韓國IT서비스學會誌
第14卷 第4號
2015年 12月, pp.121-135

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2015.14.4.121

재무부실화 예측을 위한 랜덤 서브스페이스 앙상블 모형의 최적화*

민 성 환**

Optimization of Random Subspace Ensemble for Bankruptcy Prediction*

Sung-Hwan Min**

■ Abstract ■

Ensemble classification is to utilize multiple classifiers instead of using a single classifier. Recently ensemble classifiers have attracted much attention in data mining community. Ensemble learning techniques has been proved to be very useful for improving the prediction accuracy. Bagging, boosting and random subspace are the most popular ensemble methods. In random subspace, each base classifier is trained on a randomly chosen feature subspace of the original feature space. The outputs of different base classifiers are aggregated together usually by a simple majority vote. In this study, we applied the random subspace method to the bankruptcy problem. Moreover, we proposed a method for optimizing the random subspace ensemble. The genetic algorithm was used to optimize classifier subset of random subspace ensemble for bankruptcy prediction. This paper applied the proposed genetic algorithm based random subspace ensemble model to the bankruptcy prediction problem using a real data set and compared it with other models. Experimental results showed the proposed model outperformed the other models.

Keyword : Random Subspace, Ensemble, Bankruptcy Prediction, Genetic Algorithms

Submitted : April 27, 2015 1st Revision : December 1, 2015 Accepted : December 9, 2015
* 이 논문은 2015년도 한림대학교 교비학술연구비(HRF-201501-007)에 의해 연구되었음.
** 한림대학교 경영학부

1. 서 론

기업의 부도는 종업원, 투자자 등과 같은 이해 관계자뿐만 아니라 국가 경제 전반에도 영향을 미칠 수 있는 중요한 문제이다. 금융기관으로서는 정확한 부도 예측을 통해 기업 부도로 인해 발생할 수 있는 비용을 줄이고 위험을 최소화하는 것이 매우 중요한 일이라 할 수 있다. 기업의 부도 예측은 재무나 회계 분야에서 오랜 기간 연구되어 온 중요한 주제이다. 초기의 부도 예측에 관한 모형은 단일변량 분석(Beaver, 1966), 다변량 판별분석(Altman, 1968), 다중 회귀분석(Meyer and Pifer, 1970), 로지스틱 회귀분석(Ohlson, 1980) 등과 같이 주로 전통적인 통계 모형에 기반을 둔 모형이 대부분이었으나 최근에는 인공신경망(Zhang et al., 1999), 사례기반 추론(Buta, 1994) 등과 같은 다양한 인공지능 기법이 활용되고 있다.

앙상블(ensemble) 모형이란 단일 분류기보다 더 좋은 성과를 내기 위해 여러 개의 분류기들을 결합하는 것을 의미한다. 여러 개의 분류기들을 적절하게 결합할 경우 단일 분류기보다 더 좋은 성과를 내는 것으로 알려졌으며, 최근에 이와 같은 앙상블 모형에 대한 연구가 활발하게 진행되고 있다(Dietterich, 1997; Kuncheva, 2004). 앙상블 분류기의 성능에 중요한 영향을 미치는 요소로는 앙상블을 구성하고 있는 기저 분류기(base classifier)들의 예측 성과와 이들 간의 다양성(diversity)을 들 수 있다. 일반적으로 서로 다른 분류기는 각각 서로 다른 예측 오차를 보이며, 이와 같은 다양한 분류기들을 결합함으로써 보다 정확한 예측률을 보이는 앙상블 분류기를 생성할 수가 있게 된다. 앙상블 분류기의 성과 개선을 위해서는 무엇보다도 서로 다른 예측 오차를 보이는 분류기들을 생성하고 이들을 결합하는 것이 필요하다. 만약에 기저 분류기들 사이에 다양성이 존재하지 않는다면 앙상블 모형을 통한 성능 개선은 기대할 수 없게 된다(Bian and Wang, 2007; Kuncheva and Whitaker, 2003). 배깅(bagging)과 부스팅(boosting), 랜덤 서브스페이스(random subspace) 등은 다양성이 존재하는 서로 다른 분류기를 생성하기 위한 대표적인 앙상블 기법이다.

최근에는 이와 같은 앙상블 모형을 부도 예측 문제에 적용하려는 연구가 활발하게 진행되고 있다. Kim and Kim(2007)은 의사결정 트리의 일종인 CART를 기저 분류기로 하는 변형된 배깅을 SOHO의 부도 예측 문제에 적용해 보았다. Kim(2009)은 배깅과 부스팅을 기업의 부도 예측에 적용해 보았으며, 실험 결과 의사결정 트리, 인공 신경망의 단일 모형모다 이들을 결합한 앙상블 모형이 성과가 좋음을 알 수 있었다. Kim(2010)은 인공신경망을 기저 분류기로 하는 배깅과 부스팅의 성과 개선을 위해 유전자 알고리즘(Genetic Algorithms)을 이용하였다. Shin and Hong(2011)은 SVM을 기저 분류기로 하는 AdaBoost 모형을 기업 신용평가 문제에 적용해 보았다. Li et al.(2011)은 로짓 모형을 기저 분류기로 하는 랜덤 서브스페이스 모형을 여러 단일 모형과 비교 분석하였다. Louzada et al.(2011)은 폴리 배깅(poly-bagging)이라는 새로운 형태의 배깅 변형 모형을 제안하여 신용평가 문제에 적용해 보았다. Min(2014)은 배깅과 사례 선택(instance selection)을 결합하는 새로운 모형을 부도 예측 문제에 적용해 보았다. Abellan and Mantas(2014)는 배깅 앙상블 모형의 기저 분류기로 크리덜 의사결정 트리(credal decision tree) 모형을 사용할 때 좋은 성과가 난 다는 것을 보였다. Kim et al.(2015)은 기하 평균을 활용한 부스팅 알고리즘을 제안하였으며, 제안한 모형이 부도 예측 문제에서 다른 경쟁 모형 비해 좋은 예측 성과를 보였다.

이와 같이 최근에 각광 받고 있는 앙상블 모형을 부도 예측 문제에 적용하려는 다양한 연구가 있었지만 아직까지 랜덤 서브스페이스 모형을 국내 기업의 부도 예측 문제에 적용해 본 연구는 많지 않은 것이 현실이다. 또한, 랜덤 서브스페이스 앙상블 모형의 성과는 기저 분류기의 파라미터 값에 많은 영향을 받지만 이에 대한 분석은 거의 없

는 것이 현실이다.

이에 본 연구에서는 랜덤 서브스페이스 앙상블 모형을 국내 기업의 부도 예측 문제에 적용해 보았으며 기저 분류기의 파라미터와 앙상블 모형의 성과와의 관계에 대한 분석을 시도하였다. 또한, 앙상블 모형의 성능 개선을 위해 유전자 알고리즘을 이용한 랜덤 서브스페이스 앙상블의 최적화 모형을 제안하였으며 실제 국내 기업의 부도 관련 데이터를 이용해 제안한 모형의 우수성을 검증하였다.

본 논문의 구성은 다음과 같다. 다음 장에서는 앙상블 분류기에 대한 설명을 하고, 제 3장에서는 본 논문에서 제안한 모형에 대한 설명을 하였다. 제 4장에서는 본 연구에서 제안한 모형의 검증을 위한 실험 설계에 대한 설명을 하고 제 5장에서는 실험결과에 대해 서술하였다. 마지막 장에서는 요약 및 향후 연구 과제에 대해 설명하였다.

2. 앙상블 분류기

앙상블 분류기란 단일 분류기보다 더 좋은 성과를 내기 위해 여러 개의 분류기들을 결합하는 것을 의미하며 최근에 데이터 마이닝, 기계학습 분야에서 많은 관심을 끌고 있다. 앙상블 분류기를 구성하는 개별 분류기를 기저 분류기라고 부르며, 대부분의 경우에 앙상블 분류기는 그것을 구성하고 있는 기저 분류기보다 더 좋은 성과를 보이는 것으로 알려져 있다(Dietterich, 1997).

앙상블 분류기는 서로 다른 다수의 기저 분류기들의 출력 정보를 결합하여 최종적으로 분류를 하는 분류기를 의미한다. Ho(2002)는 앙상블 분류기의 성과를 최적화하는 방법을 두 가지로 분류하고 있다. 하나는 최적의 앙상블을 구성하기 위해 주어진 서로 다른 기저 분류기 풀(classifiers pool)로부터 어떤 분류기들을 선택할까하는 문제이다. 또 다른 접근 방법은 이미 선택된 기저 분류기들의 출력 정보(output)를 결합하기 위한 규칙을 최적화 하는 것이다. 전자는 최적의 분류기를 선택하는 문제로 범위 최적화(coverage optimization)라고 하며, 후자는 결정 최적화(decision optimization)라고 한다.

앙상블 분류기가 기저 분류기보다 좋은 성과를 내기 위해서는 개별 기저 분류기들의 예측 성과뿐만 아니라 기저 분류기들 간의 다양성(diversity)이 중요한 요소이다. 기저 분류기들의 예측 결과값이 모두 똑같은 값을 갖는다면 기저 분류기들 간의 다양성이 존재하지 않는다고 말할 수 있으며, 반대의 경우는 다양성이 존재한다고 말할 수 있다. 만약 앙상블을 구성하고 있는 기저 분류기들이 다양성이 전혀 존재하지 않고 모두 똑같다면 이들의 결합으로 이루어진 앙상블 모형은 개별 분류기보다 좋은 성과를 기대할 수 없을 것이다(Kuncheva and Whitaker, 2003).

기저 분류기들을 다양화시키는 방법으로는 학습 데이터에 변화를 주는 방법, 학습 파라미터에 변화를 주거나 서로 다른 학습 알고리즘을 사용하는 방법 등 여러 가지가 있다. 이 중에서 가장 대표적인 방법은 학습 데이터를 다양화시킴으로써 기저 분류기들을 다양화시키는 방법이며 배깅, 부스팅과 랜덤 서브스페이스 기법이 이에 속한다.

배깅(bagging)은 bootstrap aggregating의 약자로 단순하고 적용하기 쉬우면서도 매우 좋은 성과를 내는 것으로 알려져 있다(Brieman, 1996). 배깅은 기저 분류기를 다양화시키기 위해 학습 데이터를 복원 추출 방법에 의해 N번 랜덤하게 샘플링을 하여 N개의 학습 데이터를 생성시킨 후, 이들 각각의 서로 다른 학습 데이터를 이용하여 각각의 기저 분류기를 학습시킨다. 이와 같은 방법으로 다양성을 갖는 서로 다른 다수의 기저 분류기가 생성되게 되며, 이들 기저 분류기들의 결과값(outputs)은 다수결 투표(majority vote)와 같은 방식에 의해 결합되게 된다.

부스팅은 학습 데이터에 변화를 줌으로써 기저 분류기를 다양화시킨 다는 점에서 배깅과 유사하지만, 앞 단계에서 생성된 기저 분류기의 예측 성과에 따라 다음 단계에서 데이터가 선택될 확률이

영향을 받게 된다는 차이가 있다(Freund et al., 1996).

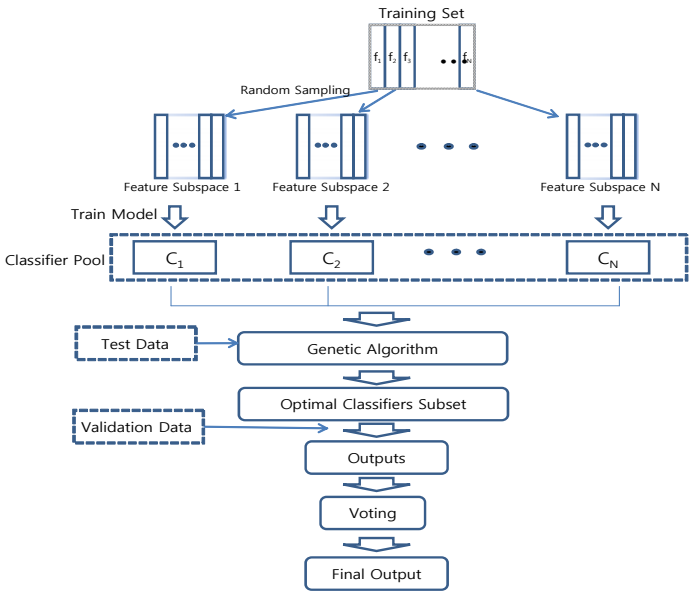
랜덤 서브스페이스 기법은 Ho(1998)에 의해 제안된 앙상블 구축 기법으로 배깅에서처럼 학습 데이터에 변화를 줌으로써 기저 분류기의 다양성을 확보하고자 하는 접근방법이다. 하지만, 랜덤 서브스페이스 기법은 배깅과 달리 입력변수 집합(feature set)으로부터 랜덤하게 선택한 서로 다른 입력변수 부분집합(feature subspace)을 구성한 다음 이들 각각을 이용해 서로 다른 분류기를 학습시키고, 이들의 결과를 통합하는 방법이다. 배깅과 랜덤 서브스페이스 기법과의 차이는 배깅은 사례 공간(instance space)에서의 변화를 주는 것이고 랜덤 서브스페이스는 입력변수 공간(feature space) 상에서의 변화를 주는 것이다.

3. 연구 모형

본 연구에서는 탐색적 연구를 통해 랜덤 서브스페이스 앙상블 모형과 기저 분류기의 관계에 관한 심층적인 분석을 수행하였다. 또한, 기존의 랜덤 서브스페이스 앙상블 모형의 성능 개선을 위해 유전자 알고리즘을 이용한 최적화 모형을 제안하였다. 유전자 알고리즘은 경영분야 뿐만 아니라, 공학분야 등 다양한 분야의 최적화 문제 해결에 성공적으로 적용되어 왔다(Kim, 2010; Kim and Lee, 2011; Lee et al., 2012). 본 연구에서는 랜덤 서브스페이스 기법을 통해 생성된 기저 분류기 풀(classifiers pool)에서 최적의 기저 분류기 조합을 선택하기 위해 유전자 알고리즘이 활용되었다.

<Figure 1>은 본 논문에서 제안한 모형의 전반적인 절차를 보여주고 있다. <Figure 1>에서 보는 바와 같이 학습 데이터의 입력변수 집합으로부터 랜덤하게 선택한 서로 다른 입력변수 부분집합을 구성하고 이들 각각의 서로 다른 학습 데이터 셋을 이용하여 다양성을 갖는 서로 다른 다수의 기저분류기 셋 $\{C_1, \cdots, C_n\}$을 생성한다. 일반적인 랜덤 서브스페이스 앙상블 모형의 경우 위의 단계에서 생성된 분류기 셋 $\{C_1, \cdots, C_n\}$을 가지고 앙상블을 구성하게 되며, 개별 분류기의 출력 값을 특정한 전략에 따라 결합하게 된다.

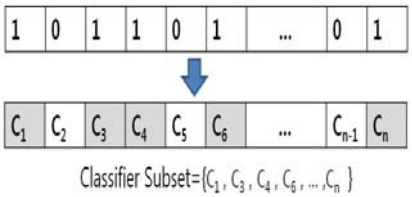
본 연구에서는 일반적인 랜덤 서브스페이스 앙



〈Figure 1〉 The Overall Architecture of the Proposed Model

상블 모형의 성능 개선을 위해 유전자 알고리즘을 이용한 랜덤 서브스페이스 앙상블 모형을 제안하였다. 본 연구에서 제안한 모형은 일반적인 랜덤 서브스페이스 앙상블 모형의 성능 개선을 위해 유전자 알고리즘을 적용하였으며, 이를 통해 기저분류기 풀에서 최적의(또는 근사 최적의) 분류기 집합을 찾아내고, 최적의 분류기 집합으로 구성된 앙상블 모형을 최종 모형으로 사용하였다.

본 연구에서는 랜덤 서브스페이스의 기법으로 생성된 다양한 분류기 풀 중에서 앙상블 분류기의 성과를 최적화하는 분류기를 선택하는 범위 최적화 문제를 해결하기 위해 최적화 문제에서 많이 활용되고 있는 유전자 알고리즘을 활용하였다.



〈Figure 2〉 Encoding for GA

<Figure 2>는 최적의 앙상블 분류기 선택을 위해 사용된 유전자 알고리즘의 염색체 구조를 나타내고 있다. <Figure 2>에서 C_1은 기저 분류기 풀에서 첫 번째 기저 분류기를 의미하고 C_2는 두 번째 기저 분류기를 의미한다. n개의 기저 분류기 풀 중에서 최적의 분류기 조합을 선택하기 위해 유전자 알고리즘을 사용하였으며, 분류기 조합에 대한 염색체는 0과 1의 값을 갖는 이진열(binary string) 형태로 표현하였다. 유전자 알고리즘 염색체의 비트는 분류기의 총 수와 같은 n개의 비트(bit)로 구성되어 각각 분류기와 대응되도록 설계하였다. 각각의 비트는 해당되는 분류기의 선택 유무를 알려주게 된다. 즉 각각의 비트에서 1의 값은 해당되는 기저 분류기가 선택되었다는 것을 의미하고 0의 값은 해당되는 분류기가 선택되지 않았다는 것을 의미한다. <Figure 2>에서 C_1에 해당하는 염색체 비트의 값은 1로 이는 첫 번째 분류기 C_1가 선택되었다는 것을 의미하고 C_2에 해당하는 염색체 비트의 값은 0으로 이는 두 번째 분류기 C_2가 선택되지 않았다는 것을 의미한다.

<Table 1>은 랜덤 서브스페이스 앙상블의 전반적인 절차를 보여주고 있다. 본 논문에서는 랜덤 서브스페이스 모형의 성과 개선을 위해 유전자 알고리즘을 이용해 최적의 기저분류기 조합을 선택하는 앙상블 모형을 제안하였다. 본 연구에서 제안한 모형의 전반적인 흐름은 <Table 2>에 나와 있다.

유전자 알고리즘에서 적합도 함수로는 예측 정확도를 사용하였으며, 과적합을 피하기 위해 사용한 테스트용 데이터 셋에서의 예측률을 구하여 사용하였다. 적합도 함수로 사용한 예측 정확도는 식 (1)과 같이 구할 수 있다. 여기서 N은 테스트용 데이터의 총 데이터 수를 의미하고, 테스트용 데

〈Table 1〉 Steps of Random Subspace Method

Random Subspace Ensemble
1. Partition Data Set(Training Data Set(T_1), Validation Data Set(V))
2. Generate a new training data set with randomly selected f' features from T_1 (f' < F(total number of features in T_1))
3. Repeat step 2 to generate n new training data sets ==> $T_1(RS)_1$, $T_1(RS)_2$, …, $T_1(RS)_n$
4. Train a classifier for each new training set (Different classifiers are generated) ==> C_1, …, C_n
5. Apply the classifiers generated in step 4 to the validation data set (n different output data) ==> O_1, …, O_n
6. Combine the output data(O_1, …, O_n) by a combining method (In this paper, we use majority voting scheme as a combining method)

〈Table 2〉 Steps of the Proposed Model

GA based Random Subspace Ensemble
1. Partition Data Set(Training Data Set(T_1), Test Data Set(T_2), Validation Data Set(V))
2. Generate a new training data set with randomly selected f' features from T_1 (f' < F(total number of features in T_1))
3. Repeat step 2 to generate n new training data sets ==> $T_1(RS)_1$, $T_1(RS)_2$, …, $T_1(RS)_n$
4. Train a classifier for each new training set (Different classifiers are generated) ==> C_1, …, C_n
5. Define the chromosome (The chromosome for the classifiers pool is encoded as a form of binary string)
6. Determine parameters of GA
7. Generate the initial population
8. Select the classifiers pool for each chromosome
9. Apply the classifiers pool generated in step 8 to the test data set(T_2)
10. Combine the output data by a combining method ==> Calculate the fitness values
11. Perform GA operations and create a new generation
12. Repeat from 8 to 10 until the termination criteria are satisfied
13. Select the optimal classifiers pool
14. Apply the optimal classifiers pool to the validation data set(V)
15. Combine the output data(O_1, …, O_n) by a combining method (In this paper, we use majority voting scheme as a combining method)

이터 중 i번째 데이터에서의 예측치가 실제값과 일치할 경우 H_i 값은 1의 값을 갖고, 그렇지 않을 경우 0의 값을 갖는다.

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

본 연구에서는 앙상블 모형의 기저 분류기로 k 최근접 이웃(k nearest neighbor : KNN) 알고리즘을 사용하였다. KNN 기법은 단순하지만 매우 효과적인 기법으로 데이터마이닝 분야에서 많이 사용되고 있다. KNN은 가장 가까운 이웃(neighbor)의 학습 데이터의 정보를 바탕으로 분류를 하는 알고리즘으로 데이터의 변화에 매우 강건한(robust) 것으로 알려져 있으며, 반면에 입력 변수의 변화에는 매우 민감한(sensitive)것으로 알려져 있다. 그러므로, KNN을 기저 분류기로 사용할 경우에는 입력변수에 변화를 줌으로써 다양성이 존재하는 KNN을 생성할 수 있게 되어 앙상블의 성능 개선을 기대할 수 있게 된다.

가까운 이웃을 찾기 위해 사용되는 거리 척도(distance measure)로는 코사인 거리(cosine distance), 유클리디안 거리(Eulidean distance) 등이 가장 많이 사용되고 있다. 본 연구에서는 사례간의 거리를 계산하기 위해 유클리디안 거리를 사용하였으며 두 사례 a, b 간의 유클리디안 거리를 구하는 식은 아래 식 (2)와 같다. 식에서 d_{ab}는 사례 a와 b의 유클리디안 거리 값을 의미한다. 또한, f는 입력변수의 총 수를 의미하며, x_{ai}는 사례 a의 i번째 입력변수에 해당하는 값을 의미하고, x_{bi}는 사례 b의 i번째 입력변수에 해당하는 값을 의미한다. w_i는 i번째 입력변수에 대한 가중치를 의미한다. 본 연구에서는 모든 입력변수에 대해 같은 가중치(equal weighting)를 부여하여 거리를 계산하였다.

$$d_{ab} = \sqrt{\sum_{i=1}^{f} w_i \cdot (x_{ai} - x_{bi})^2} \quad (2)$$

4. 실험 설계

본 연구에서 제안한 모형의 검증을 위해 국내 비외감 기업의 데이터를 사용하여 실험하였다. 데이터는 총 1,800개로 구성되어 있으며 부도 기업의 데이터와 비부도 기업의 데이터는 같은 수인 900개로 구성되어 있다. 데이터는 학습용 데이터(training data)와 테스트용 데이터(test data), 그리고 검증용 데이터(validatiaon data)로 분류하여 실험을 하였다. 학습용 데이터는 모형의 학습을 위한 데이터로 이용하였으며, 테스트용 데이터는 유전자 알고리즘을 이용한 최적의 분류기를 선택할 때 과적합(overfitting)을 피하기 위해 사용되었으며, 최종적으로 모형의 검증을 위해 검증용 데이터를 사용하였다.

본 연구에서 수행한 실험은 크게 탐색적 연구를 위한 실험과 모형의 검증을 위한 실험으로 분류할 수 있다. 우선, 탐색적 연구를 위한 실험에서는 학습용 데이터와 테스트용 데이터를 6대 4의 비율로 사용하여 실험하였다. 하지만, 본 연구에서 제안한 최적화 모형의 검증을 위해서는 보다 정확한 분석을 위해 10-겹 검증(10-fold cross validation) 방법으로 실험을 하였으며, 전체 표본 수 1,800개를 동일한 수(180개)의 10개의 fold로 나눈 후, 9개의 fold는 학습용과 테스트용 데이터로 활용하고 나머지 1개의 fold는 검증용 데이터로 활용하였다.

랜덤 서브스페이스와 본 논문에서 제안한 모형의 실험의 경우 각각의 검증용 데이터에 대해 10회 반복하여 실험을 수행하였으며 실험결과는 10회 실험 결과의 평균값을 대푯값으로 사용하였다.

기업의 부도 여부를 예측하기 위해 재무비율을 입력 변수로 사용하였다. 본 연구에서는 수익성, 안정성, 성장성, 활동성 및 현금 흐름으로 분류된 총 131개의 재무비율을 대상으로 단일 표본 t검정(independent-samples t-test)을 실시하여 p-value 값이 0.05보다 큰 변수는 제외하고 나머지 변수를 대상으로 후진 선택법(backward selection)을 이용한 로지스틱 회귀분석을 이용하여 최종 변수를 선정하였으며 그 결과는 <Table 3>에 나와 있다.

〈Table 3〉 Input Variables

Category	Description
Profitability	EBITDA to Sales
	Financial Expenses to Sales
	Financial Expenses to Debt
	Ordinary Income Rate
	Ordinary Income to Sales
	Net income to sales
	Interest Expenses to Net income
	Ordinary Income to Capital
	Ordinary Income to Total Asset
Stability	Fixed Asset to Owener's Equity
	Quick Asset to Current Liability
	Debt Ratio
	Current Liabiltiy to Total Asset
	Current Asset to Current Liabiltity
	Capital suplus+retained (earnings-divdend)/total assets
	Borrowings to EBITDA
	Borrowings to Sales
	Cash Ratio
Growth	Coefficient of variation of sales
Cash Flow	Cash flow after interest payment to sales
	Cash Flow to Financial Expenses
Activity	Sales to net change in working capital
	Total assets turnover period
	Sales to net change in account receivable

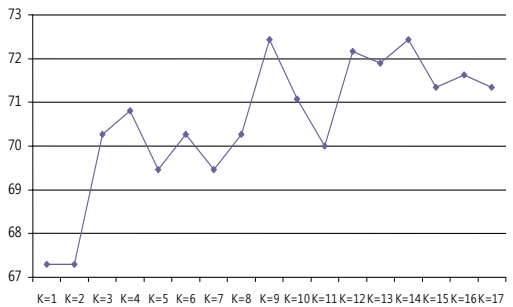
5. 실험 결과

5.1 탐색적 분석

본 연구에서는 앙상블 모형의 기저 분류기로 KNN을 사용하였다. KNN의 성과는 파라미터 k에 따라 크게 영향을 받으므로 본 연구에서는 예비 실험을

통해 k와 단일 모형의 성과와의 관계를 살펴 보았다. k에 대한 민감도 분석 결과는 <Figure 3>에 나와 있다. <Figure 3>에서 보는 바와 같이 k에 따라 단일 모형 KNN의 성과가 크게 차이가 나는 것을 알 수 있으며, k = 9일 때 가장 좋은 결과를 보임을 알 수 있다.

랜덤 서브스페이스 모형의 성과에 영향을 주는 대표적인 파라미터로는 앙상블을 구성하는 기저 분류기의 총 수와 선택된 입력변수의 수가 있다. 본 연구에서는 실험의 복잡성을 줄이기 위해 모형의 수는 100으로 고정하고 실험을 하였으며 전체 입력변수의 수(F = 24) 중에서 랜덤하게 선택한 입력변수의 수(f)에 따른 성과를 비교 분석하였으며 그 결과는 <Figure 4>와 같다. 여기서 KNN은 k = 9인 KNN의 단일 모형을 의미하며, RS는 k = 9인 KNN을 기저 분류기로 사용하는 랜덤 서브스페이스 앙상블 모형을 의미한다. BC는 랜덤 서브스페이스 앙상블 모형을 구성하고 있는 기저 분류기들을 의미한다. <Figure 4>에서의 값은 각 모형에서의 예측률을 의미하며, BC에서의 값은 랜덤 스페이스 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률을 의미한다. 단일 모형인 KNN은 <Table 3>에 있는 모든 입력 변수를 사용하여 모형을 구성하였으며, 랜덤 서브스페이스 모형의 경우 전체 입력 변수(F = 24) 중에서 램덤하게 f개 만큼 선택하여 기저 분류기를 구성한 후 이들을 다수결 투표(majority vote) 방식에 의해 결합하였다. 본

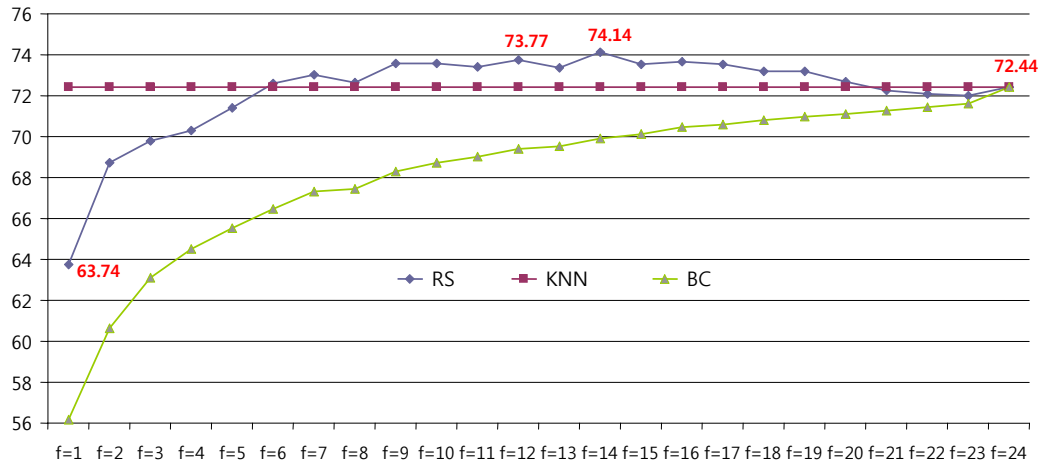


〈Figure 3〉 The Classification Accuracy of the KNN Model at Different k Values(%)

실험에서는 비복원 추출 방식을 이용하여 랜덤하게 입력 변수를 선택하였다. 앞에서 살펴 본 바와 같이 랜덤 서브스페이스 앙상블 모형의 성과는 선택된 입력 변수의 수(f)에 영향을 받으며 그 결과는 <Figure 4>와 <Table 4>와 같다.

실험 결과에서 알 수 있듯이 기저 분류기들의 평균 예측률은 f의 값이 커질수록 증가하는 것을 알 수 있다. <Figure 4>에서 RS의 값이 BC의 값보다 큰 것을 알 수 있으며 이는 랜덤 서브스페이스 앙상블 모형의 효과라고 할 수 있다. 즉, 기저 분류기들을 결합함으로써 기저 분류기 평균 예측률 보다 성과가 좋은 앙상블 모형을 구성할 수 있음을 알 수 있다. f = 1인 경우, 즉 입력 변수 한 개로만 기저 분류기를 구성하는 경우에는 앙상블 모형의 예측률이 63.74%로 모든 입력 변수를 사용한 단일 모형보다 성과가 좋지 않음을 알 수 있다. 한 개의 입력 변수만을 사용한 기저 분류기들의 평균 예측률은 56.17%로 24개 전체 입력 변수를 모두 사용한 단일 모형의 예측 성과보다 매우 낮기 때문에 이들의 결합을 통한 앙상블 모형의 성과 개선이 있었지만 전체 변수를 사용한 단일 모형보다는 성과가 좋지 않았다. 랜덤 서브스페이스 앙상블 모형에서 비복원 추출로 기저 분류기를 구성했으므로 f = 24인 경우는 모든 변수를 사용한 동일한 기저 분류기들로 구성되어 있으며 이들의 평균 예측률은 단일 모형인 KNN의 값과 같음을 알 수 있다. 또한, 모두 동일한 기저 분류기들을 결합할 경우 전혀 성과 개선이 없음을 알 수 있다. 이와 같이, 앙상블 모형의 성과는 기저 분류기들의 평균 예측률과 다양성이 중요함을 알 수 있다.

<Figure 4>에서 보는 바와 같이 f가 6 이상인 경우부터 랜덤 서브스페이스 앙상블 모형의 성과가 KNN 단일 모형보다 더 좋아 지는 것을 알 수 있다. Ho(1998)의 실험에 의하면 f의 값이 전체 입력 변수 총수의 0.5에 해당할 경우 가장 좋은 결과 또는 가장 좋은 결과와 근사한 결과를 냈다고 보고하고 있다. <Figure 4>에서 보는 바와 같이 전체 입력 변수의 총 수(24)의 반인 f = 12에서 두



〈Figure 4〉 Sensitivity Analysis of the Random Subspace Ensemble Model at Different f Values

〈Table 4〉 Experimental Results of the Random Subspace Ensemble Model

f	f = 1	f = 2	f = 3	f = 4	f = 5	f = 6	f = 7	f = 8	f = 9	f = 10	f = 11	f = 12	f = 13	f = 14	f = 15	f = 16	f = 17	f = 18	f = 19	f = 20	f = 21	f = 22	f = 23	f = 24
RS	63.74	68.73	69.77	70.30	71.41	72.59	73.01	72.66	73.58	73.59	73.42	73.77	73.34	74.14	73.52	73.65	73.53	73.20	73.17	72.68	72.26	72.07	72.00	72.44
BC	56.17	60.66	63.10	64.50	65.53	66.47	67.30	67.44	68.30	68.71	69.04	69.39	69.52	69.92	70.12	70.46	70.61	70.83	71.00	71.11	71.27	71.44	71.64	72.44
Q	0.22	0.32	0.41	0.48	0.51	0.55	0.58	0.62	0.64	0.67	0.70	0.73	0.75	0.77	0.80	0.82	0.84	0.86	0.88	0.91	0.93	0.95	0.98	1.00
KNN	72.44																							

번째로 좋은 결과를 보였으며, f = 14에서 가장 좋은 결과를 보임을 알 수 있다. 전반적으로 전체 변수의 총수의 0.5에 해당하는 f = 12 부근에서 좋은 결과를 보임을 알 수 있다. 즉, 본 연구에서의 실험 결과 랜덤 서브스페이스에 관한 선행 연구 결과와 일관된 결과를 보임을 알 수 있다.

앞에서 살펴본 바와 같이 앙상블을 구성하고 있는 기저 분류기들이 서로 얼마만큼 다양성을 가지고 있는 가는 앙상블의 성과에 중요한 영향을 미친다. 본 논문에서는 앙상블 모형의 예측 정확도 뿐만 아니라 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률, 기저 분류기들 간의 다양성 지수 등을 같이 살펴보았다. 기저 분류기들 간의 다양성을 측정하기 위한 많은 방법들이 제안되어 왔으며, 본 연구에서는 가장 대표적인 다양성 척도 중의 하나인 Q-통계량을 살펴보았다.

Q-통계량은 Yule의 Q 통계량에 의해 다음과 같이 구할 수 있다(Yule, 1900; Kuncheva and Whitaker, 2003). 두 개의 분류기 f_i와 f_j가 있다고 가정하자. 아래 <Table 5>는 두 개의 분류기의 예측 오차에 대한 일치 정도를 나타내 주고 있다. N_a는 두 분류기 모두 정확하게 예측한 데이터(또는 사례)의 수를 의미하고 N_d는 두 분류기 모두 오분류한 사례의 수를 의미한다. N_b와 N_c는 두 분류기의 예측 결과가 불일치를 보인 경우를 의미한다. 즉, N_b는 분류기 f_i는 옳게 예측하고 분류기 f_j는 오분류한 경우를 뜻한다. N_c는 분류기 f_i는 오분류하고 분류기 f_j는 옳게 분류한 경우를 뜻한다. 이때 두 분류

〈Table 5〉 The Table of the Relationship between a Pair of Classifiers

	f_j *correct*	f_j *wrong*
f_i *correct*	N_a	N_b
f_i *wrong*	N_c	N_d

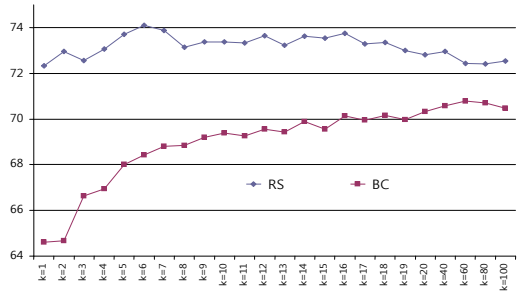
기 f_i와 f_j의 Q-통계량은 아래와 같이 식 (3)에 의해 계산될 수 있다.

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

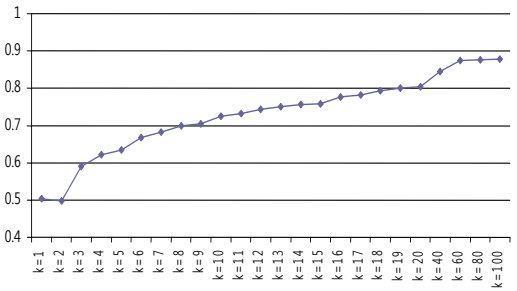
Q-통계량의 값은 -1과 1사이에 존재하게 되며 통계적으로 독립적인 분류기들의 Q 통계량의 값은 0(Q = 0)이 될 것이다. 또한, 두 분류기들이 정확하게 같은 패턴으로 분류를 한다면 이들 사이의 Q값은 양수(Q > 0)의 값을 갖게 될 것이며, 서로 다른 패턴으로 오분류하는 분류기간의 Q값은 음수(Q < 0)가 될 것이다. <Table 4>에서의 Q-통계량 값은 위와 같은 방법으로 구한 각 분류기의 Q-통계량의 평균값을 의미한다.

<Table 4>는 f의 변화에 따른 랜덤 서브스페이스 앙상블 모형의 예측률, 기저 분류기의 평균 예측률, Q통계량을 보여주고 있다. 실험 결과 Q-통계량은 f가 증가함에 따라 커짐을 알 수 있으며, 동일한 기저 분류기로 구성된 f = 24인 경우에는 1이 됨을 알 수 있다. <Table 4>에서 보는 바와 같이 Q-통계량이 커짐에 따라 기저 분류기를 결합하여 앙상블 모형을 구성할 경우의 성과 개선의 폭이 줄어드는 것 알 수 있다.

<Figure 5>는 KNN의 파라미터 k에 따른 앙상블 성능의 변화를 보여주고 있다. 그림에서 RS는 k에 따른 랜덤 서브스페이스 앙상블의 예측률을 나타내고 있다. BC는 랜덤 서브스페이스 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률의 k에 따른 변화를 보여주고 있다. <Figure 6>은 랜덤 서브스페이스 앙상블을 구성하고 있는 기저 분류기들의 평균 Q-통계량의 k에 따른 변화를 보여주고 있다. 그림에서 보는 바와 같이 앙상블의 성과는 k에 영향을 많이 받는 것을 알 수 있다. 본 연구에서의 실험 결과 f = 12로 고정했을 때 k가 증가함에 따라 랜덤 서브스페이스 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률과 Q-통계량이 모두 증가하는 것을 알 수 있다. 앙상블의 예측 성과는 분류기들의 평균 예측률이 높을수록 좋으며, Q-통계량 값이 0에 가까울수록 좋은 것으로 알려져 있다. 그러므로, 본 실험 결과 평균 예측률 측면에서는 k값이 클수록 좋지만, Q-통계량 측면에서는 k의 값이 작을수록 좋다는 것을 알 수 있다. 그러므로, 앙상블 모형의 성과 측면에서 k의 값이 크거나 작을수록 좋은 것이 아니고 어느 중간 지점에서 최적의 값을 갖는 다는 것을 유추할 수 있게 된다. 본 실험에서는 <Figure 5>에서 보는 바와 같이 f = 12로 고정하였을 경우에 k = 6일 때 랜덤 서브스페이스 앙상블의 예측률이 가장 좋은 값을 보였다.



〈Figure 5〉 Sensitivity Analysis of the Random Subspace Ensemble Model at Different k Values



〈Figure 6〉 Q-statistics of the Base Classifiers of the Ensemble at Different k Values

5.2 모형 검증

본 논문에서는 일반적인 랜덤 서브스페이스 앙상블 모형의 성능 개선을 위해 유전자 알고리즘을

사용한 최적화 모형을 제안하였다. 본 논문에서 제안한 모형의 우수성을 검증하기 위해 실제 기업의 부도 예측을 위한 데이터를 사용해 실험하였으며 각 모형별 최종 예측 성과는 <Table 6>과 같다.

〈Table 6〉 Final Experimental Results

	Accuracy(%)	BC(%)	Q
KNN	72.44	-	-
RS_KNN	73.71	69.31	0.7323
GARS_KNN	75.16	69.73	0.7336

본 연구에서는 선행 연구에서 가장 많이 추천하고 있는 파라미터 값인 전체 입력 변수의 0.5에 해당 하는 f = 12일 경우를 채택하여 최종실험을 하였다. 즉, 본 논문에서 제안한 유전자 알고리즘을 이용하여 최적의 분류기를 선택하는 모형의 경우 KNN의 파라미터 k = 9를, RS의 파라미터 f = 12을 이용하여 실험하였다. 앞에서 설명한 바와 같이 10개의 fold로 데이터를 나눈 후 실험을 하였으며 9개의 fold는 학습 및 테스트용 데이터 나머지 1개는 검증용 데이터로 사용하였으며, 검증용 데이터로 사용할 fold를 변경해 가며 실험을 하였다.

<Table 6>의 값은 검증용 데이터에서의 결과값의 평균값이다. 여기서 KNN은 단일 KNN 모형을 의미하며, RS_KNN은 KNN을 기저 분류기로 사용하는 랜덤 서브스페이스 앙상블 모형을 의미한다. 또한 GARS_KNN은 KNN을 기저 분류기로 하는 랜덤 서브스페이스 모형의 최적화 모형을 의미한다. Accuracy 값은 각각의 모형의 최종 예측률을 의미하며, BC는 앙상블 모형을 구성하고 있는 기저 분류기의 평균 예측률, Q는 평균 Q통계량을 의미한다. <Table 6>에서 보는 바와 같이 랜덤 서브스페이스 앙상블 모형의 성과가 단일 모형의 성과보다 더 좋은 것을 알 수 있다. 또한, 랜덤 서브스페이스 앙상블 모형의 기저 분류기 중에서 유전자 알고리즘을 이용하여 찾아낸 최적의 분류기 조합으로 이루어진 앙상블 모형인 GARS_KNN의 값이 가장 좋은 예측 정확도를 보임을 알 수 있다. 또한, GARS_KNN과 RS_KNN의 BC와 Q를 <Table 4>의 결과를 기준으로 볼 때 BC의 값은 증가했지만 Q의 값은 거의 동일한 수준으로 유지되고 있는 것을 알 수 있다. 즉 유전자 알고리즘이 랜덤 서브스페이스 모형의 기저 분류기 풀에서 다양성 지수는 유지하면서 예측률이 높은 분류기들을 선택했음을 알 수 있다. 또한, 이로 인해 유전자 알고리즘을 통해 선택된 기저 분류기 조합으로 구성된 앙상블 모형의 예측 성과가 크게 향상되었음을 알 수 있다.

각 모형간의 성과 차이에 대한 통계적 유의성을 검토하기 위해 비모수적인 방법인 Wilcoxon 부호순위 검정을 이용하였다. <Table 7>은 검정 결과를 보여주고 있다. <Table 7>에서 보는 바와 같이 랜덤 서브스페이스 앙상블 모형은 단일 모형보다 통계적으로 유의한 차이가 있는 것으로 나왔다. 또한, 본 연구에서 제안한 모형이 기존의 단일 모형, 단순한 랜덤 서브스페이스 모형보다 통계적으로 유의한 차이가 있는 것으로 나왔다. 즉, 본 연구에서 제안한 모형이 기존의 모형보다 효과적임을 알 수 있었다.

〈Table 7〉 Wilcoxon Signed-Rank Test(p-Value)

	RS_KNN	GARS_KNN
KNN	0.005	0.005
RS_KNN	-	0.005

6. 결 론

앙상블 분류기란 단일 분류기보다 더 좋은 성과를 내기 위해 여러 개의 분류기들을 결합하는 것을 의미하며 최근에 데이터마이닝, 기계학습 분야에서 많은 관심을 끌고 있다. 앙상블 분류기의 뛰어난 성능으로 다양한 분야에서 앙상블 분류기를 적용하고 있으며, 부도 예측 문제에도 다양한 연구가 시도되고 있다.

본 연구에서는 앙상블 기법 중에 랜덤 서브스페이스 기법을 국내 기업의 부도 예측 문제에 적용해 보았으며 앙상블 모형과 기저 분류기의 파라미

터 값의 변화에 따른 다양한 탐색적 분석을 수행하였다. 또한 랜덤 서브스페이스 모형의 성과 개선을 위해 유전자 알고리즘을 이용한 최적의 분류기 조합 선택 모형을 제안하였다.

본 연구에서 제안한 모형의 우수성을 실제 국내 기업 데이터를 가지고 실험을 수행하였으며, 실험 결과 랜덤 서브스페이스 앙상블 모형이 기존의 단일 분류기 모형보다 우수한 성과를 보임을 알 수 있었다. 또한, 유전자 알고리즘을 이용하여 랜덤 서브스페이스 앙상블의 최적 조합을 선택한 모형이 단일 분류기, 일반 랜덤 서브스페이스 모형보다 우수한 성과를 보임을 알 수 있었다.

본 연구의 한계와 향후 연구 방향을 정리하면 다음과 같다. 본 연구에서 제안한 모형은 예측 정확도 측면에서 개선이 있었지만, 모형의 복잡성은 더 증가하였다. 일반적으로 실시간 처리가 요구되는 분야에서는 모형의 예측 정확도뿐만 아리라 모형의 간결성 또한 매우 중요하다. 하지만, 기업 부도 예측 문제의 경우 상대적으로 모형의 속도, 간결성 보다는 모형의 예측 정확도가 더 중요한 분야이므로 본 연구에서 제안한 모형은 의미가 있다고 볼 수 있다. 향후 연구 과제로는 보다 다양한 데이터에서의 추가 적인 분석 및 검증이 더 필요할 것으로 보인다. 또한, 본 연구에서는 KNN을 기저 분류기로 사용한 앙상블 모형을 제안하였지만 향후에는 다른 분류 모형을 기저 분류기로 하는 앙상블 모형의 성과 개선에 대한 연구가 필요할 것으로 생각된다.

References

Abellán, J. and C.J. Mantas, “Improving Experimental Studies about Ensembles of Classifiers for Bankruptcy Prediction and Credit Scoring”, *Expert Systems with Applications*, Vol.41, No.8, 2014, 3825-3830.

Altman, E.L., “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy”, *The Journal of Finance*, Vol. 23, No.4, 1968, 589-609.

Beaver, W., “Financial Ratios as Predictors of Failure, Empirical Research in Accounting : Selected Studied”, *Journal of Accounting Research*, Vol.4, No.3, 1966, 71-111.

Bian, S. and W. Wang, “On Diversity and Accuracy of Homogeneous and Heterogeneous Ensembles”, *International Journal of Hybrid Intelligent Systems*, Vol.4, No.2, 2007, 103-128.

Breiman, L., “Bagging Predictors”, *Machine Learning*, Vol.24, No.2, 1996, 123-140.

Buta, P., “Mining for Financial Knowledge with CBR”, *AI Expert*, Vol.9, No.10, 1994, 34-41.

Dietterich, T.G., “Machine-Learning Research : Four Current Directions”, *AI Magazine*, Vol.18, No.4, 1997, 97-136.

Freund, Y. and R. Schapire, “Experiments with a New Boosting Algorithm”, *Proceedings of the 13th International Conference on Machine learning*, 1996, 148-156.

Ho, T., “The Random Subspace Method for Construction Decision Forests”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No.8, 1998, 832-844.

Ho, T., “Multiple Classifier Combination : Lessons and Next Steps”, *Hybrid Methods in Pattern Recognition*, Kandel and Bunke, Eds. Singapore : World Scientific, 2002, 171-198.

Kim, J.I. and E.J. Lee, “A Technique to Apply Inlining for Code Obfuscation Based on Genetic Algorithm”, *Journal of Information Technology Services*, Vol.10, No.3, 2011, 167-177.

(김정일, 이은주, “유전 알고리즘에 기반한 코드 난독화를 위한 인라인 적용 기법”, *한국IT서비*

스학회지, 제10권, 제3호, 2011, 167-177.)

Kim, M.J., "A Performance Comparison of Ensemble in Bankruptcy Prediction", *Entrue Journal of Information Technology,* Vol.8, No.2, 2009, 41-49.

(김명종, "기업부실화 예측에 대한 앙상블 학습의 성과 비교", 엔트루 저널, 제8권, 제2호, 2009, 41-49.)

Kim, M.J., "Optimal Selection of Classifier Ensemble Using Genetic Algorithms", *Journal of Intelligence and Information Systems*, Vol.16, No.4, 2010, 99-112.

(김명종, "유전자 알고리즘을 이용한 분류자 앙상블의 최적 선택", *지능정보연구*, 제16권, 제4호, 2010, 99-112.)

Kim, M.J., D.K. Kang, and 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, 1074-1082.

Kim, S.H. and J.W. Kim, "SOHO Bankruptcy Prediction Using Modified Bagging Predictors", *Journal of Intelligence and Information Systems*, Vol.13, No.2, 2007, 15-26.

(김승혁, 김종우, "Modified Bagging Predictors를 이용한 SOHO 부도 예측", *한국지능정보시스템학회논문지*, 제13권, 제2호, 2007, 15-26.)

Kuncheva, L.I. and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy", *Machine Learning*, Vol.51, No.2, 2003, 181-207.

Kuncheva, L.I., "Combining Pattern Classifiers : Methods and Algorithms", *Wiley-Interscience*, 2004.

Lee, E.J., Y.R. Seo, S.Y. Yoon, H.R. Jang, and H.K. Bahn, "Power Scheduling of Smart Buildings in the Smart Grid Environment Using IT Optimization Techniques", *Journal of Information Technology Services*, Vol.11, 2012, 41-50.

(이은지, 서유리, 윤소영, 장혜린, 반효경, "IT 최적화 기술을 이용한 지능형전력망 환경의 스마트 빌딩 스케줄링", *한국IT서비스학회지*, 제11권, 특집호, 2012, 41-50.)

Li, H., Y.C. Lee, Y.C. Zhou, and J. Sun, "The Random Subspace Binary Logit(RSBL) Model for Bankruptcy Prediction", *Knowled ge-Based Systems*, Vol.24, No.8, 2011, 1380-1388.

Louzada, F., O. Anacleto-Junior, C. Candolo, and J. Mazucheli, "Poly-Bagging Predictors for Classification Modelling for Credit Scoring", *Expert Systems with Applications*, Vol.38, No.10, 2011, 12717-12720.

Meyer, P.A. and H. Pifer, "Prediction of Bank Failures", *The Journal of Finance*, Vol.25, No.4, 1970, 853-868.

Min, S.H., "Bankruptcy Prediction Using an Improved Bagging Ensemble", *Journal of Intelligence and Information Systems*, Vol.20, No.4, 2014, 121-139.

(민성환, "개선된 배깅 앙상블을 활용한 기업부도 예측", *지능정보연구*, 제20권, 제4호, 2014, 121-139.)

Ohlson, J., "Financial Ratios and the Probabilistic Prediction of Bankruptcy", J*ournal of Accounting Research*, Vol.18, No.1, 1980, 109-131.

Shin, T. and T. Hong, "Corporate Credit Rating Based on Bankruptcy Probability Using AdaBoost Algorithm-Based Support Vector Machine", *Journal of Intelligence and Information Systems*, Vol.17, No.3, 2011, 25-41.

Yule, G.U., "On the association of attributes in

statistics : with illustrations from the material of the childhood society, and Philosophical Transactions of the Royal Society of London", *Series A, Containing Papers of a Mathematical or Physical Character*, 1900, 257-319.

Zhang, G., Y.M. Hu, E.B. Patuwo, and C.D. Indro, "Artificial Neural Networks in Bankruptcy Prediction : General Framework and Cross-Validation Analysis", *European Journal of Operational Research*, Vol.116, No.1, 1999, 16-32.

About the Authors

Sung-Hwan Min (shmin@hallym.ac.kr)

Sung-Hwan Min received the Ph.D. degree in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST). He is an associate professor in the School of Business at Hallym University. His current research interests include data mining, recommender systems and artificial intelligence applications for business.