부도 예측을 위한 앙상블 분류기 개발†

(Developing an Ensemble Classifier for Bankruptcy Prediction)

민 성 환*
(Sung-Hwan Min)

요 약 분류기의 앙상블 학습은 여러 개의 서로 다른 분류기들의 조합을 통해 만들어진다. 앙상블 학습은 기계학습 분야에서 많은 관심을 끌고 있는 중요한 연구주제이며 대부분의 경우에 있어서 앙상블 모형은 개별 기저 분류기보다 더 좋은 성과를 내는 것으로 알려져 있다. 본 연구는 부도 예측 모형의 성능개선에 관한 연구이다. 이를 위해 본 연구에서는 단일 모형으로 그 우수성을 인정받고 있는 SVM을 기저 분류기로 사용하는 앙상블 모형에 대해 고찰하였다. SVM 모형의 성능 개선을 위해 bagging과 random subspace 모형을 부도 예측 문제에 적용해 보았으며 bagging 모형과 random subspace 모형의 성과 개선을 위해 bagging과 random subspace의 통합 모형을 제안하였다. 제안한 모형의 성과를 검증하기 위해 실제 기업의 부도 예측 데이터를 사용하여 실험하였고, 실험 결과 본 연구에서 제안한 새로운 형태의 통합 모형이 가장 좋은 성과를 보임을 알 수 있었다.

핵심주제어 :서포트벡터머신, 부도예측, 배깅, 랜덤서브스페이스, 앙상블

Abstract An ensemble of classifiers is to employ a set of individually trained classifiers and combine their predictions. It has been found that in most cases the ensembles produce more accurate predictions than the base classifiers. Combining outputs from multiple classifiers, known as ensemble learning, is one of the standard and most important techniques for improving classification accuracy in machine learning. An ensemble of classifiers is efficient only if the individual classifiers make decisions as diverse as possible. Bagging is the most popular method of ensemble learning to generate a diverse set of classifiers. Diversity in bagging is obtained by using different training sets. The different training data subsets are randomly drawn with replacement from the entire training dataset. The random subspace method is an ensemble construction technique using different attribute subsets. In the random subspace, the training dataset is also modified as in bagging. However, this modification is performed in the feature space. Bagging and random subspace are quite well known and popular ensemble algorithms. However, few studies have dealt with the integration of bagging and random subspace using SVM Classifiers, though there is a great potential for useful applications in this area. The focus of this paper is to propose methods for improving SVM performance using hybrid ensemble strategy for bankruptcy prediction. This paper applies the proposed ensemble model to the bankruptcy prediction problem using a real data set from Korean companies.

Key Words : Support vector machines; Bankruptcy prediction; Bagging; Random Subspace; Ensemble

† 이 논문은 2011년도 한림대학교 교비 학술연구비(HRF-2011-017)에 의해 연구되었음
* 한림대학교 경영학부

1. 서 론

기업의 부도를 예측하는 것은 회계나 재무 분야에서 오랜 기간 동안 연구되어온 중요한 연구 주제이다. 기업의 부도는 주주, 종업원, 금융 기관 등의 이해관계자뿐만 아니라 경제 전반에도 영향을 미칠 수 있는 매우 중요한 일이다. 그러므로, 보다 정확한 부도 예측 모형을 개발하는 일은 매우 중요한 일이라 할 수 있다. 그 동안 부도예측 모형의 성과를 개선하기 위해 많은 통계 기법들이 적용되어 왔다. [1]은 부도 예측을 위해 재무비율에 대한 단일변량 판별 분석를 제안했다. 그 후 [2]는 다변량 판별분석 적용하여 재무비율을 이용한 z-score 예측모형을 제시하였다. 그 이후 판별분석[3], 다중회귀분석[4], 로지스틱 회귀분석[5,6,7]과 같은 통계적 기법들을 이용해 기업의 부도를 예측하려는 많은 연구가 있어 왔다. 그러나, 전통적인 통계적인 기법은 선형성, 정규성 등과 같은 엄격한 가정들을 하고 있어 부도 예측과 같은 복잡한 현실 세계 문제에의 적용을 제한하고 있다.

최근에는 보다 정확한 부도 예측 모델 개발을 위해 전통적인 통계적 기법 뿐만 아니라 다양한 데이터 마이닝 기법을 사용하고 있다. 귀납적 학습(Inductive learning) [8,9], 사례기반추론(Case-based reasoning) [10,11], 그리고 인공신경망[12,13,14], SVM(Support Vector Machine)[15,16,17,18]까지 많은 연구자들의 연구가 이어지고 있다.

한편, 개별 단일 모형들의 성과 개선을 위해 여러 개의 서로 다른 분류기(Classifier)들을 적절하게 결합하여 앙상블(Ensemble) 모형을 구축하려는 연구가 기계 학습(machine learning) 분야에서 많은 관심을 끌고 있다. 앙상블 모형은 대부분의 경우에 있어서 개별 분류기들의 성과보다 더 좋은 성과를 내는 것으로 알려져 있다. 부도 예측에도 앙상블 학습을 적용하려는 연구가 최근에 활발하게 진행되고 있다. [19]는 유전자 알고리즘을 이용하여 분류기 앙상블의 최적을 찾으려고 하였다. 의사결정 트리와 인공신경망을 기저 분류기로 하여 각각 단일모형, boosting, bagging을 수행한 결과 유전자 알고리즘을 이용한 boosting, bagging의 최적 분류기를 선택하는 모형이 가장 우수한 성과를 보였다. [20]은 의사결정 트리의 일종인 CART를 기저 분류기로 하는 변형된 bagging 모형을 부도 예측 문제에 적용하였다.

분류기의 앙상블은 여러 개로 구성된 기본 분류기들로 구성되어 있다. 이들은 각각 개별적으로 모형을 학습시키며 이들 개별 분류기의 결과는 여러 다양한 방법에 의해 합쳐진다. 대부분의 경우에 있어서 이들 앙상블 모형은 개별 기저 분류기보다 더 좋은 성과를 내는 것으로 알려져 있다 [21]. 앙상블 모형의 성과가 개별 분류기들의 성과보다 더 좋아지게 하기 위해서는 기저 분류기들의 성과가 좋아야 하며 또한 기저 분류기들의 예측 결과값이 가능한 한 다양성을 가져야 한다. 즉 앙상블을 구성하는 기저 분류기들의 예측 결과값이 가능하면 서로 다른 결과값을 가져야 한다. 기저 분류기의 다양성을 확보하기 위한 많은 전략들이 개발되어 왔으며 [22]는 기저 분류기들의 다양성을 확보하기 위한 전략으로 서로 다른 분류 모형 사용하기, 서로 다른 입력 변수(attribute subsets) 사용하기 그리고 서로 다른 훈련 데이터 사용하기를 제안하고 있다. 이 중에서 서로 다른 훈련데이터 사용하기와 서로 다른 입력 변수 사용하기는 가장 대표적인 앙상블 전략으로 여겨지고 있으며 많이 활용되고 있는 방법이다. 서로 다른 훈련데이터 셋을 사용함으로써 앙상블의 성과를 개선하고자 하는 접근방법으로는 bagging과 boosting이 있으며, 서로 다른 속성 집합을 사용함으로써 기저 분류기의 다양성을 확보하고 이를 통해 앙상블의 성과를 개선하기 위한 접근방법으로는 random subspace 방식이 있다.

앙상블 학습 중에 bagging은 가장 대표적인 통합모형 방법으로 많은 선행 연구에서 단일 모형보다 더 좋은 성과를 내는 것으로 알려져 왔다. 하지만, 의사결정 트리(Decision Tree)나, 인공신경망(Neural Network)을 기저 분류기로 하는 bagging에 대한 연구는 많았지만 성과가 우수한 SVM의 bagging 통합에 대한 연구는 상대적으로 많지 않았다. 선행 연구에 의하면 bagging은 모형이 안정적인 SVM 보다는 의사결정 트리와 같은 기저 분류기 모형의 통합에서 좋은 성과를 보여왔다. [23]에 의하면 bagging은 의사결정 트리와 같이 모델이 불안정한 기저 분류기를 통합할 때는 그 효과가 좋지만 SVM에 같이 기저 분류기가 안정적인 모형일 때는 그 효과가 떨어진다.

본 논문은 부도 예측 모형의 성능개선에 관한 연구이다. 이를 위해 본 연구에서는 단일 모형으로 그 우수성을 인정받고 있는 SVM의 앙상블 모형에 대해 고찰하고자 한다. SVM 모형의 성능 개선을 위해 bagging

과 random subspace 모형을 부도 예측 모형에 적용해 보았다. 또한 bagging 모형과 random subspace 모형의 성과 개선을 위해 새로운 Ensemble 모형을 제안하고 제안한 모형의 성과를 검증하기 위해 실제 기업의 부도 예측 데이터를 사용하였다. 본 논문에서 제안한 모형의 성과를 검증하기 위해 기존의 단일 모형, 전통적인 bagging 앙상블 모형, random space 앙상블 모형, 그리고 본 연구에서 제안한 bagging과 random space와의 통합 모형의 성과를 비교해 보았다. 실험결과 본 논문에서 제안한 새로운 형태의 통합 모형이 가장 좋은 성과를 보임을 알 수 있었다.

본 논문의 구성은 다음과 같다. 다음 장에서는 본 연구에서 기저 분류기로 사용한 SVM에 대해 설명하고 앙상블에 대해 설명을 한다. 3장에서는 본 연구에서 부도 예측 모형으로 제안한 SVM 앙상블 모형에 대해 설명한다. 그리고, 4장, 5장에서는 제안한 다양한 모형의 검증을 위한 실험 설계 및 실험 결과에 대한 설명을 한다. 마지막 장에서는 본 연구의 의의와 한계점 및 향후 연구 방향에 대해 고찰해 본다.

2. 방법론

2.1 SVM

SVM(Support Vector Machine)은 [24]에 의해 개발된 방법으로 패턴인식과 같은 공학 분야에서 널리 사용될 뿐만 아니라 최근에는 다양한 분야의 예측 문제에도 사용되고 있다. [25]는 이동 차량에서의 실시간 번호판 인식 문제를 SVM 모형을 이용해 해결하려고 하였으며, [26]은 근적외선 얼굴인식 문제에 그리고, [27]은 매장문화재에 대한 분류 문제에 SVM을 적용하여 좋은 성과를 거두었다.

신경망 기법을 포함하여 대부분의 패턴 인식을 위한 전통적인 기법들이 학습 데이터의 수행도를 최적화 하기 위한 경험적인 위험을 최소화(Empirical Risk Minimization)하는데 기초한 반면, SVM은 고정되어 있지만 알려지지 않은 확률 분포를 갖는 데이터에 대해 잘못 분류하는 확률을 최소화하는 구조적인 위험(Structural Risk Minimization)을 최소화하는 것에 기초하고 있다.

SVM은 Structural Risk Minimization의 이론으로부터 발전한 이론이다. SVM은 두 개의 클래스를 분류할 수 있는 초평면(hyperplane)을 찾아냄으로써 두 개의 클래스를 분류하는 방법이다. SVM은 두 개의 클래스를 분류할 수 있는 여러 개의 초평면 중에서 클래스간의 최소거리를 최대로 하는 초평면을 찾는다. (두 클래스간의 마진(magin)을 최대화하는 초평면을 찾는다.)

SVM은 선형분류 뿐만 아니라 비선형 분류도 효율적으로 수행할 수 있다. 비선형 분류 문제에서는 클래스를 최적으로 분리할 수 있는 초평면(optimal separating hyperplane)이 존재하지 않을 수 있는데 이 경우 SVM은 커널 트릭(kernel trick:커널 대치)을 사용함으로써 이 문제를 해결한다. 커널 트릭을 통해 원래의 데이터의 입력공간(input space)을 보다 높은 차원의 feature space로 매핑(mapping; 사상)시키는 것이 가능해지고 이를 통해 SVM은 더 높은 차원의 새로운 feature space에서 최적의 분리가능한 초평면(optimal separating hyperplane)을 찾을 수 있게 된다. SVM에서 자주 사용하는 커널 함수로는 아래와 같은 것이 있다.

- 선형 커널(Linear kernel):
  $k(x, y) = x \bullet y$
- 다항식 커널(Polynomial kernel): d의 degree를 가짐: $k(x, y) = (x \bullet y + 1)^d$
- RBF 커널(Gaussian radial basis functionl kernel): 파라미터 γ를 가짐:
  $k(x, y) = \exp(-\gamma||x - y||^2)$

2.2 앙상블(Ensemble) - Bagging, Random Subspace

여러 모형을 통합하는 앙상블을 구성하는 기본 모형을 기저 분류기(classifier)라고 하고 base learner라고 부른다. 분류기의 앙상블 학습은 여러 개의 서로 다른 분류기들의 조합을 통해 만들어진다. 앙상블 학습의 구축은 기계학습(machine learning) 분야에서 많은 관심을 끌고 있는 중요한 연구주제이다. 많은 선행연구에 의하면 앙상블 학습은 많은 경우에 있어서 단일 모형보다 우수한 성과를 보였다.

앙상블 모형의 성과가 개별 분류기들의 성과보다 더 좋아지게 하기 위해서는 기저 분류기들의 성과가 좋아야 하며 또한 기저 분류기들의 예측 결과값이 가

능한 한 다양성을 가져야 한다. 즉 앙상블을 구성하는 기저 분류기들의 모형에 의한 예측 결과값이 가능하면 서로 다른 결과값을 가져야 한다. 분류기의 앙상블을 만들기 위한 많은 방법이 있다. 가장 대표적인 방법 중의 하나는 패턴에 변화를 주는 방법이다. 원래의 학습데이터 셋에 변화를 줌으로써 새로운 형태의 학습데이터를 생성할 수 있다. 이런 대표적인 방법이 bagging 방법이다[23]. Bagging 방법에서 새로운 형태의 데이터 셋은 원래 학습데이터의 부분집합들이다. Bagging은 bootstrap aggregating의 준말로 [23]에 의해 처음 소개되었다. Bagging은 가장 대표적인 앙상블 학습 알고리즘 중의 하나로 방법은 매우 단순하지만 그 성과는 매우 좋은 것으로 알려져 있다.

입력변수(feature)의 변화도 새로운 형태의 학습 데이터를 발생시킬 수 있다. 이를 통해 앙상블 학습이 가능하다. 이와 같은 형태의 입력변수의 변화를 주는 기법 중의 하나는 random subspace방법이다[28]. Random subspace방법을 통해 전체 feature set의 부분집합을 생성할 수 있고 이를 통해 서로 다른 학습데이터가 발생될 수 있다. 앙상블 학습을 구성할 수 있는 또 다른 방법 중의 하나는 분류기 모형의 변화를 주는 것이다. 이 경우 앙상블은 서로 다른 형태의 분류기로 구성되거나, 같은 유형의 모형이지만 서로 다른 파라미터 셋팅에 의해 변화를 준 분류기로 앙상블을 구성한다. 이 모든 경우 동일한 학습데이터를 통해 모형이 구축된다. 끝으로 위의 앙상블 구성방법의 조합을 통한 앙상블 구성이 있을 수 있다.

크기가 N인 학습데이터 T가 있다고 가정하면, bagging은 T로부터 크기 N'인 (N>N')표본을 랜덤하게 복원추출을 통해 새로 생성하게 된다. 이렇게 K번 반복하게 되면 K개의 새로운 학습데이터를 얻게 되며 K개의 서로 다른 학습데이터를 이용해 기저 분류기를 학습시키면 K개의 분류 모형이 나오게 된다. 이렇게 학습된 분류 모형을 다양한 방법에 의해 통합을 하게 된다. 앙상블 학습을 구성한 후 각각 모형의 결과를 종합하는 여러 가지 방법이 있으나 본 연구에서는 majority voting 방식을 채택하여 사용하였다.

Random subspace 기법은 [28]에 의해 제안된 앙상블 구축 기법이다. 이것은 원래의 feature set F에서 랜덤하게 일정 비율의 속성을 선택하여 학습데이터를 구축한다. 이렇게 K번 반복하게 되면 K개의 새로운 학습데이터를 얻게 된다. 기저 분류기 모형은 이들 K개의 각각의 학습데이터에서 학습되며 이를 통해 K개의 서로 다른 학습 모형이 생성되고 이들 서로 다른 학습모형을 다양한 방법에 의해 통합하는 방법이다.

Random subspace 방법은 데이터의 입력변수가 중복적인 것이 많거나 불필요 한 것이 많을 때 좋은 분류기를 선택할 수 있다는 장점이 있다. 또한, 학습데이터의 수가 데이터의 차원(dimensionality)에 비해 매우 적을 때 random subspace를 통해 차원을 줄임으로써 실험 데이터의 수가 매우 적을 때 발생할 수 있는 문제(small size problem)도 해결할 수 있는 것으로 알려져 있다.

3. 연구 모형

본 연구는 부도 예측 모형의 성능개선에 관한 연구이다. 이를 위해 본 연구에서는 단일 모형으로 그 우수성을 인정받고 있는 SVM의 앙상블 모형에 대해 고찰하고자 한다. 본 연구에서는 SVM 모형의 성능 개선을 위해 여러 가지 앙상블 학습 방법 중에서 가장 대표적인 bagging과 random subspace 방법을 부도 예측 모형에 적용해 보았다. 또한 bagging과 random subspace 모형의 성과 개선을 위해 새로운 통합 앙상블 모형을 제안하고 제안한 모형의 성과를 검증하기 위해 실제 기업의 부도 예측 데이터를 사용하여 검증해보았다.

본 연구에서 부도 예측을 위해 사용한 SVM 모형은 총 다섯 가지 모형이다. 다양한 SVM 앙상블 모형의 성과 개선을 검증하기 위해 단일 SVM 모형을 기본 모형으로 사용하여 실험하였다. 그리고, 단일 SVM의 성과 개선을 위해 총 네 가지 형태의 SVM 앙상블 모형을 부도 예측을 위해 사용하였다. 우선 첫 번째 앙상블 모형은 가장 대표적인 앙상블 형태인 bagging 모형이다. 두 번째 앙상블 모형은 random subspace 모형이다. 본 연구에서는 단순 앙상블 모형인 bagging과 random subspace의 통합 모형을 제안하였다. 우선 Hybrid 1(Bagging+RS) 모형은 bagging과 random subspace를 각각 독립적으로 수행한 후 통합하는 모형이다. Bagging 방식에 의해 K개의 서로 다른 학습데이터를 생성하고, 이들 각각으로부터 K개의 SVM 학습 모형을 생성시킨다. 그리고, random subspace 방식에 의해서도 feature set이 서로 다른 새로운 학습데

이터 K개를 생성하고, 이들 각각으부터 또 다른 K개의 SVM 학습 모형을 생성시킨다. 그리고, 이들 2K개의 모형을 검증용 데이터에 적용하여 결과값을 voting 방식에 의해 통합한다.

Hybrid 2(Bagging+RS) 모형은 bagging과 random subspace를 동시에 진행하는 모형이다. 우선 bagging 방식에 의해 학습용 데이터로부터 크기 N'인 표본을 랜덤하게 복원추출 방법에 의해 추출하고, 이렇게 생성된 새로운 학습데이터(Bootstrap sample)로부터 random subspace 방식에 따라 랜덤하게 크기 f인 feature set을 선택하여 최종적으로 학습할 데이터를 생성하게 된다. 이와 같은 방식을 K번 반복하여 bagging과 random space를 동시에 수행한 새로운 형태의 학습데이터 K개를 생성하고, 이들 각각의 학습데이터를 사용해 SVM 모형을 학습시켜 K개의 SVM 모형을 생성시킨다. 그리고, 검증용 데이터를 K개의 SVM 모형에 적용해 결과값을 도출하고 이것들을 voting 방식에 의해 통합한다. 부도 예측을 위해 본 연구에서 사용한 앙상블 모형의 전반적인 절차에 대한 설명은 다음과 같다.

[1] SVM Bagging Ensemble 모형(SVM-Bagging)의 절차

1. 데이터 분할 (학습용 데이터 T, 검증용 데이터 V)
2. 학습용 데이터(T)에서 크기 N'(<N: N은 T의 크기)인 표본을 랜덤하게 복원추출 방법을 통해 생성
3. 위의 2를 K번 반복하여 K개의 새로운 학습데이터(Bootstrap Sample) 생성
   ==> $T(B)_1, T(B)_2, \cdots, T(B)_k$
4. K개의 새로운 학습데이터 각각에 대해 SVM 모형 학습(K개의 SVM 학습 모형 생성됨)
   ==> $SVM_1, \cdots, SVM_K$
5. 검증용 데이터 V에 대해 위의 4에서 생성된 SVM 모형 적용 (K개의 결과값 생성)
   ==> $O_1, \cdots, O_K$
6. 결과값($O_1, \cdots, O_K$)을 다양한 통합 전략에 의해 통합 (본 연구에서는 majority voting 방식 채택)

[2] SVM Random Space Ensemble 모형(SVM-RS)의 절차

1. 데이터 분할 (학습용 데이터 T, 검증용 데이터 V)
2. 학습용 데이터 T의 총 가능한 feature의 수를 F라고 할 때, 학습용 데이터(T)에서 크기 f (<F)인 feature를 랜덤하게 선택하여 새로운 표본(Random Subspace Set)을 생성
3. 위의 2를 K번 반복하여 K개의 새로운 학습데이터 생성 ==> $T(RS)_1, T(RS)_2, \cdots, T(RS)_k$
4. K개의 새로운 학습데이터 각각에 대해 SVM 모형 학습(K개의 SVM 학습 모형 생성됨)
   ==> $SVM_1, \cdots, SVM_K$
5. 검증용 데이터 V에 대해 생성된 SVM 모형 적용 (K개의 결과값 생성) ==> $O_1, \cdots, O_K$
6. 결과값($O_1, \cdots, O_K$)을 다양한 통합 전략에 의해 통합 (본 연구에서는 majority voting 방식 채택)

[3] Hybrid1(Bagging + Random Space) Ensemble 모형(Hybrid1(Bagging+RS))의 절차

1. 데이터 분할 (학습용 데이터 T, 검증용 데이터 V)
2. 학습용 데이터(T)에서 크기 N'(<N: N은 T의 크기)인 표본을 랜덤하게 복원추출 방법을 통해 생성
3. 위의 2를 K번 반복하여 K개의 새로운 학습데이터(Bootstrap Sample) 생성
   ==> $T(B)_1, T(B)_2, \cdots, T(B)_k$
4. K개의 새로운 학습데이터 각각에 대해 SVM 모형 학습(K개의 SVM 학습 모형 생성됨)
   ==> $SVM_1, \cdots, SVM_K$
5. 총 가능한 feature의 수를 F라고 할 때, 학습용 데이터(T)에서 크기 f (<F)인 feature를 랜덤하게 선택하여 새로운 표본(Random Subspace Set)을 생성
6. 위의 5를 K번 반복하여 K개의 새로운 학습데이터 생성 ==> $T(RS)_1, T(RS)_2, \cdots, T(RS)_k$
7. K개의 새로운 학습데이터 각각에 대해 SVM 모형 학습(K개의 SVM 학습 모형 생성됨)
   ==> $SVM_{K+1}, \cdots, SVM_{2K}$
8. 검증용 데이터 V에 대해 4와 7에서 생성된 SVM 모형 적용 (총 2K개의 결과값 생성)
   ==> $O_1, \cdots, O_{2K}$
9. 결과값($O_1, \cdots, O_{2K}$)을 다양한 통합 전략에 의해 통합 (본 연구에서는 majority voting 방식 채택)

[4] Hybrid 2(Bagging + Random Space) Ensemble 모형(Hybrid2(Bagging+RS))의 절차

1. 데이터 분할 (학습용 데이터 T, 검증용 데이터 V)
2. 학습용 데이터(T)에서 크기 N'(<N: N은 T의 크기)인 표본을 랜덤하게 복원추출 방법을 통해 생성
3. 위의 2에서 생성된 새로운 학습 데이터 Bootstrap Sample T(B)i에서 크기 f (<F: F는 T의 총 feature의 수)인 feature 를 랜덤하게 선택하여 새로운 표본을 생성 ==> T(B+RS)i
4. 위의 2, 3을 K번 반복하여 K개의 새로운 학습 데이터 생성 ==> T(B+RS)1, … ,T(B+RS)K
5. K개의 새로운 학습데이터 각각에 대해 SVM 모형 학습(K개의 SVM 학습 모형 생성됨) ==> SVM1, …, SVMK
6. 검증용 데이터 V에 대해 생성된 SVM 모형 적용 (K개의 결과값 생성) ==> O1, …, OK
7. 결과값(O1,…,OK)을 다양한 통합 전략에 의해 통합 (본 연구에서는 majority voting 방식 채택)

4. 실험설계

본 연구에서 제안한 모형의 우수성을 검증하기 위해 국내 비외감 기업의 데이터를 이용해 실험을 수행하였다. 실험에 사용된 데이터는 자산규모가 10억에서 70억 사이인 기업으로 산업분야가 중공업인 기업의 데이터로 구성되어 있다. 실험에 사용한 데이터는 총 1218 개로 구성되어 있으며 이중 609 개의 데이터가 부도 기업 데이터 609 개의 데이터가 비부도 기업의 데이터로 이루어져 있다. 데이터는 1999년부터 2002년 사의의 기업 재무데이터를 사용하였다.

데이터는 학습용 데이터(training set), 검증용 데이터(validation set)로 나누어 실험을 하였다. 학습용 데이터는 모형의 학습을 위한 데이터로 사용되었으며, 모형의 일반화 정도 측정 및 비교 검증을 위해 검증용 데이터를 사용하였다. 본 연구에서는 10-겹 검증(10-fold cross validation) 방법으로 실험을 하였으며 10-겹 검증 실험을 10회 수행하여 모형의 성과를 비교 평가하였다. 실험결과는 모든 실험 결과의 평균값을 계산하여 대표값으로 사용하였다. 본 연구에서 사용된 변수는 안정성, 수익성, 성장성, 활동 및 현금 흐름으로 분류된 131 개의 재무 비율을 사용하였다. 입력변수 선정을 위해 1차로 131개의 변수를 대상으로 t-test를 실시하였다. 그리고, 1차 선정된 변수를 대상으로 stepwise method를 이용한 로짓모형 방식과 선행 연구결과 등을 종합적으로 고려해 최종 변수를 <표 1>과 <표 2>와 같이 선정하였다. <표 1>에 나온 변수는 기본 SVM 모형과 bagging 모형을 위해 사용되었다. 한편, random subspace 모형은 여러 입력변수 (feature set) 중에서 일부를 랜덤하게 추출하는 것을 반복하여 서로 다른 훈련데이터를 만드는 방식으로 이를 위해 <표 2>의 25개의 입력변수를 사용하였다. 즉, random subspace 모형과 bagging과 random subspace 통합 모형을 위해 <표 2>의 입력변수를 사용하였다.

<표 1> Bagging 실험에 사용된 입력변수

변수	변수명	변수	변수명
X1	금융비용대부채비율	X6	차입금의존도
X2	현금비율	X7	매출액대비운전자금변동율
X3	금융비용대매출액	X8	총자본순이익률
X4	매출액변동계수	X9	매입채무회전기간
X5	유보액대총자산	X10	순운전자본비율

<표 2> Random Subspace, Hybrid 모형 실험에 사용된 입력변수

변수	변수명	변수	변수명
X1	금융비용대부채비율	X14	총현금흐름대부채비율
X2	현금비율	X15	자본금회전율
X3	금융비용대매출액	X16	변동비대매출액
X4	매출액변동계수	X17	경영자산회전율
X5	유보액대총자산	X18	EBIT대이자비용
X6	차입금의존도	X19	고정자산회전율
X7	매출액대비운전자금변동율	X20	금융비용대총비용
X8	총자본순이익률	X21	자본금순이익율
X9	매입채무회전기간	X22	고정비율
X10	순운전자본비율	X23	EBITDA대매출액
X11	당좌비율	X24	현금흐름단기차입금
X12	차입금대매출액	X25	재료비/매출액
X13	총자본경상이익율		

SVM 모형은 사용하는 커널에 따라 그 성과 및 결과값이 많이 달라지는 경향이 있다. 본 연구에서는 SVM의 성과 개선에 대한 연구로 SVM의 커널과 앙상블 모형의 성과와의 관계를 살펴보기 위해 각각의 커널 별로 다양한 실험을 실시하였다. 본 연구에서는 linear, polynomial, rbf 커널을 각각 사용하여 비교 분석하였다. SVM 모형의 성과는 사용하는 커널과 커널의 파라미터에 따라 큰 차이를 보이는데 본 연구에서는 여러 실험을 통해 가장 성과가 좋은 파라미터 값을 사용했다. 앙상블 모형의 경우도 파라미터에 따라 그 성과가 크게 달라지는데, bagging 모형의 경우 bootstrap의 크기(N') 와 모형의 수(K)에 따라 성과가 크게 차이가 나며 random subspace의 경우에도 선택된 입력변수의 수(f)와 모형의 수(K)에 따라 성과가 크게 달라진다. 본 연구에서는 실험의 복잡성을 줄이기 위해 여러 예비실험을 통해 성과가 좋은 파라미터 값을 대표값으로 사용하여 실험하였다.

5. 실험결과

본 연구에서는 부도 예측 모형의 성과 개선을 위하여 단일 모형으로 성과가 우수한 것으로 알려져 있는 SVM을 기저 분류기로 사용하는 다양한 앙상블 모형을 제안하였다. SVM 모형의 성능 개선을 위해 bagging과 random subspace 앙상블 모형을 부도 예측 모형에 적용해 보았다. 또한 bagging 모형과 random subspace 모형의 성과 개선을 위해 새로운 앙상블 모형을 제안하고 제안한 모형의 성과를 검증하기 위해 실제 기업의 부도 예측 데이터를 사용하였다. 각 모형별 실험 결과는 <표 3>과 <그림 1>과 같다. 여기서 Simple SVM은 여러 분류기를 통합한 앙상블이 아닌 단일 모형을 의미한다. SVM-Bagging은 bagging 방법을 통한 앙상블 모형을, SVM-RS는 random subspace 방법을 이용한 앙상블 모형을 의미한다. Hybrid1(Bagging+RS)와 Hybrid2(Bagging+RS)는 bagging과 random subspace의 통합 모형을 의미한다. Hybrid1(Bagging+RS) 모형은 bagging과 random subspace를 각각 독립적으로 수행한 후 통합하는 모형을 Hybrid 2(Bagging+RS) 모형은 bagging과 random subspace를 동시에 진행하는 모형을 의미한다. 본 연구에서는 SVM의 다양한 커널(linear, polynomial, rbf 커널)을 사용해 각각 모형의 성과를 비교하였다.

<표 3> 각 모형의 예측 정확도(%)

커널	모형	예측률
Linear	Simple SVM	71.35
	SVM-Bagging	72.43
	SVM-RS	73.32
	Hybrid 1 (Bagging+RS)	74.14
	Hybrid 2 (Bagging+RS)	74.52
Rbf	Simple SVM	72.42
	SVM-Bagging	71.62
	SVM-RS	73.52
	Hybrid 1 (Bagging+RS)	72.87
	Hybrid 2 (Bagging+RS)	74.13
Polynomial	Simple SVM	71.89
	SVM-Bagging	74.05
	SVM-RS	74.35
	Hybrid 1 (Bagging+RS)	74.65
	Hybrid 2 (Bagging+RS)	75.76

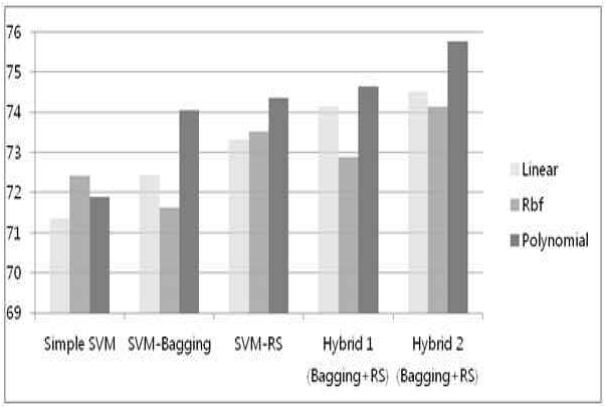
실험 결과 단일 SVM 모형에서는 rbf 커널을 사용한 SVM이 가장 좋은 결과를 보였다. 실험 결과 전반적으로 단일 모형보다는 앙상블 모형이 좋은 성과를 보임을 알 수 있었다.

<표 4>는 각 모형의 평균 예측률을 보여주고 있다. 이를 통해 커널별 비교가 아닌 각 모형의 전반적인 비교가 가능하다. <그림 2>에 보는 바와 같이 단일 모형보다는 앙상블 모형이 전반적으로 우수한 성과를 보이는 것을 알 수 있다. 각각을 살펴보면 bagging을 이용한 앙상블 모형보다는 random subspace를 이용한 앙상블 모형이 우수한 성과를 보였으며, 단일 앙상블 전략보다는 hybrid 앙상블 전략을 사용한 모형이

<표 4> 모형별 평균 예측 정확도(%)

모형	평균 예측률
Simple SVM	71.89
SVM-Bagging	72.7
SVM-RS	73.73
Hybrid 1 (Bagging+RS)	73.88
Hybrid 2 (Bagging+RS)	74.8

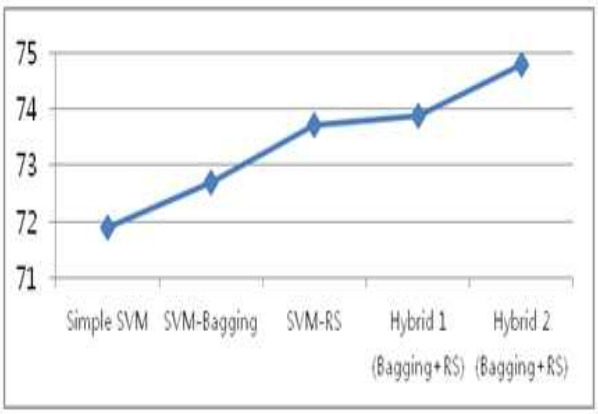
더 우수한 결과를 보였다. 전체적으로는 bagging과 random subspace를 동시에 진행한 Hybrid 2(Bagging+RS) 모형이 가장 우수한 성과를 보임을 알 수 있다.



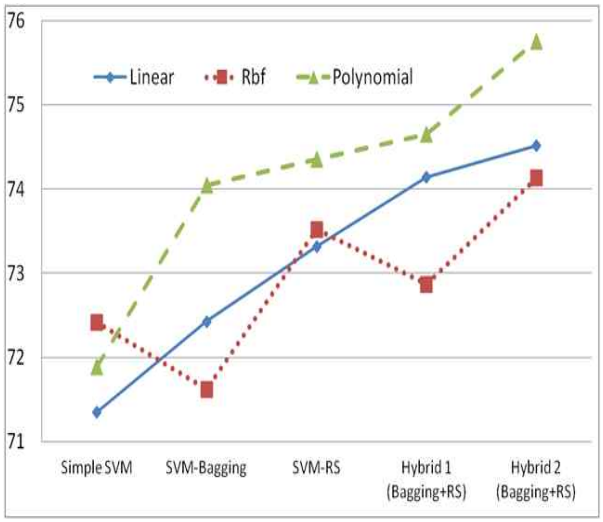
<그림 1> 모형별 예측 성과 비교(%)

<그림 3>은 커널별 예측 성과 비교를 나타내 주고 있다. <그림 3>을 보면 Linear 커널과 polynomial 커널을 사용할 경우 모형별 성과가 <표 4>와 <그림 2>에 나타난 전반적 성과와 일치하는 모습을 보임을 알 수 있으며 rbf 커널을 사용할 경우 다른 양상을 보임을 쉽게 알 수 있다. <그림 4>는 각 모형별 성과 비교를 나타내 주고 있다. <그림 4>의 A에서 보는 바와 같이 전체적으로는 Polynomial 커널을 사용하는 Hybrid2(Bagging+RS) 모형의 성과가 가장 좋은 것을 알 수 있다.

<그림 4>의 A, B에서 알 수 있듯이 Polynomial 커널을 사용할 경우 앙상블 학습으로 인한 성과 개선이

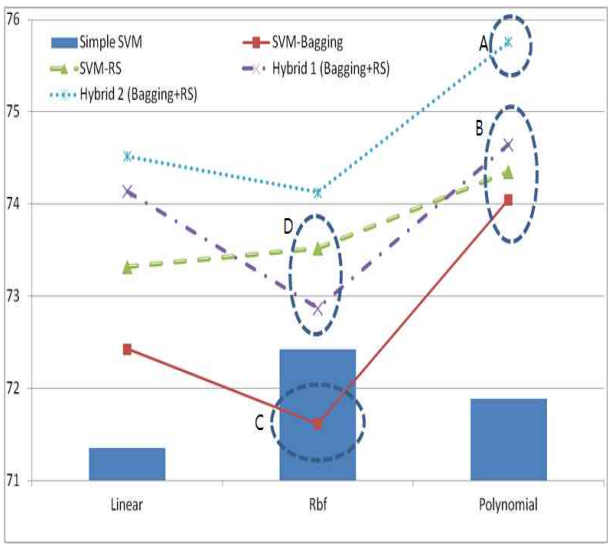


<그림 2> 모형별 평균 예측 정확도(%)



<그림 3> 커널별 예측성과 비교(%)

가장 크며, 다른 커널을 사용한 앙상블 모형보다 성과가 좋다는 것을 알 수 있다. <그림 4>의 C는 앙상블 모형이 단일 모형보다 성과가 낮은 경우를 잘 보여주고 있다. Rbf 커널을 사용하는 bagging 앙상블 모형의 경우 단일 모형보다 성과가 더 좋지 않음을 알 수 있다. D에서 보는 바와 같이 다른 커널을 사용할 경우와 달리 rbf커널을 사용할 경우는 Hybrid1 (bagging+RS) 모형보다 SVM-RS 모형의 성과가 더 좋은 결과를 보였다. 전반적으로 rbf 커널을 사용하는 SVM 앙상블 모형의 성과는 다른 커널을 사용할 때보다 성과 개선의 폭이 가장 작음을 알 수 있다.



<그림 4> 모형별 성과 비교(%)

6. 결 론

분류기의 앙상블 학습은 여러 개의 서로 다른 분류기들의 조합을 통해 만들어진다. 앙상블 학습의 구축은 기계학습(machine learning) 분야에서 많은 관심을 끌고 있는 중요한 연구주제이며 대부분의 경우에 있어서 이들 앙상블 모형은 개별 기저 분류기보다 더 좋은 성과를 내는 것으로 알려져 있다. 앙상블 모형의 성과가 개별 분류기들의 성과보다 더 좋아지게 하기 위해서는 기저 분류기들의 성과가 좋아야 하며 또한 기저 분류기들의 예측 결과값이 가능한 한 다양성을 가져야 한다. 즉 앙상블을 구성하는 기저 분류기들의 모형에 의한 예측 결과값이 가능하면 서로 다른 결과값을 가져야 한다. 기저 분류기들의 다양성을 확보하기 위한 대표적인 방법으로는 bagging과 random subpace 방법이 있으며 이들은 각기 다른 훈련 데이터 사용하거나 서로 다른 속성(attribute subsets) 사용함으로써 기저 분류기의 다양성을 확보하고 있다.

Bagging은 가장 대표적인 통합모형 방법으로 많은 선행 연구에서 단일 모형보다 더 좋은 성과를 내는 것으로 알려져 왔다. 하지만, 의사결정 트리(Decision Tree)나, 인공신경망(Neural Network)의 bagging을 이용한 앙상블 모형에 대한 연구는 많았지만 성과가 우수한 SVM을 기저 분류기로 하는 bagging 앙상블에 대한 연구는 상대적으로 많지 않았다.

본 논문은 부도 예측 모형의 성능개선에 관한 연구이다. 이를 위해 본 연구에서는 단일 모형으로 그 우수성을 인정받고 있는 SVM을 기저 분류기로 사용하는 앙상블 모형에 대해 고찰하였다. SVM 모형의 성능개선을 위해 bagging과 random subspace 모형을 부도 예측 모형에 적용해 보았으며 bagging 모형과 random subspace 모형의 성과 개선을 위해 새로운 앙상블 모형을 제안하고 제안한 모형의 성과를 검증하기 위해 실제 기업의 부도 예측 데이터를 사용하여 실험하였다. 실험 결과 본 논문에서 제안한 새로운 형태의 통합 모형이 가장 좋은 성과를 보임을 알 수 있었다.

본 연구의 한계와 향후 연구 방향을 정리하면 다음과 같다. 우선 본 연구에서 제안한 모형의 우수성을 검증하기 위해서는 보다 다양한 데이터에서의 검증이 필요할 것으로 보인다. 또한 앙상블 모형의 성과는 파라미터의 값에 따라 그 성과가 크게 차이가 있으므로, 파라미터의 영향을 통제하기 위한 보다 다양한 실험이 추가로 필요할 것으로 여겨진다. 본 연구에서 제안한 모형은 부도 예측 문제가 아닌 다른 예측 문제에도 적용 가능할 것이다. 이에 대한 검증을 위해 추가적인 연구가 필요할 것으로 여겨진다.

참 고 문 헌

[1] Beaver, W, “Financial ratios as predictors of failure, empirical research in accounting: Selected studied”, Journal of Accounting Research, 1966, pp.71-111.

[2] Altman, E. L, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, The Journal of Finance, 23(3), 1968, pp.589-609.

[3] Altman, E. L., Edward, I., Haldeman, R., & Narayanan, P. A, “new model to identify bankruptcy risk of corporations”, Journal of Banking and Finance, 1, 1977, pp.29-54.

[4] Meyer, P. A., & Pifer, H, “Prediction of bank failures”, The Journal of Finance, 25, 1970, pp.853-868.

[5] Dimitras, A. I., Zanakis, S. H., & Zopounidis, C, “A survey of business failure with an emphasis on prediction methods and industrial applications”, European Journal of Operational Research, 90(3), 1996 , pp.487-513.

[6] Ohlson, J, “Financial ratios and the probabilistic prediction of bankruptcy”, Journal of Accounting Research, 18(1),1980 , pp.109-131.

[7] Pantalone, C., & Platt, M. B, “Predicting commercial bank failure since deregulation”, New England Economic Review, 1987, pp.37-47.

[8] Han, I., Chandler, J. S., & Liang, T. P, “The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods”, Expert System with Applications, 10(2), 1996, pp.209-221.

[9] Shaw, M., & Gentry, J, “Using and expert system with inductive learning to evaluate business loans”, Financial Management, 17(3),

1998, pp.45-56.
[10] Buta, P, “Mining for financial knowledge with CBR”, AI Expert, 9(10), 1994, pp.34-41.
[11] Bryant, S. M, “A case-based reasoning approach to bankruptcy prediction modeling”, International Journal of Intelligent Systems in Accounting, Finance and Management, 6(3), 1997, pp.195-214.
[12] Bortiz, J. E., & Kennedy, D. B, “Effectiveness of neural network types for prediction of business failure”, Expert Systems with Application, 9(4), 1995, pp.503-512.
[13] Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C, “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis”, European Journal of Operational Research, 116(1), 1999 , pp.16-32.
[14] Coakley, J. R., & Brown, C. E, “Artificial neural networks in accounting and finance: Modeling issues”, International Journal of Intelligent Systems in Accounting, Finance and Management, 9(2), 2000, pp.119-144.
[15] Fan, A., & Palaniswami, M, “Selecting bankruptcy predictors using a support vector machine approach”, Proceeding of the international joint conference on neural network, Vol. 6, 2000, pp. 354-359.
[16] Van Gestel, T., Baesens, B., Suykens, J., Espinoza, M. Baestaens, D.-E., Vanthienen, J., et al. “Bankruptcy prediction with least squares support vector machine classifiers, computational intelligence for financial engineering”, 2003, proceeding 2003. IEEEE international conference on 2003, pp.1-8.
[17] Min, S.,& Lee, J., “Hybrid genetic algorithms and support vector machines for bankruptcy prediction”, Expert Systems with Applications, Volume 31, Issue 3, October 2006, pp.652-660.
[18] 신택수, 홍태호, “AdaBoost 알고리즘 기반 SVM을이용한 부실 확률분포 기반의 기업신용평가”, 지능정보연구, 17권 3호(2011), pp.25~41.
[19] 김명종, “유전자 알고리즘을 이용한 분류기 앙상블의 최적 선택”, 지능정보연구 제16권 제4호 2010, pp. 99~112.
[20] 김승혁, 김종우, “Modified Bagging Predictors를 이용한 SOHO 부도 예측”, 한국지능정보시스템학회논문지 제13권 제2호, 2007, pp.15~26.
[21] Dietterich, T. G, “Machine-learning research: Four current directions”, AI Magazine, 18(4), 1997, pp.97-136.
[22] Kuncheva L.I, “Combining classifiers: Soft computing solutions”, in: S.K. Pal and A. Pal (Eds.) Pattern Recognition: From Classical to Modern Approaches, World Scientific Publishing Co., Singapore, 2001, 427-452
[23] Breiman, L, “Bagging predictors”, Machine Learning, 24(2), 1996, pp.123-140.
[24] Vapnik, V. N, “The nature of statistical learning theory”, New York: Springer, 1995.
[25] 박창석, 김병만, 서병훈, 김준우,이광호,"이동 차량에서의 실시간 자동차 번호판 인식",한국산업정보학회논문지, v.9, no.2, 2004년, pp.32-43
[26] 원철호, 이상헌, 이태균,"인터랙티브 TV 컨트롤 시스템을 위한 근적외선 영상의 얼굴 인식", 한국산업정보학회논문지, v.15, no.5, 2010년, pp.11-17
[27] 유혜경, 이진영, 나종화,"매장문화재 예측을 위한 통계적 분류 분석", 한국산업정보학회논문지, v.14, no.3, 2009년, pp.106-113
[28] Ho, T. K, “The random subspace method for constructing decision forests”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 1998, pp.832-844.

민 성 환 (Sung-Hwan Min Hong)

• 동국대학교 산업공학과 공학사
• 고려대학교 산업공학과 공학석사
• KAIST 경영공학 공학박사
• 한림대학교 경영학부 부교수
• 관심분야: 데이터마이닝, e-비지니스, 고객관계관리