韓國IT서비스學會誌
第15卷 第1號
2016年 3月, pp.153-166

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2016.15.1.153

랜덤화 배깅을 이용한 재무 부실화 예측

민 성 환*

Randomized Bagging for Bankruptcy Prediction

Sung-Hwan Min*

■ Abstract ■

Ensemble classification is an approach that combines individually trained classifiers in order to improve prediction accuracy over individual classifiers. Ensemble techniques have been shown to be very effective in improving the generalization ability of the classifier. But base classifiers need to be as accurate and diverse as possible in order to enhance the generalization abilities of an ensemble model. Bagging is one of the most popular ensemble methods. In bagging, the different training data subsets are randomly drawn with replacement from the original training dataset. Base classifiers are trained on the different bootstrap samples. In this study we proposed a new bagging variant ensemble model, Randomized Bagging (RBagging) for improving the standard bagging ensemble model. The proposed model was applied to the bankruptcy prediction problem using a real data set and the results were compared with those of the other models. The experimental results showed that the proposed model outperformed the standard bagging model.

Keyword : Bagging, Bankruptcy Prediction, Ensemble

Submitted : January 20, 2016 1st Revision : March 15, 2016 Accepted : March 18, 2016

* 한림대학교 경영학과

1. 서 론

앙상블 기법은 다수의 분류기들을 구성하고, 이들 분류기들의 결과값(outputs)을 특정 방식에 의해 결합함으로써 새로운 표본에 대해 분류를 하는 학습 알고리즘을 말한다. 이와 같이 다수의 분류기를 이용하는 앙상블 기법은 단일 분류기보다 분류 정확도를 개선시킬 수 있는 것으로 알려져 있다(Dietterich, 1997). 앙상블 모형의 성과 개선을 위해서는 앙상블을 구성하고 있는 기저 분류기들의 예측 정확도뿐만 아니라 기저 분류기들 간의 다양성이 매우 중요하다. 기저 분류기들이 다양성을 갖게 된다면 어느 하나의 분류기가 오분류를 하더라도 다른 분류기들이 이를 보완해 줄 수 있을 것이며, 이를 통해 전체적으로 앙상블 모형의 정확도는 높아지게 된다. 반대로 기저 분류기들의 다양성 지수가 높지 않을 경우에는 여러 분류기가 동시에 오분류를 하게 되어 이들의 결합을 통한 성과 개선은 기대할 수 없을 것이다. 극단적인 예로 완전히 일치하는 예측 결과를 보이는 기저 분류기를 결합한 앙상블 모형은 개별 분류기 보다 더 좋은 성과를 전혀 기대할 수 없을 것이다. 이와 같이 앙상블 모형을 구성하고 있는 기저 분류기들의 다양성은 앙상블 모형의 성과에 영향을 주는 매우 중요한 요소이다(Kuncheva and Whitaker, 2003; Bian and Wang, 2007).

지금까지 앙상블 모형을 구성하기 위한 많은 기법들이 개발되어 왔으며 그 중에서 가장 대표적인 방법이 기저 분류기들을 학습시킬 학습 데이터의 다양화를 통한 기저 분류기 다양화 방법이다. 여기에 속하는 대표적인 예로는 배깅(bagging) (Breiman, 1997)과 부스팅(boosting) (Freund and Schapire, 1996)을 들 수 있다. 이와 같은 기법의 주된 개념은 서로 다른 학습 데이터를 발생시키기 위해 원 학습 데이터를 다루는 것이며, 각각의 기저 분류기들을 서로 다른 학습 데이터를 가지고 학습시키게 된다. 이와 같은 방법에 의해 각각의 기저 분류기들은 서로 다른 모습을 보이는 다양성을 갖게 되고 이 다양성은 앙상블 모형의 예측 정확도를 개선시키는데 도움이 된다.

배깅은 서로 다른 학습 데이터를 발생시키기 위해 부트스트랩 샘플링(bootstrap sampling) 방법을 이용한다. 이를 통해 N개의 데이터로 구성된 원 학습 데이터로부터 복원추출 방식에 의해 랜덤하게 N개의 학습 데이터로 구성된 서로 다른 새로운 형태의 학습 데이터를 발생시키게 되며 이를 통해 다양성이 존재하는 서로 다른 기저 분류기를 생성하게 되고 이들을 다수결 투표(majority voting)와 같은 특정한 전략에 의해 결합하게 된다. 배깅은 가장 대표적인 앙상블 기법으로 모형이 비교적 단순하면서도 성과가 좋아 다양한 분야에서 성공적으로 적용되고 있다.

본 논문은 앙상블 모형에서 가장 많이 활용되고 있는 배깅의 성능 개선에 관한 연구이다. 이를 위해 본 논문에서는 기존 배깅 알고리즘을 수정한 새로운 형태의 변형된 배깅 모형을 제안하였다. 또한, 제안한 모형의 검증을 위해 국내 기업의 부도 예측 문제에 적용하여 기존의 배깅 모형과 비교해 보았다.

본 논문의 구성은 다음과 같다. 다음 장에서는 재무 부실화 예측 모형에 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 실험 결과에 대해 서술하였다. 마지막 장에서는 요약 및 향후 연구 과제에 대해 설명하였다.

2. 재무 부실화 예측 모형

기업의 재무 부실화를 예측하는 것은 산업계와 학계에서 모두 중요하게 다뤄지고 있는 연구 주제이다. 초창기 재무 부실화 예측 모형은 주로 단일변량 분석, 다변량 판별 분석, 다중 회귀 분석, 로지스틱 회귀 분석과 같이 통계적인 모형에 기반을 둔 모형이 대부분이었다(Beaver, 1966; Altman, 1968; Meyer and Pifer, 1970; Ohlson, 1980). 하지만,

이들 통계적인 기법은 대부분 선형성, 정규성 등과 같은 엄격한 가정들을 하고 있어 현실 문제에서의 적용에 제한이 따르고 있다. 그 뒤로 전통적인 모형보다 더 좋은 예측성과를 내기 위해 귀납적 학습, 인공신경망, 사례기반 추론과 같은 다양한 데이터 마이닝 기법들을 재무 부실화 문제에 적용해 보려는 연구가 활발하게 진행되었다(Bryant, 1997; Shaw and Gentry, 1988; Zhang et al., 1999; Kim and Jhee, 2012; Lu et al., 2015; Yu, 2014; Zhang et al., 2013).

한편, 최근에는 기존의 단일 모형보다 더 좋은 성과를 내는 것으로 알려져 있는 앙상블 모형을 재무 부실화 예측 분야에 적용해 보려는 다양한 연구가 활발하게 진행되고 있다. 앙상블 모형이란 단일 분류기보다 더 좋은 성과를 내기 위해 다수의 기저 분류기들을 구성하고, 그들 기저 분류기들의 결과값을 특정 방식에 의해 결합하는 것을 의미한다. 이와 같은 앙상블 모형을 적절하게 구성하게 될 경우 단일 분류기 보다 더 좋은 성과를 내는 것으로 알려져 있으며, 이로 인해 앙상블 모형은 다양한 분야에서 많은 관심을 끌고 있다(Dietterich, 1997).

Kim and Kim(2007)은 성능 평가용 데이터에서 평균 이상의 예측 정확도를 가지는 기저 분류기만을 선택하여 배깅을 구성하는 변형된 배깅 모형을 제안하여 SOHO의 부도 예측 문제에 적용해 보았으며 실험 결과 제안한 모형이 기존 모형 보다 우수한 성과를 보였다. Kim(2009)은 의사결정 트리, 인공신경망 모형, SVM(Support Vector Machines) 모형을 기저 분류기로 하는 배깅과 부스팅 앙상블 모형을 기업의 부도 예측 문제에 적용해 보았다. 실험 결과 의사결정 트리와 인공 신경망 모형을 기저 분류기로 사용할 경우에는 앙상블 모형이 단일 모형보다 통계적으로 유의한 성과 개선이 있었으며, SVM을 기저 분류기로 사용할 경우에는 앙상블 학습을 통한 유의적인 성과 개선이 나타나지 않았다. Li et al.(2011)는 이진 로짓 모형을 기저 분류기로 하는 랜덤 서브스페이스 부도 예측 모형을 제안하였으며, 제안한 모형과 다른 전통적인 단일 모형과 예측성과를 비교 분석하였다. 분석 결과 제안한 앙상블 모형이 단일 모형보다 우수한 성과를 보였다. Choi and Lim(2013)은 여러 가지 커널 함수에 따른 SVM 모형들을 결합하여 앙상블 모형을 구축하고 이를 부도 예측 문제에 적용해 보았으며, 실험 결과 제안한 모형이 단일 모형보다 높은 예측성과를 보였다. Min(2014)은 배깅과 사례 선택(instance selection)을 결합하는 새로운 모형을 부도 예측 문제에 적용해 보았으며 제안한 모형이 기존의 배깅 모형의 성능을 개선하는데 효과가 있음을 알 수 있었다. Min(2015)은 유전자 알고리즘을 이용한 랜덤 서브스페이스 앙상블 모형의 최적화 모형을 부도 예측 문제에 적용해 보았다. Kim et al.(2015)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 예측성과를 보였다.

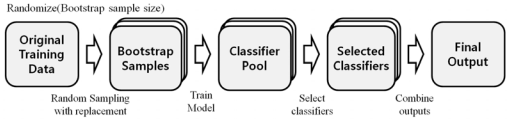
3. 연구 모형

본 연구는 배깅의 성능 개선에 관한 연구이다. 전통적인 배깅 모형에서 부트스트랩 샘플(bootstrap sample)의 크기는 원 학습 데이터의 크기와 같은 크기로 고정한다. 본 논문에서는 앙상블을 구성하고 있는 기저 분류기간의 다양성을 높이기 위해 기존의 전통적인 배깅 모형을 변형한 모형을 제안한다. 제안한 모형에서는 각각의 부트스트랩 샘플을 구성할 때 기존처럼 부트스트랩 샘플의 크기를 원 학습 데이터의 크기로 고정하는 것이 아니고, 랜덤하게 크기에 변화를 주어 부트스트랩 샘플링을 하게 된다. 이와 같은 방식은 기저 분류기간의 다양성 지수를 높여주어 결국 앙상블 모형의 최종 성과에 긍정적인 요인으로 작용할 것으로 기대된다. 하지만, 랜덤하게 부트스트랩 샘플링 사이즈를 조절할 경우 기본 배깅보다 작은 크기의 샘플이 선택되어 기저 분류기의 예측률 저하가 있을 수도 있을 것이다. 본 논문에서는 이를 보완하기

위해 기저 분류기 풀(classifier pool)에서 정확도 순으로 일정 비율 이상인 것만을 선택하여 앙상블을 구성하는 선택적 배깅(selective bagging) 모형을 제안하였다. 이와 같은 방법으로 기존의 배깅 모형보다 다양성 지수와 기저 분류기의 평균 예측률 측면에서 좋은 값을 얻게 되어 결과적으로 앙상블 모형의 성과 개선에 기여할 것이다.

배깅은 가장 대표적인 앙상블 기법 중의 하나로 학습 데이터의 다양화를 통해 다양한 기저 분류기를 생성한다. 서로 다른 학습 데이터를 발생시키기 위해 배깅은 부트스트랩 샘플링 방법을 이용한다. 이를 통해 N개의 데이터로 구성된 원 학습 데이터로부터 복원추출 방식에 의해 랜덤하게 N'개의 학습 데이터로 구성된 서로 다른 형태의 새로운 형태의 학습 데이터를 발생시키게 된다. 이와 같이 구성된 새로운 형태의 학습 데이터를 K번 반복해서 생성하고, 이들 K개의 서로 다른 학습 데이터를 이용해 다양성이 존재하는 서로 다른 기저 분류기를 생성하게 되며, 이들 기저 분류기들을 특정한 방식에 의해 결합하게 된다. 이때 전통적인 배깅 방식의 경우 N'는 N과 같은 크기로 하여 배깅 앙상블 모형을 구성하게 된다.

본 연구에서 제안한 새로운 형태의 배깅 변형 모형인 랜덤화 배깅(RBagging)은 기저 분류기의 다양성 지수를 높이기 위해 N'의 크기를 N으로 고정하여 부트스트랩 샘플링을 하는 것이 아니고, 랜덤하게 크기를 결정하는 것이다. 이와 같은 방법으로 보다 더 다양한 형태의 학습 데이터가 생성되게 되며, 이를 이용하여 기저 분류기를 구성하고 결합하게 된다. 이는 기존 전통적인 배깅 방식보다 다양성 지수가 높은 기저 분류기를 발생시킬 것으로 기대되며, 이는 앙상블 모형의 기본 전략 목표와도 일치하는 방향이라고 할 수 있다. 다만, 랜덤하게 부트스트랩 샘플의 크기를 결정하게 되어 N'의 값이 전통적인 배깅 방식보다 더 작은 값을 갖게 되므로 학습 데이터 셋의 크기는 일반 배깅보다 전반적으로 작아지게 될 것이다. 이로 인해 각 기저 분류기의 개별 성과는 낮아질 것으로



〈Figure 1〉 The Overall Process of the Proposed Model

예상되며, 이를 보완하기 위한 RBagging의 수정 모형인 RBagging(p)모형도 제안하여 이들의 성과를 같이 비교하여 보았다.

RBagging 모형이 랜덤하게 N'의 크기를 결정하고, 이에 따라 발생시킨 다수의 학습 데이터 셋에서 발생시킨 모든 기저 분류기들을 앙상블의 구성 요소로 사용하는 것과 달리, RBagging(p) 모형의 경우 RBagging 방식으로 구성된 기저 분류기 풀에서 예측률 기준으로 상위 p%에 있는 기저 분류기들을 선택하여 배깅 앙상블을 구성하는 선택적인 배깅(selective bagging) 모형이라고 할 수 있다. 본 논문에서 제안한 모형인 RBagging과 RBagging(p)의 전반적인 절차는 <Figure 1>에 나와 있으며 자세한 내용은 다음과 같다.

1단계 : 데이터 준비 및 입력 변수 설정

(a) 데이터 준비

전체 데이터를 학습용 데이터와(T) 검증용 데이터(V)로 분할한다. 학습용 데이터는 다시 모형 구축을 위한 데이터(T_1)와 선택적 배깅에서 예측 정확도 비교를 하기위해 사용한 데이터(T_2)로 분류한다.

(b) 파라미터 설정 : K, S', p(%)

앙상블 모형의 대표적인 파라미터로는 앙상블 크기(K)와 부트스트랩 샘플의 크기(S')가 있으며 이들은 일반적으로 실험 전에 고정하여 설정한다. 기본 배깅 모형에서 부트스트랩 샘플의 크기는 원 학습 데이터의 크기와 같게 고정하여 부트스트랩 샘플링을 진행한다. 하지만, 본 논문에서 제안한 모형인 랜덤화 배깅 모형에서는 부트스트랩 샘플의 크기를 랜덤하게 변경해 가며 부트스트랩 샘플링을 진행한다.

p(%)는 본 논문에서 제안한 랜덤화 배깅 모형의 변형 모형에서 필요한 파라미터로 선택적 배깅에서 기저 분류기 선택 기준이 된다. 즉, 기저 분류기 풀 중에서 예측 정확도 비교를 위해 사용한 데이터(T_2)에서의 예측률 기준 상위 p%인 기저 분류기만을 선택하여 새로운 선택적 배깅 모형을 구성하게 된다.

2단계 : 랜덤화

본 논문에서 제안한 새로운 형태의 랜덤화 배깅 모형에서는 부트스트랩 샘플의 크기로 사용할 S'를 매번 고정하여 부트스트랩 샘플을 발생시키지 않고 랜덤화하여 결정한다. 이를 통해 기저 분류기의 다양성 지수가 좋아질 것으로 기대되며 이는 앙상블 모형의 성과 개선에 긍정적으로 작용할 것이다. 본 논문에서 제안한 새로운 형태의 부트스트랩 샘플 크기(S')는 아래와 같이 정의한다.

$$S'_i = a+int(Rand(Size(T_1)-a)) \quad (i = 1, \cdots, K) \qquad (1)$$

여기서 $Size(T_1)$은 학습을 위해 사용한 데이터 T_1의 크기를 의미하고, a는 모형 구축을 위해 필요한 최소 데이터 수를 의미한다. $Rand(Size(T_1)-a)$는 0부터 $(Size(T_1)-a)$ 사이의 수를 랜덤하게 발생시키며, int()는 랜덤하게 발생한 수를 정수로 변환해 주는 함수를 의미한다. 그러므로, 위의 식 (1)을 통해 $[a, (Size(T_1)]$ 사이의 새로운 부트스트랩 샘플 크기가 결정되게 되며, 이를 통해 앙상블의 크기인 K개의 서로 다른 부트스트랩 샘플 크기가 결정되게 된다.

3단계 : 부트스트랩 샘플링

$$T'_i = \text{bootstrap sample}(Size(S'_i)) \text{ from } T_1 \quad (i = 1, \cdots, K) \qquad (2)$$

학습용 데이터에서 크기 S'_i인 데이터를 복원추출 방식에 의해 랜덤하게 선택하여 새로운 표본을 생성한다. 이를 통해 크기가 서로 다른 K개의 부트스트랩 샘플 군 $\{T'_1, T'_2, \cdots, T'_K\}$이 생성된다.

4단계 : 분류기 학습

이전 단계에서 생성된 학습 데이터를 이용해 분류기를 학습시켜 K개의 기저 분류기 풀 $\{C_1, \cdots, C_K\}$를 생성하게 된다.

$$C_i = \text{classifier}(T'_i) \ (i = 1, \cdots, K) \qquad (3)$$

5단계 : 분류

T_2 데이터에 대해 생성된 분류기를 적용하여 예측 결과 값을 계산한다.

$$Accuracy(C_i) = C_i(T_2) \ (i = 1, \cdots, K) \qquad (4)$$

6단계 : 선택적 앙상블 구성

$Accuracy(C_i)$를 내림차순으로 정렬한 후 상위 p(%)인 기저 분류기를 선택하여 선택적 앙상블 구성한다.

7단계 : 모형 검증

선택된 분류기에 검증용 데이터 V를 적용하여 결과 값을 구한 후 적절한 통합 전략에 의해 통합한다. 본 논문에서는 다수결 투표 방식으로 결과 값을 통합하였다.

4. 실험 설계

본 논문에서는 대표적인 앙상블 모형인 배깅 모형의 성능 개선을 위해 새로운 형태의 배깅 변형 모형인 랜덤화 배깅 모형을 제안하였다. 본 논문에서 제안한 랜덤화 배깅 모형의 우수성을 검증하기 위해 국내 부도 기업 예측 문제에 제안한 모형을 적용시켜 보았다. 실험을 위해 사용한 데이터는 총 1,800개로 부도기업과 비부도 기업이 각각 900개로 구성된 자산 규모가 10억에서 70억 사이인 국내 비외감 기업의 데이터를 사용하였다. 데

이터는 학습용 데이터와 검증용 데이터로 분류하여 실험을 하였으며, 학습용 데이터는 다시 모형 구축을 위한 데이터와 선택적 배깅을 위해 예측 정확도 비교를 위해 사용한 데이터로 분류하여 실험을 하였다. 모든 실험은 10-겹 검증(10-fold cross validation) 방법으로 실험을 하였다. 배깅, 랜덤화 배깅과 같은 앙상블 모형의 경우 각각의 데이터 셋에서 10회 반복하여 실험을 하였으며, 이들 값들의 평균을 계산하여 대푯값으로 사용하였다. 각각의 앙상블 모형에서 기저 분류기의 총 수는 100으로 고정하여 실험을 수행하였다. 본 연구에서는 앙상블 모형의 기저 분류기로 의사결정 나무(Decision Tree : DT)와 k 최근접 이웃(k Nearest Neighbor : KNN)을 사용하였다.

〈Table 1〉 Input Variables

Category	Description
Profitability	EBITDA to Sales
	Financial Expenses to Sales
	Financial Expenses to Debt
	Ordinary Income to Sales
	Net Income to Capital
	Ordinary Income to Capital
Stability	Fixed Asset to Owner's Equity
	Debt Ratio
	(Capital Surplus+Retained Earnings-Dividend)/Total Assets
	Borrowings to EBITDA
	Cash Ratio
Growth	Coefficient of Variation of Sales
Cash Flow	Cash Flow after Interest Payment to Sales
Activity	Sales to Net Change in Working Capital

기업의 부도 여부를 예측하기 위해 수익성, 안정성, 성장성, 활동성 및 현금 흐름으로 분류된 총 131개의 재무비율을 입력 변수 후보로 사용하였으며, 이 중에서 단일 표본 t검정(independent-samples t-test)을 실시하여 p-value 값이 0.05보다 큰 변수는 제외하고 나머지 변수를 대상으로 전진 선택법(forward selection)을 이용한 로지스틱 회귀분석을 이용하여 최종 변수를 선정하였으며 그 결과는 <Table 1>에 나와 있다.

5. 실험 결과

본 논문에서는 제안한 모형의 성과와 관련된 다양한 분석을 시도하였다. 이를 위해 앙상블 모형의 성과뿐만 아니라 앙상블 모형의 성과에 중요한 영향을 미치는 요소인 기저 분류기들의 평균 예측률과 기저 분류기간의 다양성 지수도 함께 비교분석을 하였다.

앙상블 모형에서 앙상블을 구성하고 있는 기저 분류기들의 다양성은 매우 중요하다. 여기서 다양성이란 기저 분류기들이 서로 다른 예측값을 내는 정도를 의미한다. 예를 들어 세 개의 분류기 C_1, C_2, C_3가 있다고 가정하고 이들 기저 분류기들의 10개의 사례에 대한 예측값이 C_1 = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0), C_2 = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0), C_3 = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)이라고 하고 실제값은 (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)이라고 가정 하자. 이때 C_1, C_2, C_3의 예측률은 각각 80%, 80%, 70%이며 기저 분류기들의 예측값은 다양성이 있다고 볼 수 없다. 예를 들어 분류기 C_1에서 옳게 분류한 사례는 1부터 8번째 사례이고 오분류한 사례는 9번째, 10번째 사례로 분류기 C_2와 같은 패턴임을 알 수 있다. 이와 같이 다양성이 거의 없는 이들 세 분류기를 다수결 투표 방식으로 결합한 앙상블 모형이 있다고 가정하면 그 앙상블 모형의 예측률은 80%가 될 것이다(1번부터 8번까지 사례는 1로 예측하여 옳게 분류하지만 9번과 10번의 사례에 대해 다수의 분류기가 0으로 예측하여 오분류를 하게 된다).

반면에 위의 예와 다른 형태의 예측 결과값을 보이는 세 개의 분류기 C'_1 = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0), C'_2 = (1, 1, 1, 1, 0, 0, 0, 1, 1, 1), C'_3 = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1)가 있다고 하자. 이때 C'_1, C'_2, C'_3의 예측률은 각각 70%, 70%, 60%로 앞의 경우

보다 예측률이 떨어지는 것을 알 수 있다. 하지만, 이들 세 분류기들은 앞의 예와 비교할 때 예측 결과값의 패턴이 서로 다른 형태라는 것을 알 수 있으며, 이는 기저 분류기들 사이의 다양성 지수가 높다고 얘기할 수 있다. 만약에 이들 기저 분류기들을 다수결 투표 방식으로 결합한다면 각 사례에 대해 세 분류기 중에 두 분류기가 1로 예측하므로 다수결 투표 방식에 의해 앙상블 모형의 예측 결과값도 모두 1이 되는 것을 알 수 있다. 결과적으로 이와 같은 앙상블 모형의 예측률은 100%가 됨을 알 수 있다. 이와 같이 앙상블 모형에서는 기저 분류기들의 다양성 지수가 매우 중요하다고 할 수 있다.

본 연구에서는 다양성 척도 중에서 가장 대표적인 Q-통계량을 사용하여 제안한 모형의 성과 개선 원인에 대한 분석을 시도하였다. <Table 2>는 분류기 C_i와 C_j가 있다고 가정했을 때 두 분류기의 예측 오차에 대한 일치 정도를 나타내 주고 있다. 이 때 두 분류기 C_i와 C_j의 Q-통계량은 식 (5)에 의해 계산될 수 있다(Kuncheva and Whitaker, 2003).

$$Q(C_i, C_j) = \frac{(N_a N_d - N_b N_c)}{(N_a N_d + N_b N_c)} \quad (5)$$

여기서 N_a는 두 분류기 C_i와 C_j가 모두 옳게 분류한 데이터의 수를 의미하고 N_d는 두 분류기 모두 잘못 분류한 데이터의 수를 의미한다. N_b와 N_c는 두 분류기의 분류 결과가 서로 다른 경우를 의미한다. 즉, N_b는 분류기 C_i는 옳게 분류하고 분류기 C_j는 오분류한 경우를 뜻한다. N_c는 분류기 C_i는 오분류하고 분류기 C_j는 옳게 분류한 경우를 뜻한다. Q-통계량의 값은 -1과 1사이의 값을 가지며, 통계적으로 독립적인 분류기들의 Q-통계량의 값은 0이 된다. 또한, 두 분류기들이 같은 패턴으로 분류를 한다면 Q-통계량 값은 양수가 되고, 반대로 두 분류기들이 서로 다른 패턴으로 분류를 한다면 Q-통계량 값은 음의 값을 갖게 된다.

본 연구에서 제안한 새로운 형태의 배깅 변형 모형인 랜덤화 배깅 모형에 대한 검증 및 다양한 분석을 위해 기저 분류기 별로 나누어서 분석을 실시하였으며 자세한 내용은 다음과 같다.

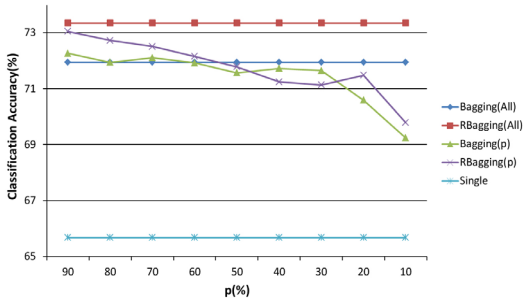
〈Table 2〉 The Table of the Relationship between a Pair of Classifiers

	C_j *correct*	C_j *wrong*
C_i *correct*	N_a	N_b
C_i *wrong*	N_c	N_d

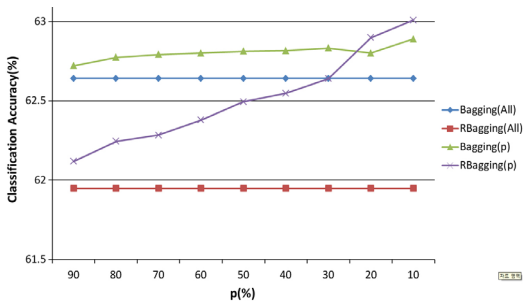
5.1 DT를 기저 분류기로 사용할 경우

DT를 기저 분류기로 사용한 경우 각각의 앙상블 모형의 여러 성과 지표 결과는 <Figure 2>, <Figure 3>, <Figure 4>에 나와 있다. 여기에서 Single은 단일 모형을 의미하고, Bagging(All)은 표준 배깅 모형을 의미하며, RBagging(All)은 본 연구에서 제안한 랜덤화 배깅 모형을 의미한다. Bagging(All) 모형과 RBagging(All) 모형은 생성한 기저 분류기 풀에 있는 모든 기저 분류기들을 사용한 앙상블 모형을 의미한다. 반면에 Bagging(p)와 RBagging(p) 모형은 생성한 기저 분류기 풀에서 일부를 선택하여 앙상블 모형을 구성한 선택적 앙상블 모형을 의미한다. 즉, Bagging 변형 모형인 Bagging(p) 모형은 배깅 앙상블 모형의 기저 분류기 중에서 예측률 기준으로 상위 p%인 분류기만을 선택하여 앙상블 모형을 구성한 모형을 의미하며, RBagging 변형 모형인 RBagging(p) 모형은 RBagging 모형의 기저 분류기 중에서 예측률 기준으로 상위 p%인 분류기만을 선택하여 앙상블 모형을 구성한 모형을 의미한다.

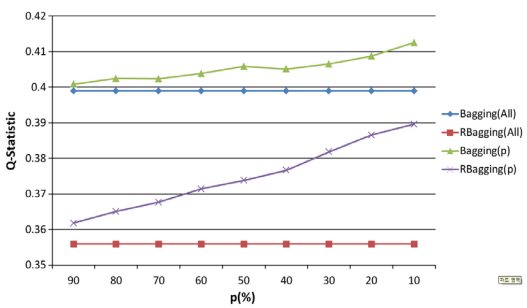
<Figure 2>는 RBagging과 Bagging 앙상블 모형의 p의 값에 따른 앙상블 예측률 변화를 보여주고 있다. 여기에서 p의 값이 10인 경우, 즉 RBagging(10)이 의미하는 것은 RBagging 앙상블 기법을 통해 기저 분류기를 발생시킨 후에, 이들 기저 분류기들 각각의 T_2 데이터에서의 예측률을 기준으로 상위 10%인 기저 분류기를 선택하여 앙상블 모형을 구축한 선택적 앙상블 모형을 의미한다.



〈Figure 2〉 Ensemble Classification Performance(%) (Base Classifier: DT)



〈Figure 3〉 The Average Classification Accuracy of base Classifiers(%)(Base Classifier : DT)



〈Figure 4〉 Diversity of the base Classifiers of the Ensembles(Base Classifier : DT)

그림에서 보는 바와 같이 DT를 기저 분류기로 사용할 경우 모든 앙상블 모형에서 단일 모형보다 성과가 좋아졌음을 알 수 있다. 또한, 본 논문에서 제안한 모형이 일반 배깅 모형보다 좋은 성과를 내고 있음을 알 수 있다.

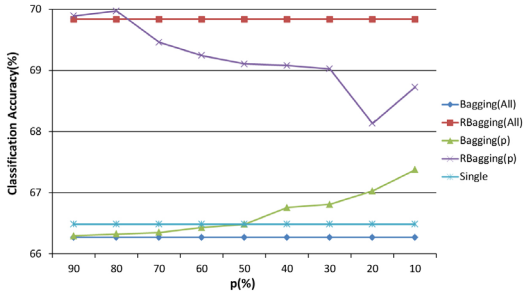
<Figure 3>와 <Figure 4>은 앙상블을 구성하고 있는 DT 기저 분류기들의 평균 예측률과 평균 Q-통계량의 값을 보여주고 있다.

<Figure 3>과 <Figure 4>에서 보는 바와 같이 p값이 작아짐에 따라 기저 분류기 평균 예측률과 Q-통계량 모두 상승하는 것을 알 수 있다. 앙상블 모형의 성과 측면에서 기저 분류기의 평균 예측률이 높을수록 좋고, 각 기저 분류기 간의 다양성 지수도 높을수록 좋다. 본 논문에서는 가장 대표적인 다양성 지표중의 하나인 Q-통계량을 통해 각 기저 분류기간의 다양성 지수를 살펴보았으며, Q-통계량의 값이 0에 가까울수록 기저 분류기간의 다양성 지수가 높다고 해석될 수 있으며, <Figure 4>과 같이 p의 값이 감소함에 따라 Q-통계량이 상승한 다는 것은, 전체 기저 분류기 풀 중에서 예측률 기준으로 상위에 있는 분류기간의 다양성 지수는 매우 낮다는 것을 의미하며 이로 인해 이들 기저 분류기를 결합할 경우의 성과 개선을 기대하기 힘들다는 것을 알 수 있다. <Figure 2>에서 보는 바와 같이 Bagging(p)와 RBagging(p) 모두 p가 감소함에 따라 앙상블 모형의 예측성과는 전반적으로 하락함을 알 수 있다.

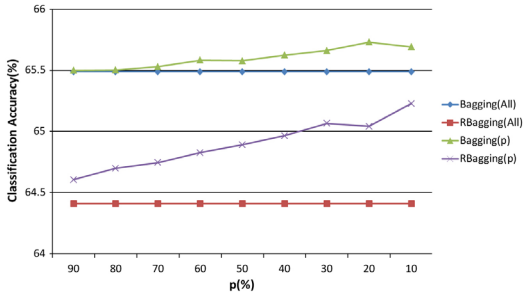
결론적으로 DT를 기저 분류기로 사용할 경우 RBagging 모형의 예측성과가 가장 좋은 것을 알 수 있다. RBagging 모형의 경우 기존의 배깅 모형과 비교해 볼 때 기저 분류기의 평균 예측률은 낮지만, Q-통계량 값이 매우 작은 값을 보여 기저 분류기 간의 다양성 지수가 매우 높다는 것을 알 수 있다. 이로 인해, 다양성 지수가 높은 기저 분류기들의 결합을 통한 앙상블 모형의 성과가 큰 폭으로 개선되었다는 것을 알 수 있다.

5.2 KNN을 기저 분류기로 사용할 경우

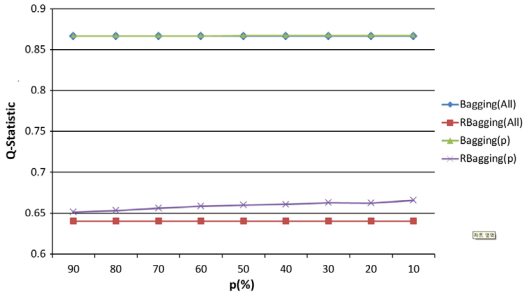
<Figure 5>는 KNN을 기저 분류기로 하는 각 앙상블 모형의 예측성과를 나타내고 있다. <Figure 6>와 <Figure 7>은 앙상블을 구성하고 있는 KNN 기저 분류기들의 평균 예측률과 평균 Q-통계량을 보여주고 있다. KNN 실험에서 파라미터 k는 1로 설정하여 실험하였다.



〈Figure 5〉 Ensemble Classification Performance(%) (Base Classifier : KNN)



〈Figure 6〉 The Average Classification Accuracy of base Classifiers(%)(Base Classifier : KNN)



〈Figure 7〉 Diversity of the base Classifiers of the Ensembles(Base Classifier : KNN)

Breiman(1997)의 실험에 따르면 배깅 앙상블 모형은 DT 모형과 같이 불안정한(unstable) 기저 분류기에서 좋은 성과를 낸다는 것을 알 수 있다. 반면에, KNN과 같이 안정적인(stable) 기저 분류기는 배깅 앙상블 모형에서의 성과 개선이 없는 것으로 나왔다. 이는 KNN이 학습 데이터의 변화에 강건한(robust) 특성이 있기 때문에 배깅 기법을 통해 기저 분류기들의 다양화가 이뤄지기가 힘들고 이로 인해 기저 분류기들을 결합한 앙상블 모형의 성과 개선에 한계가 있는 것이다. <Figure 5>에서 보는 바와 같이 본 실험에서도 선행 연구와 유사한 결과를 보임을 알 수 있다. 즉, KNN 단일 모형을 뜻하는 Single의 그래프가 일반 배깅 앙상블 모형의 성과보다 더 좋은 것을 알 수 있다. 이는 KNN을 기저 분류기로 사용할 경우 기존의 배깅 앙상블 기법으로는 성과 개선이 없다는 것을 의미한다. 반면에 본 논문에서 제안한 RBagging 모형은 단일 모형과 기본 배깅 모형보다 성과가 좋다는 것을 알 수 있다.

<Figure 7>에서 보는 바와 같이 KNN을 기저 분류기로 사용하는 배깅 모형의 경우 Q-통계량의 값이 DT를 기저 분류기로 사용하는 배깅 모형의 Q-통계량 값보다 매우 큰 것을 알 수 있다.

이는 KNN 배깅 모형의 기저 분류기들이 다양성 지수가 매우 낮다는 것을 의미하며, 이로 인해 앙상블 모형의 성과 개선이 없었다는 것을 유추해 볼 수 있을 것이다.

선택적 배깅 모형인 Bagging(p) 모형의 p의 변화에 따른 성과를 살펴보면 <Figure 6>과 <Figure 7>에서 보는 바와 같이 p값이 작아짐에 따라 기저 분류기 평균 예측률은 상승하지만, Q-통계량은 큰 변화가 없는 것을 알 수 있다. 그러므로, Bagging(p) 앙상블 모형의 경우 p값이 작을 경우 좋은 예측률을 보이는 것을 알 수 있으며 p의 값이 40인 경우부터 KNN 단일 모형보다 좋은 성과를 보이기 시작한다는 것을 그림을 통해 알 수 있다.

RBagging 모형의 경우 모든 p의 값에 대해서 단일 모형과 기존의 배깅 모형보다 더 좋은 예측 성과를 보임을 알 수 있다. RBagging 모형의 경우 기존의 배깅 모형과 비교해 볼 때 기저 분류기의 평균 예측률은 낮지만, Q-통계량 값이 매우 작은 값을 보여 기저 분류기 간의 다양성 지수가 매우 높다는 것을 알 수 있다. 이로 인해, 다양성 지수가 높은 기저 분류기들의 결합을 통한 앙상블 모형의 성과가 많이 개선되었다는 것을 알 수 있다.

RBagging(p)의 경우 Bagging(p)와 달리 p값이

감소함에 따라 전반적으로 앙상블 모형의 성과가 하락하고 있음을 알 수 있다. 이는 Bagging(p)의 경우 p값의 감소에 따른 Q-통계량의 변화가 거의 없었던 것과 달리 RBagging(p)의 경우 p값의 감소함에 따라 Q-통계량도 같이 증가함을 알 수 있으며, 이로 인해 앙상블 모형의 성과 개선의 폭이 줄어들고 있음을 알 수 있다.

5.3 최종 실험 결과

<Table 3>은 각 모형의 최종 결과를 비교해 놓은 것이다. 표에서 Best(RBagging(p))는 RBagging(p) 모형에서 가장 좋은 결과를 보인 결과값을 의미하며, Best(Bagging(p))는 Bagging(p) 모형에서 가장 좋은 결과를 보인 결과값을 의미한다.

기저 분류기가 DT일 경우 Bagging(p)의 경우 p의 값이 90일 때 가장 좋은 결과를 보였으며 RBagging(p)의 경우 p의 값이 100일 때 가장 좋은 성과를 냈다. 여기서 p = 100이 의미하는 것은 모든 기저 분류기를 사용했다는 의미로 RBagging(All)을 뜻한다. 즉, DT를 기저 분류기를 사용할 경우에는 RBagging(All) 모형이 가장 좋은 결과를 보였다. 기저 분류기가 KNN일 경우 Bagging(p)의 경우 p의 값이 10일 때 가장 좋은 결과를 보였으며 RBagging(p)의 경우 p의 값이 80일 때 가장 좋은 성과를 냈다. <Table 3>에서 보는 바와 같이 기저 분류기가 DT인 경우, KNN인 경우 모두 본 논문에서 제안한 랜덤화 배깅 모형이 단일 모형과 기존의 배깅 모형보다 좋은 예측성과를 보였다.

본 논문에서는 10-겹 검증 방법으로 실험을 하였으며 앙상블 모형의 경우 각각의 데이터 셋에서 10회 반복하여 실험을 수행하였다. 제안한 모형의 통계적 검정을 위해 t-test를 실시하였으며, 그 결과는 <Table 4>와 <Table 5>와 같다. 표에서 보는 바와 같이 본 논문에서 제안한 RBagging 모형이 DT를 기저 분류기로 사용할 때뿐만 아니라 KNN을 기저 분류기로 사용할 때도 단일 모형과 기존 배깅 모형보다 통계적으로 유의하게 우수한 성과를 냈다는 것을 알 수 있다. 즉, 본 논문에서 제안한 새로운 형태의 배깅 변형 모형이 기존 모형의 예측률을 개선하는 측면에 있어서 매우 효과적이었음을 알 수 있다.

〈Table 3〉 Final Performance Results(Classification Accuracy : %)

	Single	Bagging(All)	Best(Bagging(p))	Rbagging(All)	Best(RBagging(p))
DT	65.68	71.95	72.27	73.35	73.35
KNN	66.49	66.27	67.38	69.84	69.97

〈Table 4〉 t-test(p-Value)(Base Classifier : DT)

DT	Bagging(All)	Best(Bagging(p))	RBagging(All)	Best(RBagging(p))
Single	0.000	0.000	0.000	0.000
Bagging(All)		0.140	0.018	0.018
Best(Bagging(p))			0.045	0.045
RBagging(All)				-

〈Table 5〉 t-test(p-Value)(Base Classifier : KNN)

KNN	Bagging(All)	Best(Bagging(p))	RBagging(All)	Best(RBagging(p))
Single	0.000	0.008	0.000	0.000
Bagging(All)		0.002	0.000	0.000
Best(Bagging(p))			0.000	0.000
RBagging(All)				0.614

6. 결 론

본 논문에서는 앙상블 모형 중에 가장 대표적인 모형인 배깅 앙상블 모형의 성능 개선을 위해 새로운 형태의 변형 배깅 모형을 제안하였다. 제안한 모형에서는 각각의 부트스트랩 샘플을 구성할 때 기존 배깅 모형처럼 부트스트랩 샘플의 크기를 원 학습 데이터의 크기로 고정하는 것이 아니고, 랜덤하게 크기에 변화를 주어 부트스트랩 샘플링을 하게 된다. 이와 같은 방식은 기저 분류기간의 다양성 지수를 높여주어 결국 앙상블 모형의 최종 성과에 긍정적인 요인으로 작용하였다.

본 논문에서 제안한 모형의 우수성을 검증하기 위해 국내 기업의 부도 관련 데이터를 이용하여 다양한 분석을 수행하였다. 실험 결과 본 논문에서 제안한 RBagging 모형의 Q-통계량 값이 일반 배깅 모형의 Q-통계량 값보다 작은 것을 알 수 있다. 이는 본 논문에서 제안한 RBagging 모형이 일반 배깅 모형보다 기저 분류기 다양화 측면에서 우수한 결과를 보였다는 것을 알 수 있다. 반면에, 본 논문에서 제안한 RBagging 모형은 일반 배깅 모형보다 기저 분류기들의 평균 예측률 측면에서는 좋지 않은 결과를 보임을 알 수 있다. 이는, RBagging이 배깅 보다 평균적으로 더 적은 수의 학습 데이터로 기저 분류기를 학습 시켰기 때문으로 해석될 수 있다.

결론적으로 본 논문에서 제안한 RBagging 모형은 기존의 배깅 모형보다 기저 분류기들의 평균 예측률 측면에서는 좋지 않은 결과를 보였지만, 다양성 지수 측면에서 훨씬 더 좋은 결과를 보였으며 결과적으로 이들 기저 분류기들을 결합한 랜덤화 배깅 모형의 성과는 기존 배깅 모형보다 향상된 결과를 보임을 알 수 있었다.

본 연구의 한계와 향후 연구 방향을 정리하면 다음과 같다. 본 연구에서 제안한 모형은 기존의 모형보다 좋은 성과를 보임을 알 수 있었다. 하지만, 향후 보다 다양한 데이터에서의 검증이 더 필요할 것으로 보인다. 본 논문에서 제안한 모형은 기존의 표준 배깅 모형보다 더 작은 샘플을 랜덤하게 사용하여 기저 분류기의 다양성을 제고하였는데 향후에 표준 배깅 보다 더 많은 샘플을 랜덤하게 선택한 모형에 대한 분석도 의미 있는 연구가 될 것이라고 생각된다. 또한, 본 연구에서 사용한 기저 분류기와 다른 분류 모형을 기저 분류기로 사용할 경우 어떠한 차이가 있는지에 대한 추가적인 연구가 필요할 것으로 생각된다.

References

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.

Bryant, S.M., “A Case-Based Reasoning Approach to Bankruptcy Prediction Modeling”, *Intelligent Systems in Accounting, Finance and Management*, Vol.6, No.3, 1997, 195-214.

Choi, H.N. and 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.

(최하나, 임동훈, “앙상블 SVM 모형을 이용한 기업부도 예측”, *한국데이터정보과학회지*, 제24권

제6호, 2013, 1113-1125.)

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.

Kim, J.W. and W.C. Jhee, "Credit Card Bad Debt Prediction Model based on Support Vector Machine", *Journal of Information Technology Services,* Vol.11, No.4, 2012, 233-250.

(김진우, 지원철, "신용카드 대손회원 예측을 위한 SVM 모형", *한국IT서비스학회지*, 제11권, 제4호, 2012, 233-250.)

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., 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.

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

Lu, Y., N.Y. Zeng, X.H. Liu, and S.J. Yi, "A New Hybrid Algorithm for Bankruptcy Prediction Using Switching Particle Swarm Optimization and Support Vector Machines", *Discrete Dynamics in Nature and Society,* Vol. 2015, 2015 http://dx.doi.org/10.1155/2015/294930(Downloaded February 19, 2016.)

Meyer, P.A. and H. Pifer, "Prediction of Bank Failures", *The Journal of Finance*, Vol.25, 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 기업부도예측", *지능정보연구*, 제20권, 제4호, 2014, 121-139.)

Min, S., "Optimization of Random Subspace Ensemble for Bankruptcy Prediction", *Journal of Information Technology Services*, Vol.14, No.4, 2015, 121-135.

(민성환, "재무부실화 예측을 위한 랜덤 서브스페이스 앙상블 모형의 최적화", *한국IT서비스학회지*, 제14권, 제4호, 2015, 121-135.)

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

Shaw, M.J. and J.A. Gentry, "Using and Expert System with Inductive Learning to Evaluate Business Loans", *Financial Manage-*

ment, Vol.17, No.3, 1988, 45-56.

Yu, L., "Credit Risk Evaluation with a Least Squares Fuzzy Support Vector Machines Classifier", *Discrete Dynamics in Nature and Society,* 2014.

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, 1999, 16-32.

Zhang, Y.D., S.H. Wang, and G.L. Ji, "A Rule-Based Model for Bankruptcy Prediction Based on an Improved Genetic Ant Colony Algorithm", *Mathematical Problems in Engineering,* 2013.

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.