ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

# 부트스트랩 샘플링 최적화를 통한 앙상블 모형의 성능 개선

# Improving an Ensemble Model by Optimizing Bootstrap Sampling

민 성 환[1*]
Sung-Hwan Min

요 약

앙상블 학습 기법은 개별 모형보다 더 좋은 예측 성과를 얻기 위해 다수의 분류기를 결합하는 것으로 예측 성과를 향상시키는 데에 매우 유용한 것으로 알려져 있다. 배깅은 단일 분류기의 예측 성과를 향상시키는 대표적인 앙상블 기법중의 하나이다. 배깅은 원 학습 데이터로부터 부트스트랩 샘플링 방법을 통해 서로 다른 학습 데이터를 추출하고, 각각의 부트스트랩 샘플에 대해 학습 알고리즘을 적용하여 서로 다른 다수의 기저 분류기들을 생성시키게 되며, 최종적으로 서로 다른 분류기로부터 나온 결과를 결합하게 된다. 배깅에서 부트스트랩 샘플은 원 학습 데이터로부터 랜덤하게 추출한 샘플로 각각의 부트스트랩 샘플이 동일한 정보를 가지고 있지는 않으며 이로 인해 배깅 모형의 성과는 편차가 발생하게 된다. 본 논문에서는 이와 같은 부트스트랩 샘플을 최적화함으로써 표준 배깅 앙상블의 성과를 개선시키는 새로운 방법을 제안하였다. 제안한 모형에서는 앙상블 모형의 성과를 개선시키기 위해 부트스트랩 샘플링을 최적화하였으며 이를 위해 유전자 알고리즘이 활용되었다. 본 논문에서는 제안한 모형을 국내 부도 예측 문제에 적용해 보았으며, 실험 결과 제안한 모형이 우수한 성과를 보였다.

☞ 주제어 : 배깅, 부도 예측, 앙상블, 유전자 알고리즘

ABSTRACT

Ensemble classification involves combining multiple classifiers to obtain more accurate predictions than those obtained using individual models. Ensemble learning techniques are known to be very useful for improving prediction accuracy. Bagging is one of the most popular ensemble learning techniques. Bagging has been known to be successful in increasing the accuracy of prediction of the individual classifiers. Bagging draws bootstrap samples from the training sample, applies the classifier to each bootstrap sample, and then combines the predictions of these classifiers to get the final classification result. Bootstrap samples are simple random samples selected from the original training data, so not all bootstrap samples are equally informative, due to the randomness. In this study, we proposed a new method for improving the performance of the standard bagging ensemble by optimizing bootstrap samples. A genetic algorithm is used to optimize bootstrap samples of the ensemble for improving prediction accuracy of the ensemble model. The proposed model is applied to a bankruptcy prediction problem using a real dataset from Korean companies. The experimental results showed the effectiveness of the proposed model.

☞ keyword : Bagging, Bankruptcy Prediction, Ensemble, Genetic Algorithms

## 1. 서 론

앙상블 분류기는 개별적으로 학습된 일련의 분류기로 구성되며 이들 각각의 분류기로부터 나온 예측 결과의 결합을 통해 최종 앙상블의 예측 결과를 얻게 된다. 일반적으로 앙상블 분류기를 잘 구성할 경우 하나의 분류기를 사용할 경우 보다 더 좋은 예측 성과를 내는 것으로 알려져 있다[1,2].

앙상블 분류기가 좋은 예측 성과를 내기 위해서는 앙상블을 구성하고 있는 기저 분류기들의 예측 성과가 가능하면 좋아야 하며 또한 기저 분류기들의 예측 오차(prediction error)가 가능하면 서로 다를수록 좋은 것으로 알려져 있다. 특히, 개별 분류기들의 다양성은 앙상블의 예측 성과에 매우 중요한 요소로 알려져 있다[3,4].

앙상블 분류기의 대표적인 기법은 기저 분류기를 다양화 시키는 방식과 관련이 있다. 앙상블을 구성하게 될 기저 분류기를 다양화 시킴으로써 앙상블 모형의 성능 개선을 기대할 수 있기 때문이다. 기저 분류기를 다양화 시키는 방법으로는 학습 데이터의 다양화, 학습 알고리즘의 다양화, 학습 알고리즘 파라미터의 다양화 등의 방법이 있다. 이중에서 가장 대표적인 기법은 학습 데이터

[1] Department of Business Administration, Hallym University, Chuncheon Gangwon-do, 200-702, Korea
* Corresponding author (shmin@hallym.ac.kr)
[Received 16 February 2016, Reviewed 25 February 2016, Accepted 23 March 2016]

의 다양화를 통한 기저 분류기의 다양화 기법으로 배깅(bagging)[5], 부스팅(boosting)[6]과 랜덤 서브스페이스 기법[7] 등이 대표적인 예이다.

배깅은 단일 모형보다 더 좋은 성능을 내기 위해 여러 모형을 결합하는 앙상블 기법 중의 하나로 비교적 단순하면서도 성능이 좋아 많은 응용 분야에서 활용도가 높다. Breiman[5]에 의해 제안된 배깅 기법은 기저 분류기를 다양화시키기 위해 원 학습 데이터(original training data)로부터 부트스트랩 샘플링(bootstrap sampling) 방법을 통해 서로 다른 학습 데이터를 추출하고, 각각의 부트스트랩 샘플에 대해 학습 알고리즘을 적용하여 서로 다른 다수의 기저 분류기들을 생성시키게 되며, 최종적으로 서로 다른 분류기로부터 나온 결과를 다수결 투표(majority voting) 등과 같은 특정한 방식에 의해 결합하게 된다.

표준 배깅 방식에서는 서로 다른 학습 데이터를 추출하기 위해 복원 추출방식에 의해 랜덤 샘플링을 하게 되며, 이로 인해 어떤 데이터는 여러 번 선택되기도 하고, 어떤 데이터는 전혀 선택되지 않기도 한다. 이와 같이 랜덤하게 선택되는 학습 데이터로 인해 기저 분류기의 다양성이 발생하게 되며 이들을 결합할 경우 일반적으로 단일 분류기보다 성과가 개선되는 것으로 알려져 있다. 하지만 랜덤 샘플링으로 인한 임의성으로 인해 배깅 기법은 어떤 경우에는 성과가 매우 좋지만 어떤 경우는 그렇지 않은 경우도 존재하게 된다.

본 논문에서는 이와 같이 부트스트랩 샘플링으로 인해 배깅의 성과 편차가 생길 수 있다는 점에 착안하여 배깅 앙상블의 성과를 최고로 하는 부트스트랩 샘플링 방식을 찾기 위한 방법을 제안하였다. 본 논문에서는 성과를 최고로 하는 부트스트랩 샘플링을 위해 최적화 탐색 알고리즘으로 가장 대표적인 방법 중의 하나인 유전자 알고리즘(Genetic Algorithms)을 적용하였다. 즉, 수없이 많은 부트스트랩 샘플 집합 중에서 앙상블 모형 성과측면에서 최적(또는 근사 최적)인 부트스트랩 샘플을 선택하는 문제로 변환하여 유전자 알고리즘을 적용하였다. 본 논문에서 제안한 방법을 이용하면 앙상블 모형의 성과를 최적으로 하는 부트스트랩 샘플을 선택할 수 있을 것으로 기대되며 이를 통해 앙상블 모형의 성과 개선을 이룰 수 있을 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 2장에서는 앙상블 분류기에 대한 설명을 하고 3장에서는 본 논문에서 제안한 부트스트랩 샘플의 최적화를 통한 앙상블 모형의 성능 개선 방법에 대해 설명을 한다. 4장에서는 연구 데이터와 실험 설계에 대한 설명을 하며, 5장에서는 실험 결과와 이에 대한 분석 및 해석을 한다. 끝으로 6장에서는 연구의 결론 및 요약과 함께 향후 연구 과제에 대해 설명한다.

## 2. 앙상블 분류기

앙상블 분류기는 개별적으로 학습된 다수의 분류기로 구성되며 이들 각각의 분류기를 기저 분류기(base classifiers)라고 한다. 앙상블 분류기는 이들 기저 분류기 각각의 예측 결과를 결합함으로써 최종 앙상블의 예측 결과를 얻게 된다. 앙상블 분류기는 기저 분류기를 적절하게 구성할 경우 단일 분류기보다 성과가 좋은 것으로 알려져 있다. 이처럼 앙상블 분류기가 좋은 성과를 내기 위해서는 기저 분류기를 어떻게 구성하느냐가 중요한 영향을 미친다. 일반적으로 앙상블을 구성하고 있는 기저 분류기들의 예측 정확도가 좋을수록, 그리고 기저 분류기들의 다양성(diversity) 지수가 높을수록 앙상블의 성과가 좋은 것으로 알려져 있다. 기저 분류기들의 예측 결과값이 똑같은 결과값을 같는다면 기저 분류기들 간의 다양성이 존재하지 않는다고 말할 수 있으며, 반대의 경우는 다양성이 존재한다고 말할 수 있다.

예를 들어 모든 기저 분류기의 예측 오차를 p라고 하고, 모든 기저 분류기들이 독립적인 오차(independent errors)를 갖는다고 가정하자. 그러면, 앙상블을 구성하고 있는 기저 분류기의 총 수, N개 중에서 k개의 분류기가 오분류할 확률은 아래의 식 (1)과 같이 계산될 수 있다. 또한, 만약에 N개의 기저 분류자로 구성된 앙상블 분류기의 결합 방법을 가장 대표적인 방법인 다수결 투표방식을 사용한다고 하면, 앙상블 분류기의 예측 오차는 (2)와 같이 계산될 수 있다. 즉, 다수결 투표 방식으로 결합된 앙상블 분류기의 오차는 N개의 분류기 중에서 N/2개 이상의 기저 분류기가 동시에 오차를 보이는 확률과 같으며, 기저 분류기들의 오차가 모두 독립적이라고 가정했으므로 아래 식 (2)와 같이 계산될 수 있다. 만약 예측 오차 p가 0.5보다 작다면 (2)의 값은 N이 커짐에 따라 감소함을 알 수 있고 N이 무한대가 되면 0에 수렴하게 됨을 알 수 있다[8].

$$ {}_{N}C_{k} \times p^{k}(1-p)^{N-k} \tag{1} $$

$$ \sum_{k>N/2}^{N} {}_{N}C_{k} \times p^{k}(1-p)^{N-k} \tag{2} $$

즉, 이론적으로 기저 분류기의 예측 정확도가 0.5보다 크고, 서로 독립적이라면 앙상블의 성과는 크게 개선될 수 있다는 것을 알 수 있다. 그러나, 완벽하게 독립적인 기저 분류기를 구성하는 것은 현실적으로 불가능하다.

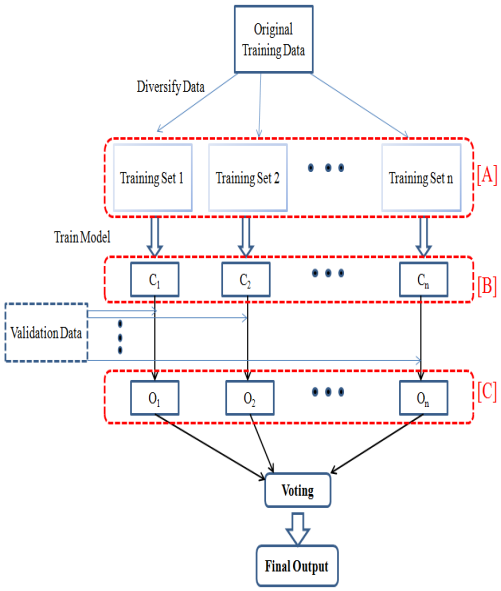
위에서 살펴본 바와 같이 앙상블의 성과 개선을 위해 기저 분류기들 간의 다양성이 매우 중요함을 알 수 있다. 즉, 앙상블 모형의 성과가 좋으려면 앙상블을 구성하고 있는 기저 분류기들이 서로 다양성을 가져야 한다. 기저 분류기들 간에 다양성을 갖는다는 것은 각각의 기저 분류기들의 예측 결과가 서로 다른 예측 결과를 낸다는 것을 의미하며 이를 통해 한 개의 분류기가 오분류를 하더라도 다른 분류기들을 결합함으로써 오분류를 줄일 수 있게 되는 것이다.

앙상블 분류기는 최근에 데이터 마이닝 분야에서 각광 받는 분야로 좋은 예측 성능으로 인해 다양한 분야의 예측 문제에 활발하게 적용되고 있다. 이와 같은 앙상블 분류기의 응용 관련 연구는 공학 분야에서 활발하게 진행되고 있지만 최근에는 재무 부실화 문제에도 앙상블 기법을 적용하려는 다양한 연구가 진행되고 있다.

Kim and Kim[9]은 가장 대표적인 앙상블 기법 중의 하나인 배깅 모형을 SOHO의 부도 예측 문제에 적용해 보았다. 또한, 기존의 배깅 모형의 성과 개선을 위해 기저 분류기 중에 예측 성과가 좋은 일부를 선택하여 앙상블을 구성하는 선택적 배깅 모형을 제안하여 의사결정 트리의 일종인 CART를 기저 분류기로 하여 실험을 수행하였다. Tsai and Wu[10] 는 인공 신경망을 기저 분류기로 하는 다양한 형태의 앙상블 모형을 재무 부실화 문제에 적용해 보았다. 이들은 다양한 파라미터에서 가장 좋은 결과를 보인 다수의 인공 신경망 분류기를 다수결 투표 방식에 의해 결합하는 앙상블 모형을 단일 모형 중에 가장 좋은 결과를 보인 모형과 비교해 보았다. 또한, 파라미터의 다양화뿐만 아니라 학습 데이터도 다양화하여 개별 분류기를 학습시킨 앙상블 모형도 이들과 같이 비교하였다. 여러 모형들을 다양한 재무 부실화 관련 데이터를 이용해 분석한  결과 앙상블 모형이 항상 단일 최고 모형보다 좋은 결과를 보이지는 않았다. Kim [11]은 의사결정 트리, 인공 신경망 모형, SVM(Support Vector Machines) 모형을 기저 분류기로 하는 배깅과 부스팅 앙상블 모형을 기업의 부도 예측 문제에 적용해 보았다. 실험결과 의사결정 트리와 인공 신경망 모형을 기저 분류기로 사용할 경우에는 앙상블 모형이 단일 모형보다 통계적으로 유의한 성과 개선이 있었지만, SVM을 기저 분류기로 사용할 경우에는 앙상블 학습을 통한 유의적인 성과 개선이 나타나지 않았다. Li et al. [12]는 이진 로짓 모형을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형을 중국 기업의 부도 예측 관련 데이터에 적용해 판별 분석, 로지스틱 회귀 분석, 프로빗  모형 등과 같은 전통적인 통계 모형과 비교해 보았다. 실험 결과 랜덤 서브스페이스 앙상블 모형이 전통적인 단일 모형보다 좋은 성과를 보였다. Min [13]은 SVM을 기저 분류기로 하는 배깅과 랜덤서브스페이스의 통합 모형을 국내 부도 예측 문제에 적용해 보았으며, 통합 앙상블 모형이 단일 앙상블 모형 보다 좋은 성과를 보임을 알 수 있었다. Marqués et al. [14] 는 의사결정 나무(Decision tree)를 기저 분류기로 하는 다양한 통합 모형을 신용 평가 문제에 적용해 보았으며, 실험 결과 배깅과 랜덤 포리스트(Random forest)를 결합한 모형이 좋은 성과를 보임을 알 수 있었다. Choi and Lim[15]은 커널 함수 다양화를 통해 서로 다른 형태의 SVM 분류기들을 생성하여 앙상블 모형을 구성하였다. 제안한 모형은 부도 예측 문제에서 단일 모형보다 높은 예측 성과를 보였다. Min [16]은 원 데이터에서 불필요한 데이터, 관련 없는 데이터를 제거하는 사례 선택 기법을 활용하여 배깅의 성능을 개선하는 새로운 형태의 모형을 제안했으며, 국내 기업의 부도 예측 관련 문제에 적용해 보았다. SVM을 기저 분류기로 하여 다양한 실험을 수행하였으며 제안한 방법이 기존의 배깅 모형의 성과 개선에 효과적이었음을 보였다. Kim et al. [17]은 기하 평균을 활용한 부스팅 앙상블 모형을 기업 부실 예측 데이터의 불균형 문제 해결에 적용해 보았으며, 제안한 모형이 데이터의 불균형 정도와 관계없이 기존의 부스팅 모형 보다 좋은 예측 성과를 보였다.

## 3. 제안 모형

본 연구는 재무 부실화 예측을 위한 배깅 앙상블 분류기의 성능 개선에 관한 연구이다. 본 논문에서는 전통적인 배깅의 성능 개선을 위해 배깅 기법의 진행 과정 중 부트스트랩 샘플링 과정을 본 논문에서 제안한 방법으로 수정한 새로운 형태의 앙상블 모형을 제안한다. 제안한 방법은 기존의 부트스트랩 샘플링 과정을 최적화 하는 효과를 가져올 것으로 기대되며, 이를 통해 전체 앙상블 모형의 성과가 개선될 것으로 기대된다. (그림 1)은 데이터 다양화를 통한 앙상블 모형의 일반적인 흐름을 보여주고 있다.




(그림 1) 데이터 다양화를 통한 앙상블 모형
(Figure 1) Ensemble model using diversified data set

그림에서 보는 바와 같은 형태의 앙상블 모형은 원 학습 데이터로부터 출발하게 된다. 원 학습 데이터로부터 특정 방식에 의해 다양성이 존재하는 서로 다른 학습 데이터 셋을 생성시키게 된다. 데이터 다양화를 통한 대표적인 앙상블 기법인 배깅의 경우 원 학습 데이터로부터 복원 추출방식으로 랜덤하게 샘플링을 하여 학습 데이터 셋을 다양화 시킨다. 반면에 랜덤 서브스페이스 앙상블 기법의 경우 원 학습 데이터로부터 가능한 입력 변수 집합으로부터 랜덤하게 입력 변수 집합을 선택함으로써 학습 데이터 셋을 다양화 시킨다. 이와 같은 방법을 통해 서로 다른 학습 데이터 셋이 생성이 되면, 그 다음 단계는 각각의 서로 다른 학습 데이터 셋을 이용해 특정 학습 알고리즘을 이용하여 모형을 학습시키는 것이다. 이를 통해 다양성이 존재하는 분류기 $\{C_1,\cdots, C_n\}$이 만들어지게 된다. 마지막 단계는 각각의 분류기를 통해 각각의 예측 결과값 $\{O_1, \cdots , O_n\}$을 구하고 이들을 특정한 전략에 의해 결합하는 것이다.

이와 같은 데이터 다양화를 통한 앙상블 기법의 성능 개선을 위해 지금까지 다양한 연구가 진행되어 왔다. 그 중에 대표적인 것은 서로 다른 학습 데이터 다양화 방식을 결합하는 연구와 선택적 앙상블 기법에 관한 것이다. 전자의 예는 배깅 기법과 랜덤 서브스페이스 기법을 결합을 통해 다양성을 높여 최종적으로 앙상블 모형의 성과를 개선시키려는 연구이며 여러 연구자들이 두 기법을 결합할 경우 단일 앙상블 기법보다 성능이 개선되었음을 실증적으로 보여주었다 [13, 14]. 선택적 앙상블은 (그림 1)의 [B] 부분에서 생성된 모든 기저 분류기를 이용하여 앙상블을 구성하는 것이 아니고 이 중에서 결합 시 성과를 개선시킬 것으로 기대되는 기저 분류기만을 선택하여 앙상블 모형을 구성하는 방식이다[18,19,20].

Ho[21]는 앙상블 분류기의 성과를 최적화하는 방법을 범위 최적화(coverage optimization)와 결정 최적화(decision optimization)로 분류하였는데 여기서 범위 최적화는 (그림1)의 [B]가 나타내고 있는 기저 분류기 풀(classifier pool)로 부터 최적의 분류기를 선택하는 것과 관련된 것이고 결정 최적화는 (그림1)에서 [C]가 나타내는 부분인 각 분류기들의 출력 정보(output)를 결합하기 위한 규칙을 최적화 하는 것이다.

본 논문에서는 기존 연구와 달리 [A] 부분의 최적화에 초점을 맞추어 연구를 진행한다. 선택적 앙상블 기법의 경우 이미 선택된 기저 분류기 중에서 최적의 기저 분류기를 선택하는 문제라면 본 연구는 최적의 분류기를 생성하기 위한 학습 데이터 셋의 최적 발생에 초점을 맞추었다는 점에서 기존 연구와 차별성이 있다고 할 수 있다.

표준 배깅 기법은 기저 분류기를 다양화시키기 위해 원 학습 데이터로부터 부트스트랩 샘플링 방법을 통해 서로 다른 학습 데이터 셋을 생성하게 된다. 이를 통해 기저 분류기들이 다양성을 갖게 되지만 랜덤 샘플링으로 인한 임의성으로 인해 배깅 기법은 어떤 경우에는 성과가 매우 좋지만 어떤 경우는 그렇지 않은 경우도 존재하게 된다. 본 논문에서는 이와 같이 부트스트랩 샘플링으로 인해 배깅의 성과 편차가 생길 수 있다는 점에 착안하여 배깅 앙상블의 성과를 최고로 하는 부트스트랩 샘플링 방식을 찾기 위한 방법을 제안하였다. 본 논문에서는 성과를 최고로 하는 부트스트랩 샘플링을 위해 최적화 탐색 알고리즘으로 가장 대표적인 방법인 유전자 알고리즘을 적용한다[22]. 즉, 수없이 많은 부트스트랩 샘플 집합 중에서 앙상블 모형 성과 측면에서 최적(또는 근사 최적)인 부트스트랩 샘플을 선택하는 문제로 변환하여 유전자 알고리즘을 적용한다. 본 논문에서 제안한 방법을 이용하면 앙상블 모형의 성과를 최적으로 하는 부트스트랩 샘플을 선택할 수 있을 것으로 기대되며 이를 통해 배깅 앙상블 모형의 최적화를 이룰 수 있을 것으로 기대된다. 본 논문에서 제안한 모형에 대한 자세한 설명은 다음과 같다.

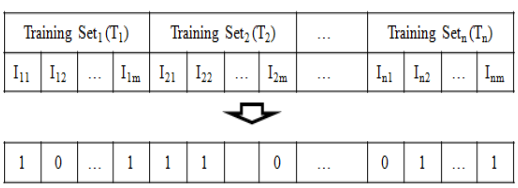
**1단계: 데이터 준비 및 입력 변수 설정**

전체 데이터를 학습용 데이터와(T) 검증용 데이터(V)로 분할한다. 학습용 데이터는 다시 모형 구축을 위한 데이터($T_A$)와 유전자 알고리즘에서 적합도 함수 계산을 위해 사용한 데이터($T_B$)로 분류한다.

**2단계: 모집단 설정**

유전자 알고리즘은 랜덤하게 선택된 염색체(chromosomes)의 모음인 모집단으로부터 시작하게 된다. 본 논문에서 각각의 염색체는 원 학습 데이터로부터 선택된 새로운 학습 데이터 집합을 나타낸다. 이를 위해 염색체는 이진열(binary string) 형태로 설계하였으며, 이는 원 학습 데이터에서 앙상블 모형을 위해 선택된 새로운 학습 데이터 집합을 의미한다.

(그림2)는 본 연구에서 사용한 유전자 알고리즘의 염색체 구조의 예를 보여주고 있다. 이진열의 각각의 비트는 0 또는 1의 값을 갖게 되며 이는 해당하는 데이터의 선택유무를 나타낸다. 원 학습 데이터가 m개의 데이터로 구성되어 있으며, n개의 서로 다른 학습 데이터를 발생시키려고 한다면 총 $n \times m$의 비트로 구성된 염색체가 필요하게 된다. 그림에서 $I_{ij}$ 에 해당하는 비트의 값이 의미하는 것은 i번째 학습 데이터($T_i$)에서 j번째 데이터의 선택유무를 나타낸다. 예를 들어 그림에서 $I_{12}$에 해당하는 비트의 값은 0이며, 이것은 새로 발생시킬 학습 데이터 셋 중에 첫 번째에 속하는 $T_1$ 에서는 원 학습 데이터에서의 두 번째 데이터를 선택하지 않는 다는 것을 의미하고, $I_{22}$ 에 해당하는 비트의 값이 1이라는 것은 원 학습 데이터에서의 두 번째 데이터가 $T_2$ 에서는 포함된다는 것을 의미한다. 살펴본 바와 같이 각각의 염색체는 서로 다른 형태의 학습 데이터 집합 {$T_1$, …, $T_n$}을 나타내게 되며, 유전자 알고리즘은 이와 같은 염색체 다수로 구성된 모집단(population)을 발생시키게 된다. 유전자 알고리즘의 첫 단계에서는 이와 같은 모집단을 랜덤하게 발생시켜 다수의 염색체가 발생하게 된다.




(그림 2) 염색체의 예
(Figure 2) An Example of Chromosome

**3단계: 적합도 함수 계산**

2단계에서는 모집단을 구성하고 있는 다수의 염색체가 발생하며, 이들 염색체는 각각 앙상블 모형을 구성하게 될 각각의 기저 분류기를 학습시키기 위한 서로 다른 학습 데이터 집합을 의미한다. 본 논문에서는 이들 염색체를 평가하기 위한 적합도 함수로 $T_B$ 데이터에서의 앙상블 모형의 분류 정확도를 사용한다.

각 염색체에 해당하는 각각의 서로 다른 학습 데이터 {$T_1$, … ,$T_n$}을 사용해 특정 학습 알고리즘으로 학습하고 이를 통해 각각의 서로 다른 분류기{$C_1$,…, $C_n$}가 생성된다. 또한, 이들 서로 다른 분류기들 각각에 대해 유전자 알고리즘을 위해 준비한 데이터인 $T_B$ 데이터 셋에서의 예측률을 구한다. 이를 통해 서로 다른 예측 결과값 {$O_1$, …, $O_n$}이 생성되며 이들을 특정한 결합 방식에 의해 결합하여 앙상블 분류기의 예측 결과값을 계산한다. 본 논문에서는 가장 대표적인 결합 방식인 다수결 투표 결합 방식을 이용한다. 이와 같은 방식으로 나온 앙상블 모형의 예측 결과값은 해당되는 염색체에 대한 적합도 값이 되며, 이는 다음 세대로 진화할 때 중요한 정보로 활용된다.

**4단계: 유전자 연산 및 재조합**

3 단계에서 계산한 각각의 염색체에 대한 적합도 값은 선택, 교배, 돌연변이 등과 같은 유전자 연산에서 활용되며 적합도 값이 높을수록 다음 세대에서 선택될 확률이 높아지게 된다. 즉, 유전자 연산을 통해 새로 생겨난 자식(offspring) 염색체와 이전 세대에서의 부모(parents) 염색체 중에서 다음 세대로 진화할 염색체의 조합이 적합도 값을 기준으로 결정된다. 이를 통해 다음 세대로 진화하게 될 염색체의 조합인 새로운 모집단이 생겨나게 된다.

**5단계: 종료 조건 확인**

종료 조건을 확인하고 종료 조건을 만족하면 6단계로 진행하고, 그렇지 않을 경우는 앞의 3,4 단계를 반복한다.

**6단계: 최적 앙상블 모형 구성**

위의 유전자 알고리즘을 통해 최적의 학습 데이터 셋 {$T^*_1$, … , $T^*_n$}이 구해지고, 이 각각의 학습 데이터 셋을 이용해 서로 다른 분류기 {$C^*_1$, … , $C^*_n$}을 생성시킨다.

**7단계: 모형 검증**

생성된 분류기를 검증용 데이터에 적용하여 각 모형의 예측 정확도를 측정한다. →{$O^*_1$, … , $O^*_n$}

**8단계: 결합**

위의 결과 값을 특정 결합 방식에 의해 결합하여 최종 앙상블 모형의 결과값을 계산한다.

## 4. 실험 설계

본 논문에서는 앙상블 모형의 성능 개선을 위한 새로운 방법을 제안하였으며, 제안한 모형을 재무 부실화 문제에 적용해 보았다. 모형의 검증을 위해 사용한 데이터는 자산규모가 10억에서 70억 사이이고 업종은 중공업인 국내 비외감 기업의 데이터로 구성되어 있다. 데이터는 총 848개로 구성되어 있으며 부도 기업과 비부도 기업이 각각 424개로 구성되어 있다. 이 데이터는 수익성, 안정성, 성장성, 활동성 및 현금 흐름으로 분류된 총 131개의 재무비율로 구성되어 있으며, 이 중에서 최종 입력 변수로는 (표 1)에 나와 있는 7개의 변수를 사용하였다. 최종 입력 변수는 단일 표본 t검정(independent-samples t-test)과 전진 선택법(forward selection)을 이용한 로지스틱 회귀분석과 선행 연구결과 등을 종합적으로 고려하여 선정하였다. (표 1)은 최종 선정된 변수에 대한 설명과 기술통계량을 보여주고 있다.

(표 1) 입력 변수

(Table 1) Input variables

| Category | Description | Range | Mean | Std. Dev | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| Profitability | Financial Expenses to Sales | 18.19 | 4.41 | 3.89 | 1.69 | 2.99 |
| | Financial Expenses to Debt | 12.47 | 5.92 | 2.93 | 0.24 | -0.29 |
| Stability | (Capital Surplus + Retained Earnings-Dividend)/ Total Assets | 70.08 | 6.40 | 12.82 | -0.66 | 2.19 |
| | Cash Ratio | 137.78 | 20.79 | 27.87 | 2.43 | 6.30 |
| Growth | Coefficient of Variation of Sales | 90.47 | 28.47 | 22.09 | 1.24 | 0.86 |
| Cash Flow | Cash Flow after Interest Payment to Sales | 0.98 | -0.01 | 0.17 | -0.72 | 2.28 |
| Activity | Sales to Net Change in Working Capital | 48.26 | 10.90 | 11.06 | 1.69 | 2.64 |

데이터는 학습용 데이터와(T) 검증용 데이터(V)로 분할하여 실험을 수행하였다. 학습용 데이터는 다시 모형 구축을 위한 데이터($T_A$)와 유전자 알고리즘에서 적합도 함수 계산을 위해 사용한 데이터($T_B$)로 분류하였다. 검증용 데이터는 모형의 최종 검증을 위해 사용되었으며, 5장의 실험 결과는 검증용 데이터에서의 결과값을 의미한다. 본 연구에서는 10-겹 검증(10-fold cross validation) 방법으로 실험을 하였으며, 앙상블 모형의 경우 10회 반복하여 실험을 수행하였으며, 10회 실험 결과의 평균값을 대푯값으로 사용하였다.

## 5. 실험 결과

본 연구에서 제안한 모형의 검증을 위해 검증용 데이터에서의 예측 정확도뿐만 아니라 앙상블 모형의 성과에 영향을 주는 요인도 함께 분석하였다. 앞에서 살펴본 바와 같이 앙상블 모형의 성과에 영향을 주는 요인으로는 앙상블을 구성하고 있는 기저 분류기의 평균 예측률과 다양성 지수가 있다. 본 연구에서는 제안한 모형의 성과 개선에 대한 심층적인 분석을 위해 앙상블 모형의 예측 정확도뿐만 아니라 기저 분류기의 평균 예측률과 다양성 지수도 함께 살펴 보았다. 기저 분류기들 사이의 다양성을 측정하기 위한 다양한 척도들이 개발 되었지만 본 연구에서는 가장 대표적인 다양성 척도 중의 하나인 Q-통계량을 살펴보았다.

두 개의 분류기($C_i$ ,$C_j$)들의 예측의 일치 정도가 (표 2)와 같다고 가정할 때 분류기 $C_i$ , $C_j$ 사이의 Q-통계량은 아래의 식 (1)과 같이 계산될 수 있다[2].

$$Q(C_i, C_j) = \frac{(N_a N_d - N_b N_c)}{(N_a N_d + N_b N_c)} \quad (3)$$

(표2) 분류기의 예측 일치도표

(Table 2) Coincident between two classifiers

| | $C_j$ correct | $C_j$ wrong |
|---|---|---|
| $C_i$ correct | $N_a$ | $N_b$ |
| $C_i$ wrong | $N_c$ | $N_d$ |

- $N_a$,$N_d$:두 분류기의 예측 결과가 일치하는 데이터 수를 의미한다.
- $N_b$,$N_c$:두 분류기의 예측결과가 불일치하는 데이터 수를 의미한다.

본 논문에서 제안한 방법의 효과성을 검증하기 위해

다양한 분류기를 대상으로 실험을 수행하였으며 실험 결과는 (표 3), (표 4), (표 5)와 같다. (표 3)은 각 모형의 예측 성과를 나타내며, (표 4)와 (표 5)는 각 앙상블 모형을 구성하고 있는 기저 분류기들의 평균 예측률과 평균 Q-통계량을 나타내고 있다.

본 논문에서 사용한 기저 분류기로는 의사결정 트리(DT: Decision Tree), k-근접이웃(KNN: k-nearest neighbor), 선형판별분석(LDA: Linear Discriminant Analysis), 로지스틱 회귀 모형(Logit)과 Support Vector Machines(SVM)이다. SVM의 경우 커널 함수가 liner인 경우(SVM(L))와 RBF인 경우(SVM(R)) 각각에 대해 실험을 수행하였다. 표에서 Single은 단일 모형을 의미하고, Bagging은 표준 배깅 앙상블 모형을 의미하며, Proposed_M은 본 논문에서 제안한 모형을 의미한다.

(표 3)에서 보는 바와 같이 전체적으로는 단일 분류 모형의 경우 RBF 커널 함수를 사용하는 SVM 모형인 SVM(R)이 가장 좋은 성과를 냈으며, 배깅 앙상블 모형과 제안한 모형의 경우 모두 DT를 기저 분류기로 할 경우가 가장 좋은 성과를 냈다. 배깅 앙상블 모형의 경우 기저 분류기가 DT, SVM(R), KNN, Logit인 경우 단일 모형 보다 성과 개선이 있었지만, LDA나 SVM(L)의 경우 오히려 기저 분류기보다 성과가 안 좋게 나왔다. 기저 분류기에 비해 가장 큰 성과 개선이 있었던 DT의 경우는 평균 Q-통계량의 값이 가장 작은 값을 보였으며, 이것이 앙상블 모형의 성과 개선과 관련이 있는 것으로 생각해 볼 수 있을 것이다.

제안한 모형의 경우 기저 분류기와 관계없이 단일 모형보다 좋은 성과를 보였으며, 표준 배깅 모형보다도 좋은 성과를 보였다. (표 4)와 (표 5)를 살펴보면 제안한 방법을 이용하여 기저 분류기를 구성할 경우 이들 기저 분류기들의 평균 예측률과 Q-통계량 값이 일반 배깅 모형보다 개선되는 것을 볼 수 있다. 이를 통해 이들 기저 분류기를 결합한 새로운 형태의 앙상블 모형의 성과가 기존의 배깅 앙상블 모형보다 좋았다는 것을 알 수 있다.

(표 3) 각 모형의 예측 정확도(%)
(Table 3) Classification accuracy of each model (%)

| | DT | KNN | LDA | Logit | SVM(L) | SVM(R) |
|---|---|---|---|---|---|---|
| Single | 70.00 | 70.54 | 70.54 | 71.35 | 71.35 | 72.43 |
| Bagging | 73.89 | 70.64 | 70.38 | 71.99 | 71.30 | 72.66 |
| Proposed_M | 75.10 | 71.21 | 71.28 | 72.29 | 71.89 | 73.12 |

(표 4) 앙상블 기저 분류기 평균 예측률 (%)
(Table 4) Average classification accuracy of base classifiers (%)

| | DT | KNN | LDA | Logit | SVM(L) | SVM(R) |
|---|---|---|---|---|---|---|
| Bagging | 70.15 | 68.30 | 71.00 | 71.68 | 71.28 | 72.46 |
| Proposed_M | 70.30 | 68.40 | 71.05 | 71.75 | 71.47 | 72.65 |

(표 5) 앙상블 기저 분류기 Q-통계량
(Table 5) Average Q-statistic value of base classifiers

| | DT | KNN | LDA | Logit | SVM(L) | SVM(R) |
|---|---|---|---|---|---|---|
| Bagging | 0.761 | 0.867 | 0.985 | 0.981 | 0.980 | 0.985 |
| Proposed_M | 0.753 | 0.858 | 0.971 | 0.979 | 0.976 | 0.980 |

제안한 모형의 성과 개선에 대한 통계적 유의성을 검토하기 위해 각 기저 분류기별로 Wilcoxon 부호 순위 검정을 하였으며 그 결과는 (표 6)과 같다. 표에서 ** 표시는 1% 수준에서 유의한 차이가 있다는 것을 의미하고, * 표시는 5% 수준에서 유의한 차이가 있다는 것을 의미한다.

표준 배깅과 단일 모형을 비교할 때 기저 분류기가 DT일 경우만이 통계적으로 유의한 차이가 있는 것으로 나왔다. 반면에, 본 논문에서 제안한 모형의 경우 기저 분류기가 SVM(L)일 때를 제외하고 모든 경우에 있어서 단일 모형 보다 유의한 차이가 있는 것으로 나왔다. 이를 통해 제안한 모형이 표준 배깅 보다 단일 모형의 성과 개선에 보다 더 효과적이었다는 것을 알 수 있다. 표준 배깅과 제안한 모형을 비교했을 경우 기저 분류기가 DT, KNN, LDA 인 경우에 유의한 차이가 있는 것으로 나왔다. 하지만, Logit, SVM(L), SVM(R)을 기저 분류기로 사용할 경우에는 제안한 모형의 성과가 표준 배깅 보다 좋게 나왔지만 이 차이가 통계적으로 유의하지는 않았다. 그러나, 전반적으로 본 논문에서 제안한 모형이 기존 모형의 성과 개선에 큰 기여를 했다고 볼 수 있다.

## 6. 결 론

앙상블 분류기는 개별적으로 학습된 일련의 분류기로 구성되며 이들 각각의 분류기의 예측 결과의 결합을 통해 최종 앙상블의 예측 결과를 얻게 된다. 일반적으로 앙상블 분류기를 잘 구성할 경우 하나의 분류기를 사용할 경우 보다 더 좋은 예측 성과를 내는 것으로 알려져 있다.

(표 6)윌콕슨 부호 순위 검정(p-value)
(Table 6) Wilcoxon signed-rank test(p-value)

| | DT | | KNN | | LDA | | Logit | | SVM(L) | | SVM(R) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bagging | Proposed_M | Bagging | Proposed_M | Bagging | Proposed_M | Bagging | Proposed_M | Bagging | Proposed_M | Bagging | Proposed_M |
| Single | 0.005** | 0.005** | 0.887 | 0.005** | 0.508 | 0.022* | 0.139 | 0.017* | 0.646 | 0.093 | 0.508 | 0.038* |
| Bagging | | 0.015* | | 0.022* | | 0.028* | | 0.678 | | 0.126 | | 0.139 |

대표적인 앙상블 기법인 배깅 기법은 기저 분류기를 다양화시키기 위해 원 학습 데이터로부터 부트스트랩 샘플링 방법을 통해 서로 다른 학습 데이터 셋을 생성하게 된다. 이를 통해 기저 분류기들이 다양성을 갖게 되지만 랜덤 샘플링으로 인한 임의성으로 인해 어떤 경우에는 성과가 매우 좋지만 어떤 경우는 그렇지 않은 경우도 존재하게 된다. 본 논문에서는 이와 같이 부트스트랩 샘플링으로 인해 배깅의 성과 편차가 생길 수 있다는 점에 착안하여 배깅 앙상블의 성과를 최고로 하는 부트스트랩 샘플링 방식을 찾기 위한 방법을 제안하였다.

본 논문에서는 성과를 최고로 하는 부트스트랩 샘플링을 위해 최적화 탐색 알고리즘으로 가장 대표적인 방법인 유전자 알고리즘을 적용하였다. 즉, 수없이 많은 부트스트랩 샘플 집합 중에서 앙상블 모형 성과측면에서 최적(또는 근사 최적)인 부트스트랩 샘플을 선택하는 문제로 변환하여 유전자 알고리즘을 적용하였다. 본 논문에서 제안한 방법을 이용하면 앙상블 모형의 성과를 최적으로 하는 부트스트랩 샘플을 선택할 수 있을 것으로 기대되며 이를 통해 앙상블 모형의 최적화를 이룰 수 있을 것으로 기대된다.

본 논문에서 제안한 모형의 우수성을 검증하기 위해 국내 부도 예측 관련 데이터를 가지고 실험을 수행하였다. 실험 결과 제안한 방법이 기저 분류기들의 평균 예측률과 다양성 지수 값을 개선시키는 효과가 있었으며 이를 통해 제안한 앙상블 모형의 예측 성능 또한 일반 배깅 앙상블 모형보다  향상되었다는 것을 알 수 있었다.

본 논문에서는 제안한 방법을 부도 예측 문제에 적용하여 효과성을 검증하였지만, 제안한 모형은 다른 분류 문제에서도 적용될 수 있을 것으로 기대된다. 향후 이에 대한 추가적인 검증 및 연구가 필요할 것이다.

## 참 고 문 헌 (Reference)


[1] T. G. Dietterich, “Machine-learning research: Four current directions,” AI Magazine, Vol.18, No.4, 1997, pp. 97-136.
http://dx.doi.org/10.1609/aimag.v18i4.1324

[2] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.
http://www.amazon.com/Combining-Pattern-Classifiers-Methods-Algorithms/dp/0471210781

[3] S. Bian, W. Wang, “On diversity and accuracy of homogeneous and heterogeneous ensembles,” International Journal of Hybrid Intelligent Systems, Vol.4, No.2, 2007, pp.103-128.
http://content.iospress.com/articles/international-journal-of-hybrid-intelligent-systems/his00044

[4] L. I. Kuncheva, C. J. Whitaker, “Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy,” Machine Learning, Vol.51, No.2, 2003, pp. 181-207.
http://dx.doi.org/10.1023/A:1022859003006

[5] L. Breiman, “Bagging predictors,” Machine Learning, Vol. 24, No.2, 1996, pp. 123-140.
http://dx.doi.org/10.1007/BF00058655

[6] Y. Freund, R. Schapire, “Experiments with a new boosting algorithm,” Proceedings of the 13th International Conference on Machine learning, 1996, pp. 148-156.
http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.1040

[7] T. Ho, “The random subspace method for construction decision forests,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, pp.832-844.
http://dx.doi.org/10.1109/34.709601

[8] L. Hansen, Salamon, P, “Neural network ensembles,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, No.10, 1990, pp. 993-1001.
http://dx.doi.org/10.1109/34.58871

[9] S.H. Kim, J.W., Kim, “SOHO Bankruptcy Prediction Using Modified Bagging Predictors,” Journal of Intelligence and Information Systems, Vol.13, No.2,

2007, pp. 15-26.
http://koreascience.or.kr/article/ArticleFullRecord.jsp?cn=JJSHBB_2007_v13n2_15

[10] C. Tsai, J. Wu. “Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring,” Expert Systems with Applications, Vol.34, No.4, 2008, pp. 2639-2649.
http://dx.doi.org/10.1016/j.eswa.2007.05.019

[11] M. Kim, “A Performance Comparison of Ensemble in Bankruptcy Prediction,” Entrue Journal of Information Technology, Vol.8, No.2, 2009, pp. 41-49.
http://scholar.ndsl.kr/schDetail.do?cn=NART50339718

[12] H. Li, Y.-C. Lee, Y.-C. Zhou, J. Sun, “The random subspace binary logit (RSBL) model for bankruptcy prediction,” Knowledge-Based Systems, Vol. 24, No.8, 2011, pp. 1380 – 1388
http://dx.doi.org/10.1016/j.knosys.2011.06.015

[13] S. Min, “Developing an Ensemble Classifier for Bankruptcy Prediction,” Journal of the Korea Industrial Information Systems Research, Vol. 17, No. 7, 2012, pp. 139-148.
http://dx.doi.org/10.9723/jksiis.2012.17.7.139

[14] A.I. Marques, V. Garcia, and J. S. Sanchez. “Two-Level Classifier Ensembles for Credit Risk Assessment,” Expert Systems with Applications, Vol.39, No.12, 2012, pp. 10916 – 10922.
http://dx.doi.org/10.1016/j.eswa.2012.03.033

[15] H.N. Choi, D.H. Lim, “Bankruptcy prediction using ensemble SVM model,” Journal of the Korean Data and Information Science Society, Vol.24, No.6, 2013, 1113-1125.
http://dx.doi.org/10.7465/jkdi.2013.24.6.1113

[16] S. Min, “Bankruptcy Prediction Using an Improved Bagging Ensemble,” Journal of Intelligence and Information Systems, Vol.20, No.4, 2014, pp. 121-139.
http://dx.doi.org/10.13088/jiis.2014.20.4.121

[17] M. Kim, D. Kang, H.B. Kim, “Geometric Mean Based Boosting Algorithm with over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction,” Expert Systems with Applications, Vol.42, No.3, 2015, pp. 1074-1082.
http://dx.doi.org/10.1016/j.eswa.2014.08.025

[18] C. Hung, J-H. Chen, “A Selective Ensemble Based on Expected Probabilities for Bankruptcy Prediction,” Expert Systems with Applications, Vol.36, No.3, 2009, pp. 5297-5303.
http://dx.doi.org/10.1016/j.eswa.2008.06.068

[19] K. Li, Z. Liu, Y. Han, “Study of Selective Ensemble Learning Methods Based on Support Vector Machine,” Physics Procedia, Vol. 33, 2012, pp.1518 – 1525.
http://dx.doi.org/10.1016/j.phpro.2012.05.247

[20] Y. Guo, et al., “A Novel Dynamic Rough Subspace Based Selective Ensemble,” Pattern Recognition, Vol.48, No.5, 2014, pp. 1638-1652.
http://dx.doi.org/10.1016/j.patcog.2014.11.001

[21] T. K. Ho, “Multiple classifier combination: Lessons and the next steps,” In A. Kandel and H. Bunke, editors, Hybrid Methods in Pattern Recognition. World Scientific Publishing, 2002
http://dx.doi.org/10.1142/9789812778147_0007

[22] D. E. Goldberg, “Genetic algorithms in search, optimization and machine learning,” New York: Addison-Wesley, 1989.
http://catalogue.pearsoned.co.uk/educator/product/Genetic-Algorithms-in-Search-Optimization-and-Machine-Learning/9780201157673.page

## ◐ 저 자 소 개 ◑

**민 성 환 (Sung-Hwan Min)**
1996년 동국대학교 산업공학과 학사
1999년 고려대학교 산업공학과 석사
2005년 한국과학기술원 경영공학과 박사
2005년 - 현재 한림대학교 경영학과 부교수
관심분야: 데이터마이닝