J. Soc. Korea Ind. Syst. Eng Vol. 39, No. 1 : 105-115, March 2016
http://dx.doi.org/10.11627/jkise.2016.39.1.105
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Improving an Ensemble Model Using Instance Selection Method

Sung-Hwan Min†

Department of Business Administration, Hallym University

사례 선택 기법을 활용한 앙상블 모형의 성능 개선

민 성 환†

한림대학교 경영학과

Ensemble classification involves combining individually trained classifiers to yield more accurate prediction, compared with individual models. Ensemble techniques are very useful for improving the generalization ability of classifiers. The random subspace ensemble technique is a simple but effective method for constructing ensemble classifiers; it involves randomly drawing some of the features from each classifier in the ensemble. The instance selection technique involves selecting critical instances while deleting and removing irrelevant and noisy instances from the original dataset. The instance selection and random subspace methods are both well known in the field of data mining and have proven to be very effective in many applications. However, few studies have focused on integrating the instance selection and random subspace methods. Therefore, this study proposed a new hybrid ensemble model that integrates instance selection and random subspace techniques using genetic algorithms (GAs) to improve the performance of a random subspace ensemble model. GAs are used to select optimal (or near optimal) instances, which are used as input data for the random subspace ensemble model. The proposed model was applied to both Kaggle credit data and corporate credit data, and the results were compared with those of other models to investigate performance in terms of classification accuracy, levels of diversity, and average classification rates of base classifiers in the ensemble. The experimental results demonstrated that the proposed model outperformed other models including the single model, the instance selection model, and the original random subspace ensemble model.

Keywords : Random Subspace, Bagging, Bankruptcy Prediction, Genetic Algorithms

1. 서 론

앙상블 분류의 기본 개념은 개별 분류기보다 더 좋은 예측 정확도를 얻기 위해 다수의 분류기들을 결합하는 것이다. 이와 같은 앙상블 기법은 분류기의 일반화 성능(generalization ability)을 향상시키는 것으로 알려져 있다. 앙상블 모형은 우수한 예측성과로 인해 최근에 데이터 마이닝 분야에서 큰 관심을 끌고 있다. 앙상블 모형의 일반화 성능을 향상시키기 위해서는 앙상블 모형에 사용된 기저 분류기들이 가능하면 예측성과가 좋고 다양성(diversity)이 존재해야 한다. 여기서 다양성이란 기저 분류기들의 예측 결과가 얼마나 다른지의 정도를 의미하며, 두 기저 분류기들의 예측 결과가 일치한다면 두 분류기간의 다양성이 존재하지 않는다고 말하고, 반대의 경우에는 다양성이 존재한다고 말한다. 이와 같은 분류기간의 다양성은 앙상블 모형에서 매우 중요한 요소로, 만약 동일한 다수의

Received 22 February 2016; Finally Revised 11 March 2016; Accepted 14 March 2016
† Corresponding Author : shmin@hallym.ac.kr

분류기들을 결합한다면 이들 간의 다양성은 전혀 없다고 말할 수 있으며 이들의 결합을 통한 예측 성능의 개선은 전혀 없을 것이다. 이와 같은 이유로 앙상블 모형에서는 기저 분류기들을 다양화시키는 것이 매우 중요하다[4].

앙상블의 기저 분류기들을 다양화시키기 위한 방법으로는 학습 데이터의 다양화, 학습 알고리즘의 다양화, 파라미터의 다양화 등의 방법이 있으며 랜덤 서브스페이스 앙상블 기법은 학습 데이터의 다양화를 통해 앙상블 모형을 구성하는 대표적인 기법중의 하나이다[13]. 랜덤 서브스페이스 앙상블 모형은 랜덤하게 선택된 서로 다른 입력변수 집합을 가지고 각각의 기저 분류기들을 학습시킴으로써 기저 분류기를 다양화시키는 대표적인 앙상블 기법으로 분류기의 성능 개선에 효과적인 것으로 알려져 있다.

본 논문에서는 랜덤 서브스페이스 앙상블 모형의 예측 성능을 향상시키기 위해 사례 선택 기법(instance selection method)과 랜덤 서브스페이스 앙상블 모형을 결합하는 새로운 형태의 앙상블 모형을 제안하였다. 사례 선택은 원 학습 데이터(original training data)로부터 불필요하거나 예측 모형에 악영향을 주는 사례를 제거하고 핵심적인 사례를 선택하는 것을 말한다[10]. 사례 선택 기법과 랜덤 서브스페이스는 각각 데이터 마이닝 분야에서 자주 사용되고 있는 기법으로 다양한 응용 분야에서 효과적으로 적용되고 있다. 하지만, 이들 두 개 기법의 결합에 초점을 둔 연구는 거의 없는 것이 현실이다. 본 논문에서는 최적의 사례 집합을 선택하기 위해 유전자 알고리즘을 활용하였으며, 선택된 사례 집합은 랜덤 서브스페이스 앙상블 모형의 입력 변수로 사용되었다. 본 연구에서 제안한 모형을 검증하기 위해 개인과 기업의 재무 부실화 예측 문제에 제안한 모형을 적용해 보았다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 사례 선택, 앙상블 분류기에 대한 설명을 하고 재무 부실화에 관한 관련 연구를 살펴보았다. 제 3장에서는 본 논문에서 제안한 사례 선택 기법과 랜덤 서브스페이스 앙상블 기법의 결합 모형에 관해 설명을 하였다. 제 4장에서는 연구 데이터에 대한 설명과 실험 설계에 대한 설명을 하였으며, 제 5장에서는 실험 결과와 이에 대한 분석 및 해석을 하였다. 끝으로 제 6장에서는 연구의 결론 및 요약과 함께 향후 연구 과제에 대해 설명하였다.

2. 문헌 연구

2.1 사례 선택

사례 선택은 데이터 마이닝에서 활용되고 있는 효과적인 기법중의 하나로 원 데이터로부터 불필요하거나 예측 모형 구축에 악영향을 주는 사례를 제거하고 핵심적인 사례를 선택하는 것이다. 이와 같은 방식에 의해 선택된 사례들을 이용하여 모형을 구축할 경우 모든 사례를 이용할 경우와 비교할 때보다 유사하거나 더 좋은 예측 성과를 낼 수 있는 것으로 알려져 있다[10, 15]. 이와 같은 절차는 데이터를 축소한다는 점과 모형의 예측성과를 개선시킬 수 있다는 점에서 특징 변수 선택(feature selection)과 유사하지만 큰 차이가 있다. 사례 변수 선택은 사례 공간(instance space) 상에서의 데이터 축소를 의미하고 특징 변수 선택은 특징 변수 공간(feature space) 상에서의 데이터 축소를 의미한다. 즉, 특징 변수 선택은 불필요하거나 중복적인 특징 변수를 제거하고 핵심적인 특징 변수를 선택하는 것이다.

최초의 사례 선택 기법 중의 하나는 농축된 최근접 이웃 규칙(Condensed Nearest Neighbor Rule : CNN)이다[12]. CNN의 소개 이후 사례 선택 기법에 관한 많은 연구가 진행되었다. Derrac et al.[7]은 사례 선택 기법을 특징 변수 선택과 유사한 방식에 의해 선택 전략에 따라 크게 두 그룹으로 구분하였다. 그 중 하나는 wrapper 방식이고 다른 하나는 filter 방식이다. wrapper 방식은 사례 선택을 위해 사용하는 성과 측정을 위해 분류기를 사용하는 접근 방식이고 반면에 filter 접근 방식은 분류기와는 독립적으로 사례 선택 절차가 진행되는 방식이다. wrapper 기반 사례선택 기법은 분류기의 유형에 따라 다시 두 그룹으로 분류될 수 있다. 하나는 프로토타입 선택(prototype selection) 기법이고 다른 하나는 학습 데이터 셋 선택 기법이다. 프로토타입 선택 기법은 k-최근접 이웃(k-nearest neighbor) 방식과 같이 게으른 학습(lazy learning) 모형에 적용되는 방식으로 프로토타입에 기반한 방식이며, 학습 데이터 셋 방식은 로지스틱 회귀 분석이나 인공신경망 모형과 같은 일반적인 예측 모형에 적용되는 방식이다[10].

2.2 앙상블 분류기

앙상블 분류기는 단일 분류기의 예측 정확도를 높이기 위해 개별적으로 학습된 서로 다른 분류기들을 결합하는 접근 방식이다. 이와 같은 앙상블 기법은 분류기의 일반화 성능을 향상시키는 데 매우 유용한 것으로 알려져 있다[8]. 앙상블 모형의 일반화 성능을 향상시키기 위해서, 기저 분류기들은 가능하면 예측성과가 좋아야 하고, 또한 가능하면 서로 다른 예측 오차를 내야 한다. 즉 다양성이 존재해야 한다. 분류기 사이의 다양성은 앙상블 모형에서 매우 중요한 요소이다. 만약에 동일한 패턴으로 예측 오차를 내는 분류기들을 결합한다면 이들의 결합을 통해 어떠한 성과개선도 기대할 수 없을 것이다. 반면에 분류기 사이에 다양성이 존재한다면 어느 하나의

분류기가 오분류를 하더라도 나머지 분류기들이 옳게 분류하여 이들의 결합을 통해 오분류를 줄일 수 있을 것이다. 이와 같이 앙상블 모형에서 기저 분류기들의 다양성은 모형의 성과와 관련이 있는 중요한 요소이다.

앙상블 모형의 기저 분류기들을 다양화시키는 몇 가지 방법은 아래와 같이 정리할 수 있다.

- 학습 데이터의 다양화 : 학습 알고리즘은 동일하게 사용하는 대신 학습 데이터를 다양화시킴으로써 기저 분류기들을 다양화시키는 방식이다. 여기에 속하는 대표적인 예로는 배깅(bagging), 부스팅(boosting), 랜덤 서브스페이스(random subspace) 방식이 있다.
- 학습 알고리즘의 다양화 : 학습 데이터는 동일하게 사용하는 대신에 학습 알고리즘을 다양화 시키는 방식이다.
- 학습 알고리즘의 파라미터 다양화 : 학습 데이터와 학습 알고리즘을 동일하게 두고 학습 파라미터를 다양화함으로써 기저 분류기를 다양화 시키는 방식이다.
- 하이브리드 방법 : 기저 분류기를 다양화시키기 위해 여러 가지 기법을 결합하는 경우가 이에 속한다.

배깅은 bootstrap aggregation의 약자로 원 학습 데이터로부터 복원추출 방식에 의해 서로 다른 학습 데이터를 랜덤하게 선택함으로써 학습 데이터를 다양화시키고 이를 통해 기저 분류기를 다양화시키는 대표적인 앙상블 기법이다[5]. 부스팅은 배깅과 유사하지만 이전에 생성된 분류기의 예측 정확도 정보를 활용하여 다음에 사용할 새로운 학습 데이터를 선택한다는 점이 다르다. 즉, 이전 분류기에 의해 오분류된 데이터의 가중치를 높이고, 옳게 분류된 데이터의 가중치는 낮추는 방법을 통해 서로 다른 가중치를 가지고 다음에 사용할 학습 데이터를 선택하게 된다[9].

앙상블 모형의 기저 분류기는 입력 변수를 서로 다르게 선택함으로써 다양화될 수도 있다. 랜덤 서브스페이스 기법은 이와 같은 접근 방법의 앙상블 기법중의 하나이다. 랜덤 서브스페이스 기법은 각각의 기저 분류기들을 위해 원 입력 공간(original feature space)으로부터 랜덤하게 입력 변수집합을 선택하며 배깅은 원 사례 집합(original instance space)로부터 랜덤하게 사례 집합을 선택한다. 랜덤 서브스페이스 앙상블은 이와 같이 랜덤하게 선택된 서로 다른 입력 변수 집합을 가지고 각각의 기저 분류기들을 학습시키고, 이를 통해 서로 다른 각각의 기저 분류기들이 앙상블 모형을 구성하게 되며, 이들 각각의 분류기의 예측값들은 다수결 투표와 같은 결합 방식에 의해 결합되게 된다[13].

랜덤 서브스페이스 앙상블 모형의 성과에 영향을 주는 대표적인 파라미터로는 앙상블의 크기와 선택된 입력 변수의 수가 있다. 선행 연구에 의하면 원 입력 변수의 50% 정도를 랜덤하게 선택하여 기저 분류기를 발생시키고 앙상블 모형을 구성할 경우에 앙상블 모형의 예측성과가 최고이거나 최고에 근접한다고 알려져 있다[13].

2.3 재무 부실화 예측

재무 부실화 예측은 회계 및 재무 분야에서 매우 중요한 주제이다. 재무 부실화와 관련된 비용은 매우 크기 때문에, 재무 부실화와 관련된 예측을 정확하게 하는 것은 금융 기관에서 매우 중요한 문제이다. 지난 수십 년간 많은 연구자들이 재무 부실화 예측과 관련된 연구를 진행해오고 있으며, 초창기의 연구는 주로 전통적인 통계 모형을 활용하여 재무 부실화를 예측하려는 시도가 대부분이었다[2, 3, 23, 26]. 그 뒤로 사례기반추론(Case-based reasoning(CBR), 귀납적 학습 방법, support vector machine과 인공신경망과 같은 새로운 모형들을 재무 부실화 예측 문제에 적용해 보는 연구가 뒤따랐다[6, 22, 24, 28, 35].

최근에는 재무 부실화 예측 모형의 성능을 개선시키기 위해 앙상블 모형을 활용하려는 연구가 활발하게 진행되고 있다. Tsai and Wu[29]는 부도 예측 모형에서 다수의 인공 신경망 모형을 사용할 경우의 효과성에 대한 연구를 수행하였다. 본 논문에 따르면 다수의 인공 신경망 모형을 결합한 모형이 단일 모형 중 가장 좋은 성과를 낸 값보다 좋은 결과를 내지는 않았다. Nanni and Lumini[25]는 부도 예측과 신용 평가 문제 해결을 위해 다양한 형태의 앙상블 모형을 적용해 보았다. 이들의 실험 결과 Levenberg-Marquardt 인공신경망을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형이 가장 좋은 성과를 보였다. Hung and Chen[14]은 기대확률(expected probability)에 기반을 둔 선택적 앙상블 모형을 기존의 다수결 투표 방식의 앙상블 모형과 비교해 보았다. 부도 예측 관련 데이터를 이용해 실험한 결과 제안한 모형의 예측성과가 기존의 앙상블 모형보다 우수한 결과를 보였다. Kim and Kang[16]은 부도 예측을 위한 인공 신경망 모형의 예측성과를 향상시키기 위해 대표적인 앙상블 기법인 배깅과 부스팅 앙상블 기법을 활용하였으며 이들 앙상블 기법을 통한 예측성과 개선을 실증적으로 보였다. Louzada et al.[20]는 연속적인 샘플링 과정을 반복하여 기저 분류기들을 결합하는 폴리 배깅이라는 배깅 변형 모형을 제안하여 재무 부실화 문제에서의 효과성을 실증적으로 분석하였다. Li et al.[19]는 부도 예측 문제에서 로지스틱 회귀 모형을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형이 전통적인 개별 분류 모형 보다 예측성과 면에서 우수한 결과를 낸다는 것을 보였다. Wang and Ma[30]는 두 개의 앙상블 기법을 결합하는 혼합 앙상블 모형을 재무 부실화를 예측하기 위

한 문제에 적용해 보았다. 이들은 대표적인 앙상블 기법인 배깅과 랜덤 서브스페이스 앙상블 기법을 결합하였으며, 결합한 혼합 앙상블 모형이 배깅과 랜덤 서브스페이스 단일 앙상블 모형보다 좋은 성과를 낸다는 것을 실증적으로 보였다. Marqués et al.[21]는 신용 평가 데이터를 이용해 서로 다른 학습 알고리즘을 갖는 분류기들을 다양한 앙상블 기법에 적용해 보며 성과를 비교 분석하는 연구를 수행하였다. Abellán and Mantas[1]는 부도 예측과 신용 평가 문제에 앙상블 기법을 적용해 보았다. 이들의 실험 결과에 따르면 크리덜 의사결정 트리(creedal decision tree)를 기저 분류기로 사용하는 배깅 앙상블 모형이 다른 비교 모형보다 좋은 결과를 보였다. Kim et al.[17]는 기하 평균을 기반으로 하는 부스팅 앙상블 모형을 부도 예측 문제에 적용해 보았으며, 제안한 모형이 예측성과 개선에 효과적이었음을 보였다.

3. 연구 모형

사례 선택 기법과 랜덤서브스페이스 기법은 데이터 마이닝에서 유용하게 사용되는 기법으로 많은 응용 분야에서 매우 효과적인 것으로 알려져 있다. 그러나, 이들 사례 선택 기법과 랜덤 서브스페이스의 결합에 초점을 둔 연구는 아직까지 거의 없는 것이 현실이다. 이에 본 연구에서는 예측 모형의 성과 개선에 효과적인 이 두 기법을 결합하는 새로운 형태의 앙상블 모형을 제안하였다. 우선 사례 선택 기법을 위해 유전자 알고리즘이 사용되었다. 유전자 알고리즘은 효과적인 탐색 기법으로 다양한 분야의 최적화 문제 해결에 성공적으로 적용되어 왔다[27, 32, 33, 34]. 본 논문에서는 유전자 알고리즘을 통해 불필요한 사례를 제거하고, 핵심적인 사례를 선택하여 모형 성과 측면에서 최적인 사례 집합을 선택하게 된다. 또한, 이와 같은 방식으로 선택된 최적의 사례 집합은 랜덤 서브스페이스 앙상블 모형의 입력 데이터로 사용되게 된다. 제안한 하이브리드 앙상블 모형은 크게 유전자 알고리즘 기반 사례 선택 모듈과 랜덤서브스페이스 앙상블 모듈의 두 부분으로 구성되어 있다.

첫 번째 모듈에서는 유전자 알고리즘이 랜덤 서브스페이스 앙상블 모형의 입력 변수로 사용하게 될 최적의 사례 집합을 선택하는 데 이용된다. 이 단계에서는 기저 분류기의 예측성과를 높이기 위해 핵심적인 사례가 선택되고 불필요하거나 예측 모형에 악영향을 주는 사례가 제거된다.

유전자 알고리즘은 최적 (또는 근사 최적)의 해를 찾기 위한 효율적인 탐색 알고리즘이다. 유전자 알고리즘은 더 좋은 해집합을 생성하게 될 일련의 염색체 집합인 모집단을 계속 유지하면서 최적해를 탐색해 나간다[11]. 본 논문에서는 최적의 사례 집합을 찾기 위해 가능한 해집합을 염색체로 암호화하였으며 유전자 알고리즘을 통해 최적의 염색체를 탐색하게 된다.

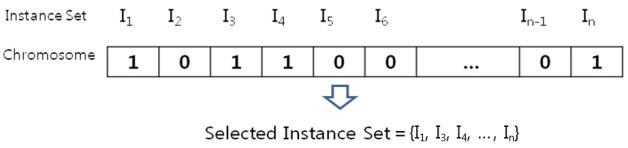
두 번째 모듈에서는 이전 모듈에서 선택된 최적의 사례 집합이 랜덤 서브스페이스 앙상블 모형의 입력 변수로 사용된다. 본 논문에서는 의사결정 트리 모형이 랜덤 서브스페이스 앙상블 모형의 기저 분류기로 사용되었으며, 다수결 투표 전략이 결합 모형으로 사용되었다. 본 논문에서 제안한 모형에 대한 자세한 설명은 다음과 같다.

1단계 : 데이터 준비 및 각종 파라미터 설정

데이터를 학습용 데이터(T)와 검증용 데이터(V)로 분할하고 학습용 데이터는 다시 모형 구축을 위한 데이터(T_1)와 유전자 알고리즘에서 적합도 계산을 위해 사용한 데이터(T_2)로 분류한다. T_2를 따로 둔 것은 유전자 알고리즘의 최적화 탐색 시 과적합(overfitting)을 줄이기 위함이다.

2단계 : 모집단 설정

유전자 알고리즘에서 모집단(population)은 다수의 염색체(chromosomes)로 구성되어 있으며 초기의 모집단은 랜덤하게 발생하게 된다. 본 논문에서 각각의 염색체는 최적의 사례 집합을 선택하기 위해 <Figure 1>과 같이 이진열(binary string) 형태로 설계되었다. 그림에서 염색체의 각각의 비트는 대응되는 사례의 선택 유무를 나타낸다. 즉, 비트의 값이 1이면 대응되는 사례가 선택된다는 것을 의미하고, 비트의 값이 0이면 대응되는 사례가 선택되지 않는다는 것을 의미한다. 이와 같은 방식으로 염색체가 이진열 형태로 암호화되었으며 이것은 원 학습 데이터로부터 선택된 특정한 사례 집합을 의미하게 된다.



<Figure 1> Example of a Chromosome

3단계 : 적합도 함수 계산

이전 단계에서는 다수의 염색체가 생성되며, 이들 염색체는 각각 학습 데이터로부터 선택된 특정한 사례 집합을 의미한다. 각각의 염색체는 적합도 함수(fitness function)에 의해 평가된다. 본 논문에서는 모형의 예측 정확도가 적합도 함수로 사용되었으며 식 (1)과 같이 적합도 함수를 수식으로 표현할 수 있다.

$$F = \frac{\sum_{i=1}^{n} H_i}{n} \quad (1)$$

여기서 H_i는 i번째 데이터의 예측값과 실제값이 일치할 경우, 즉 옳게 예측할 경우 1의 값을 갖게 되고, 그렇지 않은 경우, 즉 오분류를 한 경우 0의 값을 갖는 함수이며, n은 적합도 계산을 위해 사용한 데이터의 총 수를 의미한다. 이와 같은 방식으로 구한 적합도 값은 다음 세대로 진화할 때 중요한 정보로 활용된다.

4단계 : 유전자 연산 및 재조합

2단계에서 계산한 각각의 염색체에 대한 적합도 값은 재생, 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, 4단계를 반복한다.

6단계 : 최적 사례 집합 구성

위의 유전자 알고리즘을 통해 최적의 사례 집합이 구해지며 이는 다음 단계의 입력 변수로 사용된다.

7단계 : 입력 변수 집합 선택

원 학습 데이터(T) 대신 앞의 단계에서 구한 최적의 사례 집합 모음인 새로운 학습 데이터(T')를 랜덤 서브스페이스 앙상블 모형의 학습 데이터로 사용한다. T'의 입력 변수 공간{$feat_1$, …, $feat_F$}에서 비복원 추출 방식으로 랜덤하게 f개의 입력 변수를 선택하여 새로운 입력변수 집합 FS_i를 생성시키고 이것을 앙상블 사이즈(K)만큼 반복하여 K개의 서로 다른 입력 변수 집합을 생성한다. → {FS_1, …, FS_K}

8단계 : 분류기 학습 및 앙상블 구성

이전 단계에서 생성된 서로 다른 입력 변수 집합 FS_i로 구성된 학습 데이터를 이용해 각각의 새로운 기저 분류기 C_i를 생성시킨다. → {C_1, …, C_K}

9단계 : 모형 검증

검증용 데이터(V)를 이용해 모형을 검증한다. 8단계에서 생성된 각각의 분류기의 예측 결과값 {O_1, …, O_K}을 적절한 전략에 의해 결합한다. 본 논문에서는 다수결 투표 방식에 의해 결합하였다.

4. 실험 설계

본 연구에서 제안한 모형을 검증하기 위해 개인과 기업의 재무 부실화 예측 문제에 제안한 모형을 적용해 보았다. 개인의 재무 부실화 관련 데이터로는 2011년 Kaggle 컴피티션에서 사용된 데이터를 사용하였다[31]. 이 데이터는 총 15만 명의 데이터로 구성되어 있으며, 개인의 재무 부실을 나타내는 종속변수와 10개의 입력변수로 구성되어 있다. 본 연구에서는 15만 개의 데이터 중에서 재무 부실을 겪게 되는 데이터와 그렇지 않은 데이터를 각각 3,000개씩 랜덤하게 추출하여 총 6,000개의 데이터로 실험을 수행하였다.

기업의 재무 부실화 관련 데이터로는 자산 규모가 10억에서 70억 사이인 국내 비외감 기업의 데이터를 사용하였다. 실험에 사용된 데이터는 총 1,800개로 구성되어 있으며, 이중 부도 기업과 비부도 기업이 각각 900개로 이루어져 있다. 본 논문에서는 기업의 부도 여부를 예측하기 위해 재무 비율을 입력 변수로 사용하였다. 본 논문에서 최종적으로 사용한 개인 재무 부실화 관련 문제의 입력 변수와 기업 재무 부실화 관련 문제의 입력 변수는 각각 <Table 1>과 <Table 2>에 나와 있다.

<Table 1> Input Variables(Kaggle Data)

Variable Name	Description
Revolving Utilization Of Unsecured Lines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits
Age	Age of borrower in years
Number Of Time 30~59 Days Past Due Not Worse	Number of times borrower has been 30~59 days past due but no worse in the last 2 years.
Debt Ratio	Monthly debt payments, alimony, living costs divided by monthy gross income
Monthly Income	Monthly income
Number Of Open Credit Lines And Loans	Number of Open loans(installment like car loan or mortgage) and Lines of credit(e.g. credit cards)
Number Of Times 90 Days Late	Number of times borrower has been 90 days or more past due.
Number Real Estate Loans Or Lines	Number of mortgage and real estate loans including home equity lines of credit
Number Of Time 60~89 Days Past Due Not Worse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
Number Of Dependents	Number of dependents in family excluding themselves(spouse, children etc.)

<Table 2> Input Variables(Corporate data)

Category	Description
Activity	Sales to net change in working capital
	Total assets turnover period
Cash flow	Cash flow after interest payment to sales
Growth	Coefficient of variation of sales
Profitability	Net income to sales
	Ordinary Income to Capital
	EBITDA to Sales
Stability	Current Asset to Current Liability
	Current Liability to Total Asset
	(Capital surplus+retained earnings-dividend)/total assets
	Borrowings to Sales
	Cash ratio
	Fixed Asset to Owner's Equity
	Fixed Liability to Owner's Equity
	Borrowings to EBITDA

<Table 3> Classification Performance of Each Model(%)-Kaggle Data

f	RS	Proposed_M	DT	ISDT
1	68.82	67.20	73.05	74.12
2	73.37	72.97		
3	74.85	75.15		
4	74.92	75.51		
5	74.79	75.81		
6	75.16	75.85		
7	74.71	76.12		
8	74.23	75.68		
9	73.83	75.03		
10	73.05	74.12		
Average	73.77	74.34	73.05	74.12

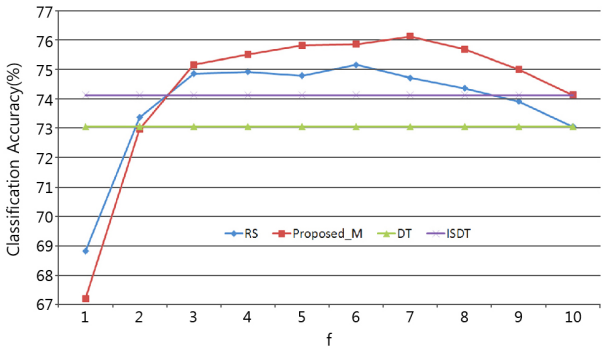
<Table 4> Classification Performance of Each Model(%)-Corporate Data

f	RS	Proposed_M	DT	ISDT
1	61.20	65.51	65.41	68.11
2	67.66	69.24		
3	69.47	70.05		
4	70.41	71.24		
5	71.22	71.85		
6	70.76	72.06		
7	71.23	72.16		
8	70.65	72.57		
9	70.58	71.70		
10	70.12	70.84		
11	70.03	70.62		
12	68.82	69.24		
13	67.72	68.35		
14	65.78	68.11		
15	65.41	68.11		
Average	68.74	70.11	65.41	68.11

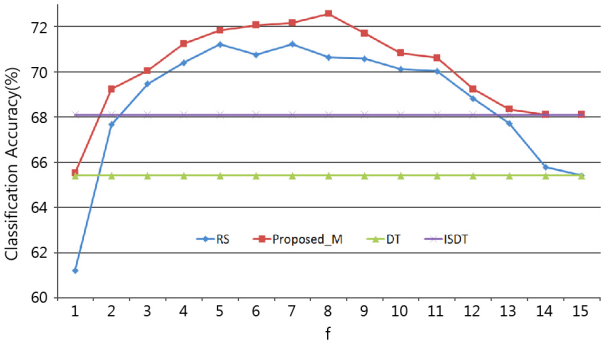
최종 데이터는 학습을 위한 데이터와 검증을 위한 데이터로 분류하여 실험을 수행하였으며, 다시 학습을 위한 데이터는 기저 분류기의 모형 구축을 위한 데이터와 유전자 알고리즘에서 적합도 값을 계산할 때 사용할 데이터로 분류하였다. 10-겹 검증(10-fold cross validation) 방법으로 실험을 수행하였으며, 앙상블 모형의 경우 10회 반복하여 실험을 수행하여 평균값을 대푯값으로 사용하였다. 앙상블 모형의 기저 분류기로는 의사결정 트리(Decision tree) 모형을 사용하였으며 각각의 앙상블 모형에서 기저 분류기의 총 수는 25로 고정하고 실험을 하였다.

5. 실험 결과

본 연구에서는 랜덤 서브스페이스 앙상블 모형의 성능 개선을 위해 사례 선택 기법과 랜덤 서브스페이스 기법을 결합한 새로운 형태의 앙상블 모형을 제안하였다. 본 연구에서 제안한 모형의 검증을 위해 제 4장에서 살펴본 바와 같이 개인 재무 부실화 관련 데이터(Kaggle data)와 국내 기업의 데이터(Corporate data)를 사용하였다. 두 데이터를 사용해 각 모형 별 예측 정확도를 비교해 보았으며 결과는 <Table 3>, <Table 4>, <Figure 2>에 나와 있다.



(A) Kaggle data



(B) Corporate data

<Figure 2> Classification Performance of Each Model at Different Feature Subset Sizes

표와 그림에서 DT는 의사결정 트리 단일 모형을 의미하고 ISDT는 사례 선택 기법을 활용한 의사결정 트리 단일 모형을 의미한다. RS는 표준 랜덤 서브스페이스 앙상블 모형을 의미하고, Proposed_M은 본 논문에서 제안한 모형을 의미한다. DT 모형은 원 학습 데이터 전체를 모형의 학습에 활용하였으며, ISDT는 유전자 알고리즘을 이용해 원 학습 데이터 중에서 모형의 성과 측면에서 불필요한 데이터를 제외하고 핵심적인 데이터를 선택한 후 이 데이터를 이용해 의사결정 트리 단일 모형을 구축한 경우를 의미한다.

DT와 ISDT는 한 개의 분류기만을 사용하는 단일 모형이며, RS와 Proposed_M은 여러 개의 분류기를 결합하여 사용한 랜덤 서브스페이스 기반 앙상블 모형을 의미한다. 랜덤 서브스페이스 기법은 전체 입력 변수 집합에서 일부를 랜덤하게 선택함으로써 데이터를 다양화시키고 이를 통해 기저 분류기를 다양화시키는 기법이다. 또한, 이와 같이 생성된 다양화된 분류기들을 결합함으로써 성과 개선을 할 수 있게 된다. 이와 같은 앙상블 모형의 경우 랜덤하게 선택된 입력 변수의 수가 앙상블 모형의 성과에 영향을 주는 중요한 파라미터이며 본 논문에서는 이에 대한 민감도 분석을 실시하였다. 즉, 전체 입력변수의 수(F) 중에서 랜덤하게 선택한 입력변수의 수(f)에 따른 랜덤 서브스페이스 앙상블 모형과 제안 모형의 성과를 살펴보았다. 단일 모형인 DT와 ISDT인 경우 모든 입력 변수를 사용하기 때문에 f의 변화에 따른 성과 차이 비교는 앙상블 모형에만 해당된다.

단일 모형인 DT와 ISDT를 먼저 비교해보면 두 데이터 셋 모두에서 DT 모형보다 ISDT 모형이 더 좋은 예측성과를 보였다. 이것은 유전자 알고리즘을 이용한 사례 선택 기법이 모형의 성과 개선에 효과적이었다는 것을 의미한다. 랜덤 서브스페이스 기법을 기반으로 하는 두 앙상블 모형의 성과는 파라미터 f에 의해 많은 영향을 받는 것을 알 수 있다. 그림에서 보는 바와 같이 f의 값이 작을 때는 앙상블 모형의 성과가 단일 모형보다 좋지 않음을 알 수 있다. 특히, f의 값이 1일 때는 모든 앙상블 기반 모형의 성과가 가장 좋지 않은 것을 알 수 있다. 이는 앙상블 모형의 기저 분류기가 입력 변수 하나만을 가지고 구성되기 때문에 기저 분류기의 예측성과가 현저하게 떨어지게 되고 그러므로 이들의 결합을 통한 앙상블 모형의 성과 개선에 한계가 있다는 것을 의미한다. 즉, 앙상블 모형의 성과가 단일 모형보다 좋기 위해서는 기저 분류기의 성과가 어느 정도 이상은 되야 되며 이를 위해서는 f의 값이 어느 정도 커야 함을 알 수 있다. <Table 3>에서 보는 바와 같이 Kaggle 데이터의 경우 f의 값이 3보다 클 경우부터 제안한 모형의 성과가 다른 비교 모형들과 비교할 때 가장 좋은 값을 나타내는 것을 알 수 있으며 기업 재무 부실화 데이터의 경우 <Table 4>에서 보는 바와 같이 f의 값이 2보다 클 경우부터 제안한 모형의 성과가 가장 좋아지는 것을 알 수 있다.

앞에서 살펴본 바와 같이 원 입력 변수의 50% 정도를 랜덤하게 선택하여 기저 분류기를 발생시키고 앙상블 모형을 구성할 경우에 랜덤 서브스페이스 앙상블 모형의 예측성과가 최고이거나 최고에 근접한다고 알려져 있다. <Figure 2>에서 보는 바와 같이 앙상블 모형의 경우 f의 값이 중간 부분에서 대체로 가장 좋은 성과를 보이는 것을 알 수 있으며 이는 본 논문의 실험 결과도 대체로 선행 연구의 결과를 따른다고 할 수 있을 것이다.

<Figure 2>에서 보는 바와 같이 f의 값이 중간 부분에서 가장 좋은 예측성과를 보이다가 그 이후로 각 앙상블 모형의 성과가 점점 감소함을 알 수 있다. 총 F개의 입력 변수 중에서 비복원 추출 방식으로 f개의 입력 변수를 랜덤하게 선택하여 기저 분류기를 구성하였는데 f의 값이 F의 값에 근접하게 되면 기저 분류기의 평균 예측률은 높아지지만 기저 분류기 사이의 다양성 지수가 크게 낮아져서 앙상블 모형의 성과 개선 효과가 줄어들게 되기 때문이다. 극단적으로 f의 값이 F의 값과 같아지게 되면 모든 기저 분류기들이 같은 입력 변수를 사용하게 되어 모두 동일한 모형이 되게 된다. 즉, 다양성 지수가 0이 되어 앙상블 모형의 성과 개선 또한 전혀 없게 되어 단일 분류기의 예측률과 같은 결과를 얻게 되는 것이다.

본 논문에서는 앙상블 모형의 예측성과뿐만 아니라 앙상블 모형의 성과에 중요한 영향을 주는 기저 분류기들의 평균 예측률과 다양성 지수도 함께 살펴보았다. 다양성 지수로는 대표적인 다양성 척도 중의 하나인 Q-통계량을 이용하였다. 분류기 C_i와 C_j사이의 예측 오차에 대한 일치 정도가 <Table 5>와 같다고 할 때 Q-통계량은 식 (2)와 같이 구할 수 있으며 앙상블 모형의 Q-통계량 값은 각 분류기 사이의 Q-통계량의 평균값을 사용하였다[18].

$$Q(C_i, C_j) = \frac{(N_a N_d - N_b N_c)}{(N_a N_d + N_b N_c)} \quad (2)$$

<Table 5> Coincident Errors between Classifier C_i and C_j

	C_j Correct	C_j Wrong
C_i Correct	N_a	N_b
C_i Wrong	N_c	N_d

<Table 6>, <Table 7>, <Figure 3>, <Figure 4>는 앙상블을 구성하고 있는 기저 분류기들의 f의 변화에 따른 평균 예측률과 Q-통계량을 보여주고 있다. 앙상블 모형의 성과 측면에서 기저 분류기의 평균 예측률은 높을 수록 좋고, 다양성 지수 또한 높을수록 좋은 것으로 알려져

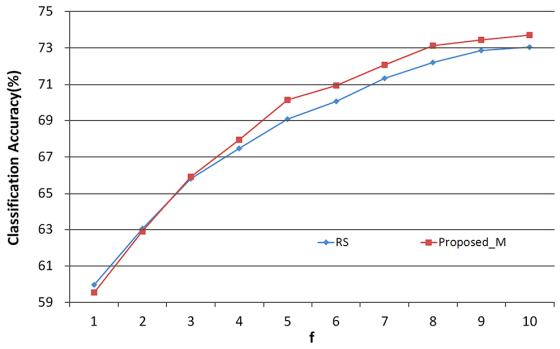
<Table 6> Performances of base Classifiers-Kaggle Data

f	Average Classification Accuracy(%)		Q-Statistic	
	RS	Proposed_M	RS	Proposed_M
1	59.96	59.53	0.342	0.395
2	63.06	62.92	0.382	0.414
3	65.82	65.92	0.444	0.452
4	67.48	67.95	0.536	0.531
5	69.09	70.14	0.628	0.615
6	70.07	70.94	0.699	0.691
7	71.33	72.07	0.785	0.793
8	72.21	73.12	0.866	0.871
9	72.87	73.45	0.925	0.938
10	73.05	73.72	1.000	1.000
Average	68.50	68.98	0.661	0.670

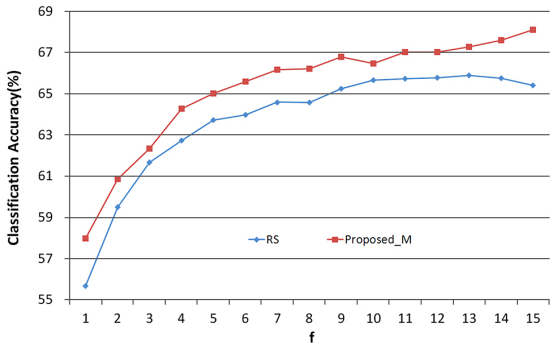
<Table 7> Performances of base Classifiers-Corporate Data

f	Average Classification Accuracy(%)		Q-Statistic	
	RS	Proposed_M	RS	Proposed_M
1	55.68	57.98	0.220	0.267
2	59.51	60.87	0.292	0.345
3	61.68	62.35	0.386	0.382
4	62.72	64.27	0.435	0.436
5	63.72	65.01	0.471	0.475
6	63.98	65.60	0.507	0.512
7	64.59	66.17	0.554	0.536
8	64.57	66.21	0.566	0.570
9	65.24	66.79	0.625	0.627
10	65.66	66.48	0.680	0.683
11	65.73	67.03	0.717	0.698
12	65.79	67.03	0.761	0.756
13	65.89	67.29	0.811	0.823
14	65.75	67.61	0.886	0.883
15	65.41	68.11	1.000	1.000
Average	63.73	65.25	0.594	0.599

있다. <Figure 3>에서 보는 바와 같이 Kaggle 데이터와 기업 데이터 모두에서 f가 증가함에 따라 기저 분류기들의 평균 예측률은 증가하는 것을 알 수 있다. 즉, 앙상블 모형의 성과 측면에서는 f의 값이 증가할수록 좋다는 것을 알 수 있다. 제안한 모형과 일반 랜덤 서브스페이스 앙상블 모형을 비교해 보면 Kaggle 데이터의 경우 f의 값이 작을 때는 일반 랜덤 서브스페이스 앙상블 모형의 기저 분류기의 평균 예측률이 높지만, f가 증가함에 따라 제안한 모형의 값이 더 좋게 나온 것을 알 수 있다. 기업 데이터의 경우는 모든 f의 값에서 제안한 모형의 값이 좋게 나온 것을 알 수 있다.

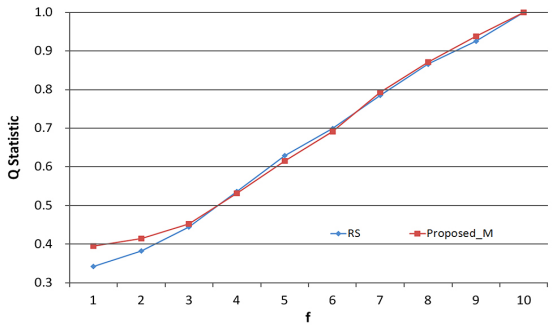


(A) Kaggle Data

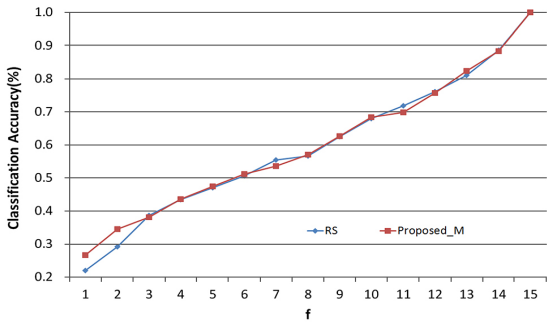


(B) Corporate Data

<Figure 3> Average Classification Accuracy of base Classifiers at Different Feature Subset Sizes



(A) Kaggle Data



(B) Corporate Data

<Figure 4> Average Q-statistic Value of base Classifiers at Different Feature Subset Sizes

<Table 8> Wilcoxon Signed-Rank Test(p-Value)

	Kaggle Data			Corporate Data		
	ISDT	RS	Propose_M	ISDT	RS	Propose_M
DT	0.038	0.005	0.05	0.005	0.005	0.005
ISDT		0.025	0.05		0.005	0.005
RS			0.015			0.005

Q-통계량의 경우도 f의 값이 증가함에 따라 값이 증가하는 것은 알 수 있다. Q-통계량의 값은 0에 가까울수록 다양성 지수가 높다는 것을 의미하며 1에 가까워진다는 것은 다양성 지수가 낮아진다는 것을 의미한다. 그러므로 Q-통계량을 기준으로 볼 때는 f의 값이 증가하는 것은 앙상블 모형의 성과에 부정적인 영향을 줄 수 있다는 것을 의미한다.

Q-통계량을 모형별로 비교해보면 제안한 모형과 일반 랜덤 서브스페이스 앙상블 모형의 값이 f의 값이 3 이후에는 거의 비슷하다는 것을 알 수 있다.

결론적으로 제안한 모형은 기존의 랜덤 서브스페이스 앙상블 모형 보다 Q-통계량 값은 거의 비슷한 수준으로 유지하면서 기저 분류기의 평균 예측률의 값은 증가한 것을 알 수 있다. 이를 통해 제안한 앙상블 모형의 성과가 기존의 랜덤 서브스페이스 앙상블 모형의 성과보다 개선되었다는 것을 유추해볼 수 있을 것이다.

제안한 모형의 성과 개선이 통계적 유의한지 알아보기 위해 윌콕슨 부호 순위 검정을 하였으며, 결과는 <Table 8>과 같다. <Table 8>에서 보는 바와 같이 제안한 모형은 다른 비교 모형보다 통계적으로 유의한 성과 개선이 있었다는 것을 알 수 있다.

6. 결 론

앙상블 분류기는 개별 모형보다 더 좋은 예측성과를 얻기 위해 개별적으로 훈련된 서로 다른 분류기들을 결합하는 것이다. 이와 같은 앙상블 분류기는 분류기들의 일반화 성능을 향상시키는 것으로 알려져 있으며, 이로 인해 최근에 다양한 분야에서 많이 활용되고 있다.

앙상블 분류기의 좋은 예측성과로 다양한 연구가 진행되고 있지만 아직까지 사례 선택 기법과 랜덤 서브스페이스 앙상블 기법을 결합하는 연구는 거의 없는 것이 현실이다. 이에 본 논문에서는 랜덤 서브스페이스 앙상블 모형의 예측 성능을 향상시키기 위해 사례 선택 기법과 랜덤 서브스페이스 앙상블 모형을 결합하는 새로운 형태의 앙상블 모형을 제안하였다. 사례 선택 기법과 랜덤 서브스페이스 기법은 각각 데이터 마이닝 분야에서 자주 활용되고 있는 기법으로 다양한 응용 분야에서 효과적으로 적용되고 있으며 각각 모형의 예측성과에 기여할 수 있는 잠재력이 있는 기법이다. 그러나, 아직까지 이들 두 기법의 결합에 대한 연구는 거의 없는 것이 현실이다. 이에 본 연구에서는 두 개 기법의 장점을 활용하여 모형의 성과를 개선하고자 두 기법을 결합하는 새로운 형태의 앙상블 모형을 제안하였으며 제안한 모형의 성과를 검증하기 위해 개인과 기업의 재무 부실화 예측 문제에 적용해 보았다. 본 논문에서 제안한 모형을 재무 부실화 문제에 적용한 결과 제안한 모형이 기존의 모형보다 예측성과 면에서 우수한 성과를 보임을 알 수 있었다.

본 연구에서는 제안한 모형을 재무 부실화 예측 문제에 적용해 보았지만 향후 다양한 분류 문제에도 적용될 수 있을 것으로 기대된다. 본 연구에서는 유전자 알고리즘을 이용한 사례 선택 기법을 랜덤 서브스페이스 기법과 결합하였다. 하지만, 향후 다양한 사례 선택 기법과 앙상블 기법과의 결합에 대한 다양한 연구도 필요할 것으로 생각된다.

References

[1] Abellan, J. and Mantas, C.J., Improving Experimental Studies about Ensembles of Classifiers for Bankruptcy Prediction and Credit Scoring, *Expert Systems with Applications*, 2014, Vol. 41, No. 8, pp. 3825-3830.

[2] Altman, E.L., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, 1968, Vol. 23, No. 4, pp. 589-609.

[3] Beaver, W., Financial ratios as predictors of failure, empirical research in accounting : Selected studied, *Journal of Accounting Research*, 1966, Vol. 4, No. 3, pp. 71-111.

[4] Bian, S. and Wang, W., On diversity and accuracy of homogeneous and heterogeneous ensembles, *International Journal of Hybrid Intelligent Systems*, 2007, Vol. 4, No. 2, pp. 103-128.

[5] Breiman, L., Bagging predictors, *Machine Learning*, 1996, Vol. 24, No. 2, pp. 123-140.

[6] Bryant, S.M., A case-based reasoning approach to bankruptcy prediction modeling, *International Journal of Intelligent Systems in Accounting*, *Finance and Management*, 1997, Vol. 6, No. 3, pp. 195-214.

[7] Derrac, J., Cornelis, C., García, S., and Herrera, F., Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection, *Information Sciences*, 2012, Vol. 186, No. 1, pp. 73-92.

[8] Dietterich, T.G., Machine-learning research : Four current directions, *AI Magazine*, 1997, Vol. 18, No. 4, pp. 97-136.

[9] Freund, Y. and Schapire, R., Experiments with a new boosting algorithm, Proceedings of the 13th, *International Conference on Machine learning*, 1996, pp. 148-156.

[10] Garcia, V., Marques, A.I., and Sanchez, J.S., On the use of data filtering techniques for credit risk prediction with instance-based models, *Expert Systems with Applications*, 2012, Vol. 39, No. 18, pp. 13267-13276.

[11] Goldberg, D.E., *Genetic algorithms in search*, *optimization and machine learning*, New York : Addison-Wesley, 1989.

[12] Hart, P.E., The condensed nearest neighbor rule, *IEEE Transactions on Information Theory*, 1968, Vol. 14, pp. 515-516.

[13] Ho, T.K., The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, Vol. 20, No. 8, pp. 832-844.

[14] Hung, C. and Chen, J.-H., A Selective Ensemble Based on Expected Probabilities for Bankruptcy Prediction, *Expert Systems with Applications*, 2009, Vol. 36, No. 3, pp. 5297-5303.

[15] Kim, K.-J. and Ahn, H., Optimization of Support Vector Machines for Financial Forecasting, *Journal of Intelligence and Information Systems*, 2011, Vol. 17, No. 4, pp. 241-254.

[16] Kim, M. and Kang, D., Ensemble with neural networks for bankruptcy prediction, *Expert System with Applications*, 2010, Vol. 37, No. 4, pp. 3373-3379.

[17] Kim, M., Kang, D., and Kim, H.B., Geometric Mean Based Boosting Algorithm with over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction, *Expert Systems with Applications*, 2015, Vol. 42, No. 3, pp. 1074-1082.

[18] Kuncheva, L.I. and Whitaker, C.J., Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning*, 2003, Vol. 51, No. 2, pp. 181-207.

[19] Li, H., Lee, Y.-C., Zhou, Y.-C., and Sun, J., The random subspace binary logit (RSBL) model for bankruptcy prediction, *Knowledge-Based Systems*, 2011, Vol. 24, No. 8, pp. 1380-1388.

[20] Louzada, F., Anacleto-Junior, O., Candolo, C., and Mazucheli, J., Poly-bagging predictors for classification modelling for credit scoring, *Expert Systems with Applications*, 2011, Vol. 38, No. 10, pp. 2717-12720.

[21] Marques, A.I., Garcia, V., and Sanchez, J.S., Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles, *Expert Systems with Applications*, 2012, Vol. 39, No. 11, pp. 10244-10250.

[22] Messier, W. and Hansen, J., Inducing rules for expert system development : an example using default and bankruptcy data, *Management Science*, 1998, Vol. 34, No. 12, pp. 1403-1415.

[23] Meyer, P.A. and Pifer, H., Prediction of bank failures, *The Journal of Finance*, 1970, Vol. 25, pp. 853-868.

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

[25] Nanni, L. and Lumini, A., An Experimental Comparison of Ensemble of Classifiers for Bankruptcy Prediction and Credit Scoring, *Expert Systems with Applications*, 2009, Vol. 36, No. 2, pp. 3028-3033.

[26] Ohlson, J., Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, 1980, Vol. 18, No. 1, pp. 109-131.

[27] Park, K.-J., Simulation Optimization of Manufacturing System using Real-coded Genetic Algorithm, *Journal of Society of Korea Industrial and Systems Engineering*, 2005, Vol. 28, No. 3, pp. 149-155.

[28] Tam, K. and Kiang, M., Managerial applications of neural networks : the case of bank failure predictions, *Management Science*, 1992, Vol. 38, No. 7, pp. 926-947.

[29] Tsai, C. and Wu, J., Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring, *Expert Systems with Applications*, 2008, Vol. 34, No. 4, pp. 2639-2649.

[30] Wang, G. and Ma, J., A hybrid ensemble approach for enterprise credit risk assessment based on Support Vec-

tor Machine, *Expert Systems with Applications*, 2009, Vol. 39, No. 5, pp. 5325-5331.

[31] www.kaggle.com/c/GiveMeSomeCredit (Give Me Some Credit).

[32] Yoo, J., Release Planning in Software Product Lines Using a Genetic Algorithm, *Journal of Society of Korea Industrial and Systems Engineering*, 2012, Vol. 35, No. 4, pp. 142-148.

[33] Yum, C.-S. and Lee, H.-J., Economic Design of Local Area Networks using Genetic Algorithms, *Journal of Society of Korea Industrial and Systems Engineering*, 2005, Vol. 28, No. 2, pp. 101-108.

[34] Yum, J.K., Nam, K.S., A Study of D-Optimal Design in Nonlinear Model Using the Genetic Algorithm, *Journal of the Korean Society for Quality Management*, 2000, Vol. 28, No. 2, pp. 135-146.

[35] Zhang, G., Hu, Y.M., Patuwo, E.B., and Indro, C.D., Artificial neural networks in bankruptcy prediction : general framework and cross-validation analysis, *European Journal of Operational Research*, 1999, Vol. 116, pp. 16-32.

ORCID

Sung-Hwan Min | http://orcid.org/0000-0003-3931-4376