퍼지이론과 SVM 결합을 통한 기업부도예측 최적화

최소윤, 안현철
국민대학교 비즈니스IT전문대학원

Optimized Bankruptcy Prediction through Combining SVM with Fuzzy Theory

So-Yun Choi, Hyun-Chul Ahn
Graduate School of Business IT, Kookmin University

요 약 기업부도예측은 재무 분야에 있어 중요한 연구주제 중 하나로 1960년대 이후부터 꾸준히 연구되어져 왔다. 국내의 경우, IMF 사태 이후 기업부도예측에 관한 중요성이 강조되고 있다. 이에 본 연구에서는 보다 정확한 기업부도예측을 위해 높은 예측력과 동시에 과적합화의 문제를 해결한다고 알려진 SVM(Support Vector Machine)을 기반으로 퍼지이론(fuzzy theory)을 활용해 입력변수를 확장하고, 유전자 알고리즘(GA, Genetic Algorithm)을 이용해 유사 혹은 유사최적의 입력변수집합과 파라미터를 탐색하는 새로운 융합모형을 제시한다. 제안모형의 유용성을 검증하기 위하여 H은행의 비외감 중공업 기업 데이터를 이용하여 실험을 수행하였으며, 비교모형으로는 로짓분석, 판별분석, 의사결정나무, 사례기반추론, 인공신경망, SVM을 선정하였다. 실험결과, 제안모형이 모든 비교모형들에 비해 우수한 예측력을 보이는 것으로 나타났다. 본 연구는 우수한 예측 성능을 가진 다기법 융합 모형을 새롭게 제안하여, 부도예측 분야에 학술적, 실무적으로 기여할 수 있을 것으로 기대된다.

주제어 : 부도예측, 서포트 벡터 머신, 퍼지이론, 유전자 알고리즘, 융합모형

Abstract Bankruptcy prediction has been one of the important research topics in finance since 1960s. In Korea, it has gotten attention from researchers since IMF crisis in 1998. This study aims at proposing a novel model for better bankruptcy prediction by converging three techniques - support vector machine(SVM), fuzzy theory, and genetic algorithm(GA). Our convergence model is basically based on SVM, a classification algorithm enables to predict accurately and to avoid overfitting. It also incorporates fuzzy theory to extend the dimensions of the input variables, and GA to optimize the controlling parameters and feature subset selection. To validate the usefulness of the proposed model, we applied it to H Bank's non-external auditing companies' data. We also experimented six comparative models to validate the superiority of the proposed model. As a result, our model was found to show the best prediction accuracy among the models. Our study is expected to contribute to the relevant literature and practitioners on bankruptcy prediction.

Key Words : Bankruptcy prediction, Support vector machine, Fuzzy theory, Genetic algorithm, Convergence model

Received 18 January 2015, Revised 20 February 2015
Accepted 20 March 2015
Corresponding Author: Hyunchul Ahn(Kookmin University)
Email: hcahn@kookmin.ac.kr

ISSN: 1738-1916

1. 서론

기업경영에는 데이터 분류분석을 필요로 하는 많은 경영학적 의사결정 문제들이 존재한다. 사람들은 데이터 분류분석을 통해 전문가의 판단을 보완할 수 있는 정보를 얻을 수 있다. 특히, 기업부도예측의 경우, 주주, 채권단, 협력업체, 노동자 등과 같은 이해관계자들에게 예측 가능한 손실을 최소화 할 수 있는 정보를 제공한다는 점에서 의미를 가진다[1].

기업의 부도는 사회적·경제적으로 큰 손실을 야기한다. 일반적으로 기업의 경제적 가치는 자산 가치 외에 서비스 가치, 인적가치, 상징적 가치 등이 포함되어 측정된다. 그러나 부도가 발생할 경우, 기업의 물적 자산 가치만으로 평가가 이루어지기 때문에 기업의 경제적 가치가 감소하게 된다. 이러한 경제적 가치의 감소는 채권자 및 주주의 부의 감소로 이어진다. 또한, 기업의 부도는 관련 업체 및 협력업체 등의 실적 저하로 연계되어 산업 생산력을 저하시키고 실업자를 발생시킨다. 부도기업이 해외에 인수될 경우에는 핵심기술이 유출되어 국가 산업 경쟁력을 약화시킬 수 있다. 따라서 이 같은 문제를 미연에 방지하고 피해를 최소화하기 위해 기업부도예측을 필요로 한다[2].

국내의 경우, 1997년 IMF를 겪으면서 기업부도예측에 대한 연구가 더욱 활발히 진행되고 있다. IMF가 발생하면서 한국 경제는 대기업들의 부도와 경영환경의 급변을 맞이하였다. 한보와 기아의 부도를 시작으로 대우, 현대 등에 이르기까지 한국 경제에 많은 영향을 미쳐온 기업들이 부도 혹은 경제적 어려움에 직면하였다. 이러한 기업부도는 기업을 시작으로 채권자, 주주 등의 기업관계자부터 은행에 이르기까지 연쇄적으로 영향을 미치며 경제에 역효과를 가져왔다. 대부분의 국내 금융기관들은 상당한 규모의 부실채권을 부담하게 되었으며, 이 과정에서 일부 금융기관은 합병이나 퇴출 등으로 사라졌다. 이후, 기업의 부도에 대한 사전 대비의 필요성이 대두되었으며, 정확한 기업부도예측의 중요성이 강조되었다.

기업부도예측을 위한 연구는 꾸준히 진행되어 왔다. 1960년대부터 시작된 부도예측연구는 초기에 판별분석(discriminant analysis, DA), 로짓분석(logistic regression, LOGIT), 프로빗분석(probit analysis) 등과 같은 통계적 기법이 주를 이뤘다. 1980년대 후반부터는 인공신경망(artificial neural network, ANN), 사례기반추론(case based reasoning, CBR), 의사결정나무(decision tree, DT) 등과 같은 인공지능기법을 통한 연구가 진행되고 있는데, 그 중에서도 높은 예측력을 가진다고 알려진 인공신경망이 부도예측연구에서 많이 사용되었다. 그러나 인공신경망은 입력패턴의 분포를 추정하기 위해 많은 양의 학습 데이터가 필요하며, 학습 과정에서 발생하는 과적합화(overfitting)의 문제로 인해 예측결과를 일반화 하는데 어려움이 존재한다. 뿐만 아니라 지역적 최소값을 피하기 위해 연구자의 경험이나 지식에 의존하여 초기화 작업이 진행되고, 결과를 해석하기 어렵다는 점 등이 한계로 지적되어 왔다[3].

본 연구에서는 이러한 인공신경망의 문제를 해결하기 위해 분류문제에서 각광받는 기법 중 하나인 SVM(support vector machine)을 이용하여 기업부도예측연구를 진행하였다. 또한 예측성과를 높이기 위한 방법으로 다양한 기법을 유기적으로 결합하고자 한다. 기존에도 다양한 모형을 결합하는 연구가 존재해 왔지만 이러한 연구들은 대체로 단일모형들의 결과에 대하여 가중치를 조절하는데 초점을 맞추고 있다. 반면에 본 연구의 제안 모형은 SVM에 퍼지이론(fuzzy theory)과 유전자 알고리즘(genetic algorithm, GA)을 결합하여, 분석을 수행하기 전 입력변수를 퍼지화하고, 유전자알고리즘을 이용하여 최적 혹은 유사최적의 입력변수집합과 π-fuzzy함수 및 SVM의 파라미터를 찾아내고자 한다.

본 연구는 제안모형의 우수성을 검증하기 위해 실제 기업신용평가에서 사용되는 재무데이터를 이용하여 실험을 진행하였다. 비교모형으로는 판별분석, 로짓분석, 의사결정나무, 사례기반추론, 인공신경망, SVM으로 총 6개의 단일모형을 이용하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 부도예측에 대한 선행연구와 제안 모형에 사용되는 기법들에 대하여 소개한다. 3장에서는 본 연구에서 제안하는 SVM과 퍼지 이론을 결합하여 파라미터와 입력변수집합을 최적화하는 모형의 과정과 각각의 비교모형에 대하여 설명한다. 4장에서 앞에서 제시한 모형의 유용성을 검증하기 위한 실험의 결과를 분석한다. 마지막으로 5장에서는 시사점 및 한계점과 함께 향후 연구의 발전 방향을 제시한다.

2. 문헌연구

2.1 부도예측에 관한 선행연구

기업부도예측에 관한 연구는 1960년대부터 꾸준히 이루어져 왔다. 기존 연구들을 살펴보면 1980년대 까지는 판별분석, 회귀분석, 로짓분석 등과 같은 통계적 기법이 주를 이뤘다. 1980년대 후반부터는 재무데이터의 복잡한 특성을 모형에 보다 잘 반영할 수 있다고 알려진 인공지능기법을 이용한 연구가 활발히 진행되기 시작하였는데, 그 중에서도 예측능력이 우수하다고 알려진 인공신경망을 활용한 연구가 가장 활발히 진행되었다.

통계적 기법을 사용하여 기업부도를 예측하는 연구는 Beaver의 연구[4]를 시초로 한다. Beaver[4]는 단변량 분석을 이용하여 수익성, 유동성 등과 관련된 재무비율의 평균값의 차이를 통해 기업채무상환 위험도를 예측하는 연구를 하였다. 그러나 이 연구는 재무비율의 평균값의 차이가 유의한지에 대해서 통계적인 검증이 이루어지지 않았으며, 또한 부도예측 시 오직 하나의 변수만을 고려하고 다른 변수에 대한 영향은 고려하지 못한다는 한계점을 지닌다. 이러한 단일변량분석의 한계점을 극복하기 위해 Altman[5]은 여러 개의 변수를 통합하여 분석하는 다중판별분석을 사용한 기업부도예측연구를 수행하였다. Altman은 22개의 재무비율 중 가장 예측 정확도가 높은 5개의 변수를 추출하여 이를 가중 평균하는 판별식을 제안하였다. 이는 개별적으로 관찰되던 변수들을 통합하고 단순화하여 명확한 형태의 부도예측을 가능하게 하였다. 그러나 변수의 분포가 다변량 정규분포여야 하고, 분석을 수행하기 위해 각 집단의 분산 및 공분산 구조가 동일하다는 기본조건을 충족시켜야 하는 한계점을 지닌다. 이러한 한계점을 보완하기 위하여 확률판별함수(probability discriminant function)가 제안되었다. 확률판별함수는 기업의 부도확률 또는 비부도확률을 퍼센트(percentage)로 제시함으로써 0과 1의 이분적 의사결정(dichotomous decision)을 회피하면서 동시에 적중률을 높일 수 있다. 확률모형을 이용한 대표적인 연구로는 Ohlson[6]의 로짓분석을 이용한 기업부도예측연구를 들 수 있다. Ohlson은 시그모이드(sigmoid) 함수를 로짓분석 모형에 도입하여 기업부도를 예측하고자 하였다. 로짓분석은 판별분석과 비교했을 때, 0과 1 사이의 값을 산출하여 확률적으로 계산할 수 있다는 장점을 가지지만, 판별점의 위치를 어디로 정하느냐에 따라 예측력이 변동되며, 실제의 기업부도예측에서는 판별점을 비교 및 선택할 수 없다는 점에서 예측력이 높다고 할 수 없다. 이 밖에도 프로빗분석[7], 다중회귀분석[8]과 같은 다양한 통계기법을 사용한 연구가 진행되었다.

1980년대 후반부터는 인공신경망, 사례기반추론, 유전자 알고리즘 등과 같은 인공지능 기법들을 기업부도예측 연구에서 활용하려는 움직임이 일어나기 시작했다. Odom and Sharda[9]의 연구는 신경망을 부도예측에 적용한 초기의 연구로 꼽힌다. 이들은 판별분석과 인공 신경망을 사용하여 재무비율들에 대한 분석을 수행하였으며, 그 결과 인공 신경망의 예측성과가 판별분석에 비해 더 우수함을 증명하였다. Tam and Kiang[10]은 은행의 부도여부를 인공신경망과 의사결정나무(ID3방법)를 이용한 2가지의 인공지능 기법과 판별분석, 로짓분석, k-최근접이웃방법(k-Nearest Neighbor, k-NN)를 이용한 3가지의 통계적 기법을 통해 비교분석하였다. 분석 결과, 인공 신경망이 다른 기법들에 비해 우수하다는 결론을 도출하였다. 이건창[11]은 판별분석, ACLS(Analog Concept Learning System), 인공 신경망을 이용하여 기법들 간의 비교분석을 수행하였다. 이 때, 신경망의 변수를 판별분석과 귀납적 학습방법학습방법을 통해 선정하였다. 예측결과, 귀납적 학습방법에 의해 선정된 변수로 구축된 인공 신경망이 가장 높은 예측력을 보였다. 이러한 부도예측 뿐 아니라, 다양한 분야에서 인공신경망의 우수한 예측 정확도는 최근까지도 많은 연구를 통해 증명되어 왔다[12,13].

그러나 초기 인공신경망은 주로 역전파 신경망(backpropagation network)의 형태를 가졌다. 역전파 신경망은 예측결과의 원인을 설명하기 어렵고, 신경망을 구축함에 있어 과적합화의 문제와 설계 시 많은 시간과 노력이 필요하다는 단점을 가지고 있다. 이러한 이유로 1990년대 중반부터는 다양한 형태의 신경망을 기업부도예측에 활용하였다. Serrano-Cinca[14]는 재무분야에서 자기조직화 신경망(Self Organizing Feature Maps)의 적용 가능성에 대하여 연구하였다. Yang and Honavar[15]의 연구는 PNN(Probabilistic Neural Networks)를 부도예측에 적용하고 이를 판별분석과 역전파 신경망과 비교하여 더욱 우수한 예측성과를 보임을 확인하였다. 이 밖에도 유전자 알고리즘을 이용해 부도규칙을 찾는 연구

[16], 부도예측을 위한 방법론으로 SVM을 적용해 보고, 그 성능을 다각도에서 검증한 연구[17], 최적화된 사례기반추론을 통해 부도예측의 성능을 높이고자 한 연구[18] 등 인공지능 기법을 이용한 다양한 부도예측 연구가 진행되었다.

2.2 퍼지이론

퍼지이론은 현상의 불확실한 상태를 그대로 표현해주는 방법으로 Zadeh[19]에 의해 1965년 처음 소개 되었다. 복잡한 현상에 대한 문제를 컴퓨터로 해결하고자 할 때, 컴퓨터는 정확히 표현하기 어려운 애매한 부분들을 단순화하여 처리한다. 이 과정에서 필연적으로 정보의 손실이 발생한다. 퍼지이론은 이러게 모호하게 표현된 데이터에서의 정보손실을 최소화하면서 데이터를 보다 유용하게 만들기 위해 퍼지집합(fuzzy set), 퍼지논리(fuzzy logic), 퍼지숫자(fuzzy number)들의 개념을 포함하고 있으며, 이를 처리하기 위한 수학적인 계산방법도 포함하고 있다. 따라서, 퍼지이론은 현실세계에서 수학적 모형으로 구축하기 어려운 문제에 적용하기 적합하다[20,21]. 퍼지이론은 소개된 이후로 많은 응용 분야에서 다양한 기법들과의 결합을 통해 좋은 성과를 보이고 있다.

퍼지집합은 데이터의 소속 여부가 불확실한 경우의 집합을 정의한다. '두어 개'라는 표현을 예로 들어 설명해 보자. 보통집합(crisp set)은 각각의 집합에 대한 소속여부를 {(2, 1.0), (3, 1.0)}와 같이 표현하는 반면, 퍼지집합은 각각의 집합에 속하는 정도를 {(2, 1.0), (3, 0.5)}로 표현한다. 이처럼 퍼지집합은 근사추론을 통해 집단의 소속여부가 아닌 허용 가능성으로 정의한다. 퍼지집합은 소속도(membership degree), 퍼지집합, 소속함수(membership function)로 구성된다. 소속도는 전체집합 내의 원소들이 각각의 퍼지 부분집합에 속하는 정도를 의미한다. 퍼지집합은 동일한 퍼지 부분집합에 대한 소속도를 가지는 원소들로 구성된 집합을 의미한다. 소속함수는 각각의 퍼지 부분집합에 대한 원소들의 소속도를 합리적으로 계산해 주는 함수를 말한다.

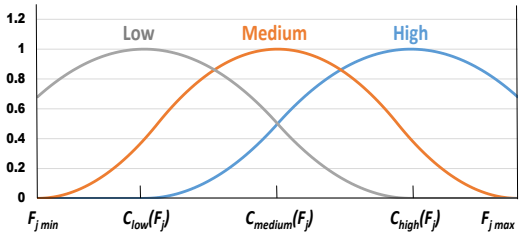
본 논문에서는 입력변수를 퍼지화하기 위한 소속함수로 김명종의 연구[22]에서 사용된 π-fuzzy 함수를 사용하였다. 이는 Pal and Pramanik[23]이 제안한 소속함수로 low, med, high라는 세 개의 퍼지집합에 대하여 0부터 1사이의 퍼지값을 계산한다.

$$\pi(F_j;c,\lambda)=\begin{cases}2(1-\frac{|F_j-c|}{\lambda})^2 & \text{for } \frac{\lambda}{2}\le|F_j-c|\le\lambda\\ 1-2(1-\frac{|F_j-c|}{\lambda})^2 & \text{for } 0\le|F_j-c|\le\frac{\lambda}{2}\\ 0 & otherwise\end{cases}$$

상기 함수식에서 F_j는 n차원에서의 j번째 투입값이고, λ는 c를 중심점으로 하는 π-fuzzy 함수의 반경을 나타낸다. F_j가 중심점 c에 위치하면 $|F_j-c|=0$이고 소속값은 최대가 된다. 즉, π-fuzzy는 1값을 가진다. 반대로 소속값이 감소하면 중심점 c로부터 멀어지므로 $|F_j-c|$의 값이 증가한다. $|F_j-c|=\lambda/2$면 소속값은 0.5이 되며 이를 교차점이라고 부른다.

$$\lambda_{medium(F_j)}=\frac{1}{2}(F_{j\max}-F_{j\min})$$
$$C_{medium(F_j)}=F_{j\min}+\lambda_{medium(F_j)}$$
$$\lambda_{low(F_j)}=\frac{1}{f_{\text{denom}}}(C_{medium(F_j)}-F_{\min})$$
$$C_{low(F_j)}=C_{medium(F_j)}-0.5\times\lambda_{low(F_j)}$$
$$\lambda_{high(F_j)}=\frac{1}{f_{\text{denom}}}(F_{\max}-C_{medium(F_j)})$$
$$C_{high(F_j)}=C_{medium(F_j)}+0.5\times\lambda_{high(F_j)}$$

F_j의 최소 및 최대의 범위는 위의 식과 같이 정의된다. f_{denom}은 매개변수로 퍼지집합의 중복(overlapping)의 규모를 통제하는 변수이다. π-fuzzy 함수에 따른 F_j의 중복 구조를 그래프로 그리면 [Fig. 1]과 같다. [Fig. 1]은 Low, Medium, High라는 3차원에서의 범위를 표현하고 있으며 f_{denom}의 값이 작아질수록 그래프의 기울기도 완만해진다.



[Fig. 1] Overlapping structure of the π-functions

2.3 Support Vector Machine

SVM은 러시아의 통계학자인 Vapnik[24]이 제안한 학습이론으로 분류문제를 해결하기 위해 최적의 분리 초평면(hyperplane)을 제공한다. 이는 입력공간과 관련된 비선형문제를 고차원의 분리 초평면에서의 선형문제로 대응시켜 나타내기 때문에 수학적인 분석이 수월하다[25]. SVM은 분류문제에 대해서 우수한 예측정확도를 보인다는 점에서 인공신경망과 같지만, 상대적으로 여러 측면에서 이점을 갖고 있다. 먼저 SVM은 '경험적 위험 최소화(empirical risk minimization)'를 추구하는 인공신경망과는 달리 '구조적 위험 최소화(structural risk minimization)'를 추구함으로써, 과적합화의 문제로부터 벗어날 수 있다. 다음으로, 인공신경망은 추정해야 할 가중치의 수가 많아 학습 시 많은 양의 데이터를 필요로 하는 반면에, SVM은 서포트 벡터(support vector)라 불리는 소수의 데이터만을 이용하여 학습에 사용하기 때문에, 일반적으로 적은 양의 학습 데이터만으로도 우수한 예측성과를 나타낸다. 마지막으로, SVM은 인공신경망에 비해 조정해야하는 파라미터의 수가 많지 않아 학습에 영향을 미치는 요소를 규명하는 것이 비교적 간단하다[26].

SVM은 두 가지 기본 아이디어를 가진다. 첫 번째 아이디어는 '최대 마진 분류(maximum margin classification)'이다. SVM은 각 그룹(class)들의 경계에 위치한 서포트 벡터로부터 가장 멀리 떨어져 있는 분류기(classifier)를 찾도록 설계된다[24,27].

두 번째 아이디어는 '고차원 공간으로의 데이터 위치 사상(mapping)'이다. 이는 선형 분류기로 분류되지 않는 비선형 문제나 저차원 공간에서의 문제를 보다 높은 차원의 공간(feature space)으로 이동시킴으로써 선형 분류기로도 분류가 가능하다[24,27]. 이 때, 고차원으로의 사상을 수행해 주는 함수를 커널함수(kernel function)라 한다. 커널함수는 원래 데이터를 고차원 공간으로 이동시킴으로써 특징공간 내에 선형으로 분리 가능한 입력 데이터셋을 만든다. 이 때 사용될 수 있는 커널함수는 여러 가지가 존재하며 어떤 커널함수를 사용하는지는 문제에 따라 상이하다. 커널함수의 선택은 SVM을 적용하는데 있어 가장 중요한 요소 중 하나로 일반적으로 많이 사용되는 커널함수로는 선형함수(linear function), 다항식함수(polynomial function), 그리고 가우시안 RBF 함수(Gaussian radial basis function)을 들 수 있다. 각각의 함수식은 아래와 같다.

선형	$K(x,y)=xy$
다항식	$K(x,y)=(xy+1)^d$
가우시안RBF	$K(x,y)=\exp(-\frac{1}{\sigma^2}(x-y)^2)$

여기서 d는 다항식 함수의 차수이고, σ^2은 가우시안 RBF 함수의 대역폭이다[28].

2.4 유전자 알고리즘

유전자 알고리즘은 확률적인 탐색이나 학습 및 최적화를 위한 기법 중 하나로, 자연에 잘 적응하는 객체는 생존하고 그렇지 못한 객체는 도태된다는 찰스 다윈(Charles Darwin)의 적자생존(survival of the fittest)의 원리와 자손의 형질은 부모의 유전자로부터 유전된다는 멘델(Mendel)의 유전법칙을 바탕으로 한다[29]. 유전자 알고리즘은 점이 아닌 개체들이 모여 이룬 개체군에 의해 병렬적 탐색이 이루어진다는 점에서 기존의 최적화 알고리즘과 차별화 된다. 또한 탐색의 방향이나 영역이 초기 설정된 값에 과도하게 의존하지 않고, 세대에 따라 확률적으로 변화한다는 점에서 전역 최적화가 가능하다는 장점을 가진다[27,30].

유전자 알고리즘은 다음과 같은 프로세스에 의해 작동된다. 먼저, 유전자 알고리즘에서 최적의 해(solution)을 찾기 위해서 주어진 문제의 해가 될 가능성을 가지는 개체 집단을 생성하는데, 이를 초기 모집단(initial population)이라고 한다. 이 과정에서 초기 모집단은 해집단 내에서 무작위로 선택되거나 경험적 방법으로 선택된다. 초기 개체 집단의 생성이 끝나면, 다음으로 생성된 개체 집단이 문제해결에 얼마나 적합한지를 평가한다. 이 때 적합도 함수(fitness function)라는 평가기준을 이용하여 각 개체(염색체)를 평가하게 된다. 이렇게 각 개체의 적합도가 평가되고 나면, 유전자 알고리즘은 평가된 개체 집단을 확률적으로 선택(selection)하게 된다. 개체들은 적합도 평가에 기초하여 도태 및 증식의 과정을 거친다. 선택된 개체들은 교배(crossover) 및 돌연변이(mutation)의 유전자 조작 과정을 통해 이전 세대(generation)와는 다른 새로운 개체들로 이루어지는 세대를 생성한다. 이렇게 생성된 새로운 세대의 개체 집단은

목표로 한 적합도 수준에 도달하거나 사전에 정해진 최대 진화 세대수에 도달할 때까지, 즉 정지조건을 만족할 때까지 선택, 교배 및 돌연변이 과정을 반복하게 된다. 이러한 과정을 거쳐 전체 생성된 세대 중 가장 최적의 적합도를 나타낸 개체 집단을 최종적으로 선택하여 그 결과를 최적 혹은 유사 최적의 해로 도출한다.

3. 실증연구

3.1 수집 및 데이터 구성

본 연구는 국내 H은행의 비외감 중공업 기업 데이터를 사용하여 실증분석을 진행하였다. 표본 기업은 자산규모가 10억 이상 70억 이하인 기업이며, 1999년부터 2001년까지 수집된 데이터를 사용하였다. 분석에 사용된 데이터는 총 1,548건으로, 학습용(training)과 검증용(validation) 데이터의 비율은 8:2로 하였다. 종속변수값은 건전과 부실로 구분되며, 여기서 부실은 부도기업 및 3개월 이상 연체기업을 의미한다. 독립변수는 총 164개 재무비율 중 독립표본 t-검정을 통해 유의한 변수를 찾아낸 뒤, 여기에 로짓분석의 단계적 변수선정방법을 통해 1차 후보변수를 선정하였다. 여기에 전문가 추천변수를 추가해 <Table 1>과 같이 총 9개 변수를 선정하였다.

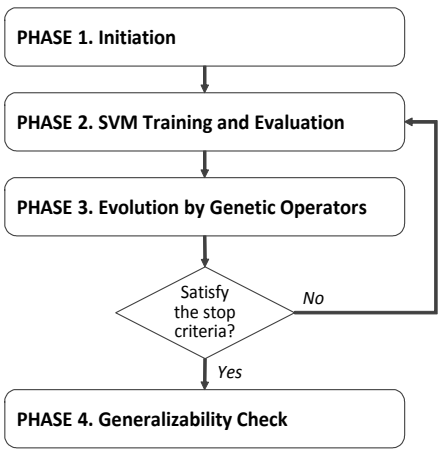
<Table 1> Selected input variables

Variable	Category	Selection by LOGIT	Selection by experts
Coefficient of sales volatility	Growth	O	O
Financial costs to sales	Profitability	O	O
Accumulated earnings to total assets	Stability	O	O
Total borrowings and bonds payable to total assets	Stability	-	O
Cash to current liabilities	Stability	O	O
Volatility of working capital to sales	Activity	O	O
Borrowings to EBITDA	Stability	O	-
Trade payable turnover period	Activity	O	-
Cashflow to debts	Cashflow	O	O

3.2 실험설계 및 분석절차

본 연구는 SVM을 개선시키기 위해 입력변수를 퍼지화하고, 유전자 알고리즘을 활용하여 π-fuzzy 함수의 f_{denom}파라미터와 입력변수집합 그리고 SVM의 커널 파라미터를 동시에 최적화하는 모형을 제안한다. 유전자 알고리즘 탐색을 위한 제어 파라미터는 해집단의 규모는 100개체(organisms), 교배율과 돌연변이율은 각각 50%와 10%로 하였다. 마지막으로 정지시점(stopping condition)은 50세대로 설정하여 총 5,000번을 반복하도록 하였다. π-fuzzy 함수의 f_{denom}의 탐색범위는 사전 실험을 통해 중복의 규모를 고려하여 $0.5 \le f_{\mathrm{denom}} \le 1.5$로 설정하였다. SVM의 커널 파라미터 탐색범위는 Tay and Cao[31]의 연구를 참고하여 C는 $10 \le C \le 100$, σ^2은 $1 \le \sigma^2 \le 100$의 값을 탐색하도록 범위를 설정하였다.

본 연구에서는 편의를 위해 제안모형을 GA Fuzzy SVM(Genetic Algorithm-based Fuzzy Support Vector Machine)라 칭하기로 한다. 분석절차는 [Fig. 2]와 같이 크게 4단계의 절차에 따라 진행된다.



[Fig. 2] GA Fuzzy SVM process

1단계. 1단계는 초기화 단계이다. 이 단계에서는 입력변수를 퍼지화하기 위해 앞에서 설명한 π-fuzzy 함수를 활용한다. 이 과정에서 9개였던 입력변수집합이 high, med, low의 3개의 퍼지집합으로 세분화 되면서 총 27개의 입력변수집합으로 증가한다. 다음으로 최적화 탐색의

대상이 되는 요인으로 SVM 커널 파라미터, π-fuzzy의 f_{denom} 파라미터, 그리고 입력변수집합을 선택하여 염색체(chromosome) 구조의 형태로 적절히 반영한다. 기본적으로 본 연구에서 제안하는 GA Fuzzy SVM 모형은 보편적으로 가장 많이 사용되는 커널함수인 가우시안 RBF 함수를 기반으로 한다. 이 같은 염색체 설계과정이 끝나면, 임의의 난수를 발생시켜 초기 모집단을 생성하는 초기화 작업이 이루어진다.

2단계. 초기화 작업이 끝나고 나면, 생성된 초기 모집단의 요인값을 실제로 이분류 SVM 모형에 적용하여 학습을 수행한다. 이 과정에서 각각의 염색체의 값이 얼마나 목적에 부합하는지를 평가하는 적합도 함수값을 산출한다. GA Fuzzy SVM 모형은 대용량의 데이터를 이용하면서도 정확한 의사결정을 지원할 수 있도록 설계되어야 한다. 이러한 목표는 분석에 사용되는 데이터마이닝 기법의 목적함수와 유전자 알고리즘의 적합도함수를 일치시키고 유전자 알고리즘이 적합도함수를 최적화하기 위해 탐색할 공간을 전체 원 데이터로 설정함으로써 구현할 수 있다. 이에 따라, 유전자 알고리즘의 적합도 함수는 학습용 데이터셋에 대한 분류 정확도(classification accuracy)로 설정하였으며, 학습용 데이터셋을 대상으로 유전자 알고리즘이 최적 혹은 유사최적의 SVM 커널 파라미터, π-fuzzy의 f_{denom} 파라미터, 그리고 입력변수집합의 선택을 탐색하도록 하였다.

3단계. 모집단에 소속된 모든 염색체들에 대한 적합도 평가가 끝나면 보다 우수한 염색체를 가진 모집단을 찾기 위해 유전자 연산을 적용한다. 먼저, 초기 모집단에서 우성인 염색체들을 선택하여 교배하고, 일정한 수준의 돌연변이를 만들어 새로운 모집단을 생성한다. 돌연변이를 통해 우연히 부모세대보다 더 우수한 염색체가 발생하기도 하지만 너무 많은 돌연변이를 생성할 경우 최적해로의 수렴이 잘 이루어지지 않을 수 있으므로 돌연변이율 설정에 유의해야한다. 이렇게 새로운 모집단의 요인값을 가지고 새로운 학습이 수행된다. 이러한 과정은 모형 설계자가 초기에 설정한 정지시점을 만족할 때까지 계속 반복됨으로써, 수십세대에 걸쳐 진화가 진행되고, 이를 통해 각 요인들에 대한 최적 혹은 유사최적의 값을 찾게 된다.

4단계. 정지조건을 만족하여 진화가 종료되면, 검증용 데이터셋을 최종적으로 선정된 요인값을 기반으로 하는 GA Fuzzy SVM 모형에 적용하여 예측성과를 점검하는 작업이 수행된다. 이 과정을 통해 제안모형의 실제로도 우수한 예측력을 보이는지 확인함으로써, 제안 모형의 일반화 가능성을 검증하게 된다.

GA Fuzzy SVM 모형 구축을 위한 소프트웨어로는 공개 소프트웨어인 LIBSVM version 2.8[32]과 상용 프로그램인 Evolver 5.5를 사용하였으며, 이 2가지 프로그램을 결합시키기 위해 Microsoft Excel VBA를 활용하여 분석 프로그램을 개발하였다.

3.3 비교모형설계

본 연구에서 제안하는 GA Fuzzy SVM 모형의 예측성능을 검증하기 위해 기존의 부도예측연구에서 사용된 로짓분석, 다중판별분석, 의사결정나무, 인공신경망, 사례기반추론, SVM의 총 6개의 비교모형[1]을 설정하고 예측정확도를 비교해 보았다. 변수는 앞에서 선정된 9개의 변수를 가지고 수행하였으며, 데이터는 학습용 데이터 80%, 검증용 데이터 20%로 하였다. 인공신경망의 경우, 최적의 학습 정지시점을 찾기 위한 과정이 추가되기 때문에, 학습용 데이터 60%, 평가용 데이터(test data) 20%, 검증용 데이터 20%로 나누었다.

로짓분석과 다중판별분석, 의사결정나무, 사례기반추론 모형의 경우, SPSS 20 소프트웨어를 사용하여 분석하였다. 로짓분석에서의 판별점은 0.5를 기준으로 하였다. 의사결정나무의 경우, CART(Classification and Regression Trees) 모형을 활용하였고, 불순도지수는 지니 지수로 측정하였다. 사례기반추론의 경우, k-NN방법을 사용하였으며, k의 범위는 1~10까지로 설정하였다.

인공신경망의 경우, 은닉층이 하나인 3층 퍼셉트론을 사용하였다. 본 연구에서는 인공신경망 모형구축을 위해 Neuroshell2 R4.0 소프트웨어를 사용하였는데, 이 때 은닉층에 포함된 노드의 수는 18개로 설정하였고, 학습률(learning rate)과 가속률(momentum rate)은 모두 0.1로 하였다. 또한 모형학습용 데이터에 과적합화 되는 것을 방지하기 위해 데이터의 평균오차가 최소값을 기록한 후

1) 비교모형으로 사용된 6가지 분류 기법들에 관한 이론적 배경과 원리에 대한 보다 상세한 설명은 전치혁[33]의 문헌을 참고하기 바란다.

<Table 2> Prediction accuracy of the models

Model	Training	Test	Validation	Settings
LOGIT	71.00%		71.29%	Enter Method
MDA	71.16%		71.94%	Enter Method
DT	71.24%		68.06%	CART (Gini) Method
ANN	71.34%	67.42%	72.58%	# of the nodes in the hidden laye r= 18
CBR (k-NN)	71.34%		68.06%	k = 6
SVM	71.65%		72.26%	Polynomial kernel, C=1, d=1
GA Fuzzy SVM	73.00%		75.48%	No. of selected features=21, f_{denom}=1.45, RBF kernel, C=99.50, σ^2=34.09

<Table 3> The results from the two-sample test for proportions

	MDA	DT	ANN	CBR	SVM	GA Fuzzy SVM
LOGIT	0.21	0.09*	0.18	0.09*	0.19	0.05**
MDA		0.07*	0.21	0.07*	0.23	0.07*
DT			0.05**	0.25	0.06*	0.01***
ANN				0.05**	0.23	0.10*
CBR(k-NN)					0.06*	0.01***
SVM						0.09*

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

50,000회가 지나면 학습이 정지되도록 하였다. 한편 SVM의 경우, 선형과 다항식, 그리고 가우시안 RBF 커널 함수를 사용하였다. SVM의 성능에 있어서는 커널함수의 상한 C와 커널 파라미터 σ^2, d가 중요한 역할을 한다고 알려져 있다[31]. 이에 본 연구는 모형의 탐색범위를 각각 $10 \leq C \leq 100$, $1 \leq \sigma^2 \leq 100$, $1 \leq d \leq 5$로 설정하였다. SVM 실험에는 LIBSVM[32]이 사용되었다.

4. 실험결과

앞에서 작성한 실험설계를 바탕으로 모형별 기업부도예측실험을 수행하였다. 아래의 <Table 2>는 제안모형인 GA Fuzzy SVM과 로짓분석, 다중판별분석, 의사결정나무, 인공신경망, 사례기반추론, SVM의 총 6개의 비교모형의 실험결과를 종합한 표이다. 실험 결과에 의하면 제안모형의 예측정확도가 75.48%로 기존 기법들에 비해 가장 우수한 예측정확도를 보이는 것을 알 수 있다. 또한 학습용 모형과 검증용 모형간의 예측정확도 차이도 적어 상당히 안정된 결과를 보여주고 있다. 이를 통해 퍼지이론의 결합과 유전자 알고리즘을 통한 파라미터 및 입력변수집합 선정이 기업부도예측모형 구축에 있어 효과적임을 증명하였다.

이어 제안모형과 비교모형들 간의 예측정확도의 차이가 통계적으로 유의한 지를 알아보기 위해 2 Sample Test for Proportions를 실시하였다. 그 결과는 다음의 <Table 3>와 같다. <Table 3>의 결과에 의하면 제안모형이 다중판별분석, 인공신경망, 그리고 SVM의 결과와 90%, 로짓분석의 결과와 95%, 의사결정나무와 사례기반추론의 결과와는 99% 신뢰수준 하에서 통계적으로 유의한 차이를 보이고 있다. 따라서 비교 모형들 간의 예측정확도의 차이가 통계적으로 유의함을 확인할 수 있다.

5. 결론

본 연구에서는 이해관계자들에게 경영학적 의사결정 문제를 해결하기 위한 정보를 제공하고, 기업부도로 인해 발생하는 사회적·경제적 손실을 미연에 방지하기 위해 보다 높은 예측정확도를 가지는 새로운 기업부도예측

모형을 제안하였다. GA Fuzzy SVM으로 명명된 본 연구의 제안모형은 설명력과 일반화 성능이 우수하다고 알려진 SVM과 퍼지이론을 결합하여 입력변수집합을 확장하고, 유전자 알고리즘을 이용해 파라미터와 입력변수집합을 최적화하고자 하였다. 제안모형의 기업부도예측 적용 가능성을 확인하고자 실제 데이터에 로짓분석, 다중판별분석, 의사결정나무, 인공신경망, 사례기반추론, SVM과 같은 비교모형과 제안모형을 동시에 적용해 본 결과, 제안모형이 기징 우수한 예측정확도를 가지는 것으로 나타났고, 그 차이가 통계적으로도 유의한 것으로 확인되었다.

본 연구의 대상이 되고 있는 비외감 기업의 경우, 공인회계사로부터 감사를 받지 않기 때문에 이들이 발표하는 회계정보에 대한 신뢰성이 상대적으로 부족하다. 따라서 비외감 기업의 경우, 재무제표로부터 산출된 재무비율 정보를 이용해 정확하게 기업의 부도를 예측하는 것이 상당히 어려운 편이다. 본 연구에서는 이러한 산업 현장의 어려움을 해결하기 위한 대안으로 퍼지이론과 SVM, 그리고 유전자 알고리즘 간의 결합을 통한 새로운 해법을 제시하였다는 점에서 학술적으로 그리고 실무적으로 의의를 갖는다.

그러나 본 연구는 다음과 같은 몇 가지 한계점을 지닌다. 첫 번째로 본 연구의 제안모형은 최적화의 영역이 상당히 제한적이라는 문제를 가진다. 제안모형인 GA Fuzzy SVM은 커널함수를 가우시안 RBF 함수로 고정하고 커널 파라미터만 최적화 하고 있다. 그러나 앞 절의 설명에 따르면 커널 파라미터와 커널 함수도 함께 최적화를 수행할 때 예측성과를 더 향상시키는 것이 가능하다. 따라서 유전자 알고리즘의 염색체 설계를 커널함수를 포함하는 형태로 재설계하여 최적화할 수 있는 후속 연구가 필요하다.

두 번째로 본 연구의 제안모형은 입력변수의 최적화만 시도하고 있다. 그러나 앞에서 설명한 내용에 따르면 적절한 학습표본의 선정도 이분류 SVM의 예측정확도를 향상시키는데 영향을 미친다. 그러므로 향후 연구에서는 입력변수집합 뿐만 아니라 학습표본집합도 함께 최적화하는 SVM 모형에 대해 실험해 보고, 예측성과가 만족할 만한 수준까지 개선이 되는지 확인해볼 필요성이 있다.

셋째로 본 연구에서 실증분석에 사용된 기업부도예측 데이터가 2000년대 초반에 수집된 데이터라는 점 역시 또 다른 한계라고 할 수 있다. 향후 수행될 연구에서는 보다 최신의 데이터를 활용한 분석을 통해, 제안모형이 현 시점의 기업부도예측에서도 유효한 성능을 보이는지 확인할 필요가 있다.

마지막으로 본 연구의 제안모형은 기업부도예측이라는 특정 분야에 대해서만 실험이 진행되었기 때문에 일반성이 충분히 검증되지 못하였다. 따라서 주가시장예측이나 고객 분류와 같은 다른 분야의 데이터를 사용하여 제안모형의 일반성을 검증하는 것이 향후 연구과제가 될 수 있을 것이다.

REFERENCES

[1] J. L. Bellovary, D. E., Giacomino, & M. D. Akers, A review of bankruptcy prediction studies: 1930 to present. Journal of Financial Education, pp. 1-42, 2007.

[2] S. Kim, C. S. Park, & S. M. Jeon, Default Decisions of FIs and Endogeneity Problems in Default Prediction. Journal of Business Research, Vol. 26, No. 1, pp. 99-132, 2011.

[3] J. M. Park, Bankruptcy Prediction using Support Vector Machine. Korea Advanced Institute of Science and Technology, Master's Thesis, 2003.

[4] W. H. Beaver, Financial ratios as predictors of failure. Journal of Accounting Research, Vol. 4, pp. 71-111, 1966.

[5] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, Vol. 23, No. 4, pp. 589-609, 1968.

[6] J. A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, pp. 109-131, 1980.

[7] M. E. Zmijewski, Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, pp. 59-82, 1984.

[8] R. O. Edmister, An empirical test of financial ratio analysis for small business failure prediction. Journal of Financial and Quantitative Analysis, Vol. 7, No. 2, pp. 1477-1493, 1972.

[9] M. D. Odom, & R. Sharda, A neural network model for bankruptcy prediction. In proceedings of the International Joint Conference on Neural networks, Vol. 2, pp. 163-168, 1990.

[10] K. Y. Tam, & M. Y. Kiang, Managerial applications of neural networks: the case of bank failure predictions. Management science, Vol. 38, No. 7, pp. 926-947, 1992.

[11] K. C. Lee, A Comparative Study on the Bankruptcy Prediction Power of Statistical Model and AI Models: MDA, Inductive Learning, Neural Network. Journal of The Korean Operations Research and Management Science Society , Vol. 18, No. 2, pp. 57-81, 1993.

[12] K. Y. Kim, G. R. Lee, & S. W. Lee, A Comparative Analysis of Artificial Intelligence System and Ohlson model for IPO firm's Stock Price Evaluation. Journal of Digital Convergence, Vol. 11, No. 5, pp. 145-158, 2013.

[13] K. K. Seo, Development of a Sales Prediction Model of Electronic Appliances using Artificial Neural Networks. Journal of Digital Convergence, Vol. 12, No. 11, pp. 209-214, 2014.

[14] C. Serrano-Cinca, Self organizing neural networks for financial diagnosis. Decision Support Systems, Vol. 17, No. 3, pp. 227-238, 1996.

[15] J. Yang, & V. Honavar, Feature subset selection using a genetic algorithm. Computer Science Technical Reports, Paper 156, 1997.

[16] K. S. Shin, & Y. J. Lee, A genetic algorithm application in bankruptcy prediction modeling. Expert Systems with Applications, Vol. 23, No. 3, pp. 321-328, 2002.

[17] K. S. Shin, T. S. Lee, & H. J. Kim, An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, Vol. 28, No. 1, pp. 127-135, 2005.

[18] H. Ahn, & K. J. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Applied Soft Computing, Vol. 9, No. 2, pp. 599-607, 2009.

[19] L. A. Zadeh, Fuzzy sets. Information and Control, Vol. 8, No. 3, pp. 338-353, 1965.

[20] S. H. Lee, K. I. Moon, & S. J. Lee, Application of Fuzzy Logic in Scenario Based Language Learning. Journal of Digital Convergence, Vol. 11, No. 2, pp. 221-228, 2013.

[21] L. A. Zadeh, Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, Vol. 1, pp. 3-28, 1978.

[22] M. Kim, The Application of Knowledge Integration Using Fuzzy Logic and Genetic Algorithms to Financial Market. Korea Advanced Institute of Science and Technology, Doctoral Thesis, 2004.

[23] S. K. Pal, & P. K. Pramanik, Fuzzy measures in determining seed points in clustering. Pattern Recognition Letters, Vol. 4, No. 3, pp. 159-164, 1986.

[24] V. Vapnik, Statistical learning theory. Wiley, New York, 1998.

[25] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, & B. Scholkopf, Support vector machines. IEEE Intelligent Systems and Their Applications, Vol. 13, No. 4, pp. 18-28, 1998.

[26] H. Ahn, K. J. Kim, & I. Han, Purchase Prediction Model using the Support Vector Machine. Journal of Intelligence and Information Systems, Vol. 11, No. 3, pp. 69-81, 2005.

[27] S. W. Kim, & H. Ahn, Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms. Journal of Intelligence and Information Systems, Vol. 16, No. 1, pp. 71-92, 2010.

[28] H. Ahn, & K. J. Kim, Corporate Bond Rating Using Various Multiclass Support Vector Machines. Asia Pacific Journal of Information Systems, Vol. 19, No. 2, pp. 157-178, 2009.

[29] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[30] Y. Cha, G. Lee, J. Lee, D. Y. Wie, Optimization of the Distribution Plan and Multi-product Capacity

using Genetic Algorithm. Journal of Digital Convergence, Vol. 12, No. 6, pp. 125-134, 2014.

[31] F. E. Tay, & L. Cao, Application of support vector machines in financial time series forecasting. Omega, Vol. 29, No. 4, pp. 309-317, 2001.

[32] C. C. Chang, & C. J. Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 27:1-27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[33] C. H. Jeon, Data Mining Techniques. Hannarae Publishing Co., Seoul, 2012.

소 윤(Choi, So Yun)

・2013년 2월 : 국민대학교 경영대학 경영정보학부
・2013년 3월 ~ 현재 : 국민대학교 비즈니스IT전문대학원 비즈니스IT 전공
・관심분야 : Technical MIS
・E-Mail : csy010@kookmin.ac.kr

안 현 철(Ahn, Hyunchul)

・1999년 2월 : KAIST 산업경영학과 (이학사)
・2002년 8월 : KAIST 테크노경영대학원 경영공학전공(공학석사)
・2006년 8월 : KAIST 테크노경영대학원 경영공학전공(공학박사)
・2009년 3월 ~ 현재 : 국민대학교 경영정보학부 부교수
・관심분야 : 경영정보시스템, 지능형의사결정지원시스템
・E-Mail : hcahn@kookmin.ac.kr