J. Inf. Technol. Appl. Manag. 23(1): 45~59, March 2016
http://dx.doi.org/10.21219/jitam.2016.23.1.045

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

유전자 알고리즘 기반 통합 앙상블 모형

민 성 환*

Genetic Algorithm based Hybrid Ensemble Model

Sung-Hwan Min*

Abstract

An ensemble classifier is a method that combines output of multiple classifiers. It has been widely accepted that ensemble classifiers can improve the prediction accuracy. Recently, ensemble techniques have been successfully applied to the bankruptcy prediction. Bagging and random subspace are the most popular ensemble techniques. Bagging and random subspace have proved to be very effective in improving the generalization ability respectively. However, there are few studies which have focused on the integration of bagging and random subspace. In this study, we proposed a new hybrid ensemble model to integrate bagging and random subspace method using genetic algorithm for improving the performance of the model. The proposed model is applied to the bankruptcy prediction for Korean companies and compared with other models in this study. The experimental results showed that the proposed model performs better than the other models such as the single classifier, the original ensemble model and the simple hybrid model.

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

Received : 2016. 01. 21. 1st Revised : 2016. 02. 15. 2nd Revised : 2016. 03. 18. Final Acceptance : 2016. 03. 22.

* Department of Business Administration, Hallym University 39 Hallymdaehak-gil, Chuncheon Gangwon-do, 24252, Korea,
Tel: +82-33-248-1841, Fax: +82-33-256-3424, e-mail : shmin@hallym.ac.kr

1. 서 론

기업의 재무부실화 예측은 경영 분야에서 중요하게 다뤄지고 있는 연구 주제로 많은 연구자들이 보다 정확한 재무부실화 예측 모형 개발을 위해 많은 노력을 기울여왔다. 재무 부실화 예측에 관한 초기의 연구는 통계적인 모형을 재무 부실화 예측 문제에 적용해 보는 연구가 대부분이었으며, 단일변량 분석, 다변량 판별 분석, 다중회귀분석, 로지스틱 회귀 분석 등과 같은 통계 모형을 기업의 부실화 예측 문제에 적용해 본 연구가 여기에 속한다[Beaver, 1966; Altman, 1968; Meyer and Pifer, 1970; Ohlson, 1980]. 그 뒤로 통계 모형보다 더 정확한 모형 개발을 위해 인공신경망, 사례기반추론, 의사결정 트리 등과 같은 다양한 데이터 마이닝 기법들을 활용한 연구가 활발하게 진행되었다[Tam and Kiang, 1992; Buta, 1994; Messier and Hansen, 1998; Zhang et al., 1999].

최근에는 좋은 예측 성과로 데이터 마이닝 분야에서 각광 받고 있는 앙상블 기법을 재무 부실화 문제에 적용하는 연구가 활발하게 진행되고 있다. Kim and Kim[2007]은 가장 대표적인 앙상블 기법중의 하나인 배깅(bagging) 모형을 SOHO의 부도 예측 문제에 적용해 보았다. 또한, 기존의 배깅 모형의 성과 개선을 위해 기저 분류기(base classifier) 중에 예측성과가 좋은 일부를 선택하여 앙상블을 구성하는 선택적 배깅 모형을 제안하여 의사결정 트리의 일종인 CART를 기저 분류기로 하여 실험을 수행하였다. Kim[2009]은 기업의 부도 예측 문제에 배깅과 부스팅(boosting) 앙상블을 적용해 보았다. 의사결정 트리, 인공신경망 등과 같은 분류기를 대상으로 다양한 실험을 하였으며, 실험 결과 의사결정 트리, 인공신경망의 단일 모형보다 이들을 결합한 앙상블 모형이 성과가 좋음을 알 수 있었다. Li et al. [2011]은 로짓 모형을 기저분류기로 하는 랜덤 서브스페이스(random subspace) 앙상블 모형을 중국기업의 부도 예측 문제에 적용해 보았으며 실험 결과 앙상블 모형의 성과가 단일 모형보다 더 좋은 성과를 보였다. Choi and Lim[2013]은 여러 가지 커널 함수에 따른 서로 다른 형태의 SVM 모형들을 결합하는 앙상블 모형을 부도 예측 문제에서 적용해 앙상블 모형의 우수성을 보였다. Min[2014]은 배깅과 사례 선택(instance selection)을 결합하는 새로운 모형을 국내 기업의 부도 예측 문제에 적용해 보았으며, 제안한 모형이 기존의 배깅 모형과 사례 선택 기법을 적용한 모형 보다 더 우수한 성과를 보임을 알 수 있었다. Kim et al.[2015]은 기존의 부스팅 알고리즘을 개선하기 위해 기하 평균 개념을 활용한 부스팅 알고리즘을 제안하여 부도 예측 문제에 적용해 보았으며 실험 결과 제안한 모형이 우수한 성능을 보였다.

한편, 최근 앙상블 관련 연구 중에서 두 개의 앙상블 기법을 결합하여 앙상블 모형의 성과를 향상시키려는 연구가 시도되고 있다. Min[2012]은 SVM을 기저분류기로 하는 배깅과 랜덤 서브스페이스의 통합 모형을 국내 부도 예측 문제에 적용해 보았으며, 통합 앙상블 모형이 단일 앙상블 모형 보다 좋은 성과를 보임을 알 수 있었다. Marques et al.[2012]는 의사결정 트리(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.

이와 같이 최근에 각광 받고 있는 앙상블 모형을 부도 예측 문제에 적용하려는 다양한 연구가 있었지만 아직까지 여러 앙상블 기법을 통합하는 연구는 많지 않은 것이 현실이다. 또한, 기존의 통합 모형도 단순히 두 개의 모형을 결합하는 모형이 대부분이다. 이에 본 논문에서는

부도 예측 모형의 성과 개선을 위해 새로운 형태의 통합 앙상블 최적화 모형을 제안하고자 한다. 기존의 앙상블 통합 모형은 단순한 두 개 앙상블 모형의 결합으로 이루어져 있지만 본 연구에서는 두 개의 앙상블 기법의 단순한 결합이 아닌 유전자 알고리즘을 이용한 최적화 모형을 제안하였다. 또한, 본 논문에서 제안한 모형의 성과를 검증하기 위해 실제 기업데이터를 이용해 제안한 모형의 우수성을 검증하였다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 앙상블 분류기에 대한 설명을 하고 제 3장에서는 본 논문에서 제안한 배깅과 랜덤 서브스페이스 앙상블 통합 모형의 최적화 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. 앙상블 분류기

앙상블 학습이란 동일한 문제를 풀기 위해 다수의 분류기를 활용하는 기법으로, 일반적으로 하나의 분류기를 사용할 때보다 좋은 성과를 낼 수 있는 것으로 알려져 있으며 이로 인해 최근에 데이터 마이닝 분야에서 많은 관심을 끌고 있다[Dietterich, 1997]. 앙상블 분류기는 다수의 기저 분류기 생성과 이들의 결합이라는 두 가지 단계로 구성되어 있다. 일반적으로 성과가 좋은 앙상블 분류기를 얻기 위해서는 앙상블을 구성하고 있는 기저 분류기가 가능하면 정확해야 하고 또한 가능하면 다양성을 가져야 한다.

Hansen과 Salamon[1990]에 의하면 기저 분류기들의 예측 오차가 0.5보다 작고, 이들의 오차가 모두 독립적이라면 이들을 다수결 투표 방식으로 결합한 앙상블 분류기의 오차는 앙상블 크기가 증가함에 따라 0에 수렴됨을 이론적으로 보여주고 있다. 즉, 이론적으로는 기저 분류기의 예측정확도가 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 위해서는 개별 분류기의 예측성과 뿐만 아니라 기저 분류기들 간의 다양성이 매우 중요함을 알 수 있다. 즉, 앙상블 모형의 성과가 좋으려면 앙상블을 구성하고 있는 기저 분류기들이 서로 다양성을 가져야한다. 각각의 기저 분류기들의 예측 결과가 서로 다른 결과를 가져야만 한 개의 분류기가 오분류를 할 때 이와 다른 결과 값을 갖는 다른 분류기들을 통해 오분류를 피할 수 있을 것이다.

앞에서 살펴본 바와 같이 앙상블 모형은 그것을 구성하고 있는 기저 분류기들을 다양화시키는 것이 매우 중요하다. 기저 분류기들을 다양화시키는 방법으로는 학습 데이터에 변화를 주는 방법, 학습 파라미터에 변화를 주는 방법, 서로 다른 학습 알고리즘을 이용하는 방법 등이 있으며 이중에서 가장 대표적인 기법은 학습 데이터를 다양화시킴으로써 기저 분류기들을 다양화시키는 방법으로 배깅[Breiman, 1996], 부스팅[Freund and Schapire, 1996]과 랜덤 서브스페이스 기법[Ho, 1998]이 여기에 속한다.

본 논문에서는 배깅과 랜덤 서브스페이스 앙상블 기법을 통합하고, 이 통합 앙상블 모형을 최적화하는 새로운 형태의 모형을 제안하였다. 배깅은 사례 공간(instance space)에서의 변화를 통해 기저 분류기를 발생시키고, 랜덤 서브스페이스는 입력변수 공간(feature space) 상에서의 변화를 통해 기저 분류기를 발생시키는 기법이다. 본 논문에서 사용한 랜덤 서브스페이스와 배깅 앙상블 기법의 전반적인 절차는 <Table 1>과 <Table 2>에 나와 있다.

〈Table 1〉 Steps of Random Subspace Ensemble Method

Random Subspace Ensemble
1. Partition Data Set(Training Data Set : T, Validation Data Set : V) 2. Generate a new training data set with randomly selected f' features from T (f' < F(total number of features in T) 3. Repeat Step 2 to generate n new training data sets → $T(RS)_1$, $T(RS)_2$, …, $T(RS)_n$ 4. Train a learning algorithm using each new training set(Different base classifiers are generated) → C_1, …, C_n 5. Apply the base classifiers to the validation data set → n different output data(O_1, …, O_n) 6. Combine the output data(O_1, …, On) by a combining method

〈Table 2〉 Steps of Bagging Ensemble Method

Bagging Ensemble
1. Partition Data Set(Training Data Set : T, Validation Data Set : V) 2. Generate a new training data set with randomly selected N' instances from T 3. Repeat Step 2 to generate n new training data sets → $T(B)_1$, $T(B)_2$, …, $T(B)_n$ 4. Train a learning algorithm using each new training set → Different base classifiers are generated(C_1, …, C_n) 5. Apply the base classifier 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

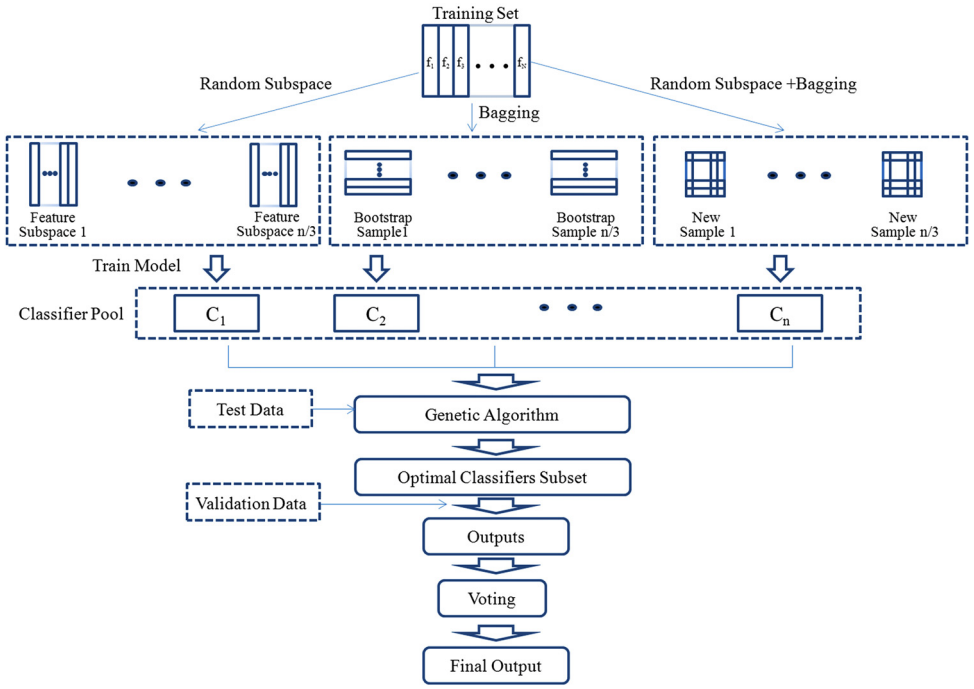
3. 연구 모형

본 연구는 기업의 부도 예측을 위한 앙상블 분류기의 성능 개선에 관한 연구이다. 이를 위해 배깅과 랜덤 서브스페이스 기법을 통합하고 이를 최적화하는 새로운 모형을 제안하였다. 최근에 각광 받고 있는 앙상블 모형을 부도 예측 문제에 적용하려는 다양한 연구가 있었지만 아직까지 여러 앙상블 기법을 통합하는 연구는 많지 않은 것이 현실이다. 또한, 기존의 통합 모형도 단순히 두 개의 모형을 결합하는 모형이 대부분이다. 이에 본 논문에서는 부도 예측 모형의 성과 개선을 위해 새로운 형태의 통합 앙상블 최적화 모형을 제안하고자 한다. 기존의 앙상블 통합 모형은 단순한 두 개 앙상블 모형의 결합으로 이루어져 있지만 본 연구에서는 두 개의 앙상블 기법의 단순한 결합이 아닌 유전자 알고리즘을 이용한 최적화 모형을 제안하였다.

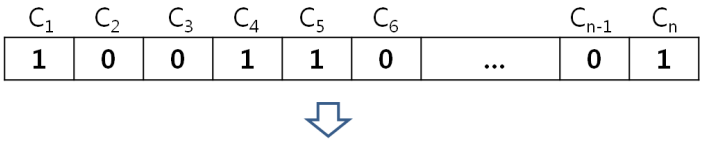
본 연구에서는 배깅과 랜덤 서브스페이스의 결합을 통해 다양성을 갖는 기저 분류기 풀을 구성한 후 이 중에서 최적의 기저 분류기 조합을 선택하기 위해 유전자 알고리즘을 활용하였으며, 전반적인 절차는 <Figure 1>과 같다.

<Figure 1>에서 보는 바와 같이 랜덤 서브스페이스와 배깅 기법을 통해 각각 서로 다른 기저 분류기를 생성하고 또한 이 두 기법의 결합을 통해 새로운 기저분류기를 생성한다. 이처럼 원 학습 데이터로부터 각각 다양한 방법에 의해 생성된 서로 다른 기저 분류기들의 조합인 기저 분류기 풀{C_1, …, C_n}을 생성한다.

일반적인 통합 앙상블 모형의 경우 원 학습데이터(original training set)로부터 랜덤하게 입력변수 집합을 선택한 후, 랜덤하게 일정 비율의 데이터를 선택하여 서로 다른 학습데이터를 생성하는 방식으로 랜덤 서브스페이스 기법과 배깅 기법을 통합한다(또는, 순서를 반대로 하여 배깅 기법을 먼저 적용한 데이터를 가지고 랜덤 서브스페이스 기법 적용할 수도 있다). 또한 이와 같은 방법으로 생성된 분류기 셋 {C_1, …, C_n}을 가지고 앙상블을 구성하게 되며, 개별 분류기의 출력 값을 특정한 전략에 따라 결합하게 된다. 본 연구에서는 기존의 통합 모형과 달리



〈Figure 1〉 The Overall Architecture of the Proposed Model



Selected Classifier Subset={C_1, C_4, C_5, ... ,C_n}

〈Figure 2〉 Example of Encoding for a Genetic Algorithm

랜덤 서브스페이스 기법과 배깅 기법을 각각 독립적으로 수행하여 생성된 기저 분류기와 두 개 기법을 동시에 진행하여 생성한 기저 분류기를 모두 기저 분류기 후보로 등록한 후 최적의 분류기 조합 선택을 위해 유전자 알고리즘을 활용하였다.

본 연구에서 제안한 방법은 기존의 통합 앙상블 모형보다 두 가지 점에서 장점이 있다고 볼 수 있다. 첫째로, 기존의 통합 앙상블 모형보다 더 다양한 형태의 기저 분류기를 대상으로 통합 앙상블 모형을 구성한다는 장점이 있다. 둘째로, 이런 다양한 기저 분류기들 중에서 최적의 조합을 선택하여 앙상블을 구성한다는 장점이 있다. 생성된 모든 기저 분류기를 결합하는 것이 아니고, 결합하였을 경우 앙상블 모형의 성과 측면에서 가장 좋은 성과를 낼 수 있는 기저 분류기들의 조합을 유전자 알고리즘을 이용하여 탐색하고 이를 통해 최종 모형을 구성한다는 점에서 기존의 단순한 통합 앙상블 모형보다 성과 측면에서 장점이 있다고 볼 수 있다.

제안한 모형에서 유전자 알고리즘의 염색체는 <Figure 2>와 같이 0과 1의 값을 갖는 이진열(binary string)형태로 표현하여 각각의 비트(bit)는 기저 분류기와 대응되도록 설계하였다. 각각

〈Table 3〉 Steps in the Proposed Model

Genetic algorithm based hybrid ensemble model
1. Partition Data Set(Training Data Set(T_A), Test Data Set(T_B), Validation Data Set(V)) 2-1. Generate a new training data set with randomly selected f' features from T_A (f' < F(total number of features in T_A) 2-2. Generate a new training data set with randomly selected N' instances from T_A 3-1. Repeat Step 2-1 to generate n/3 new training data sets → $T_A(RS)_1$, $T_A(RS)_2$, …, $T_A(RS)_{n/3}$ 3-2. Repeat Step 2-2 to generate n/3 new training data sets → $T_A(B)_1$, $T_A(B)_2$, …,$T_A(B)_{n/3}$ 3-3. Generate a new training data sets with randomly selected N' instances from $T_A(RS)_i$(i = 1, …, n/3) → $T_A(RS+B)_1$, $T_A(RS+B)_2$, …, $T_A(RS+B)_{n/3}$ 4. Train a learning algorithm using each new training set generated in Step 3 → Different n base classifiers are generated(C_1, …, C_n) 5. Define the chromosome corresponding to the classifier pool(C_1, …, C_n) generated in Step 4. (The chromosome for the classifier pool is encoded as a form of binary string) 6. Determine parameters of GA 7. Generate the initial population 8. Select the classifier subset for each chromosome 9. Apply each classifier subset generated in step 8 to the test data set(T_B) 10. Calculate the fitness value of each classifier subset 11. Repeat GA operations and create a new generation 12. Repeat from step 8 to 11 until the termination criteria are satisfied 13. Select the optimal classifier subset 14. Apply the optimal classifier subset to the validation data set(V) 15. Combine the output data(O_1, …, O_n) by the majority voting scheme

의 비트에서의 값은 해당되는 분류기의 선택 유무를 알려주게 된다. 예를 들면, <Figure 2>에서 첫 번째 비트의 값은 1로 이는 대응되는 분류기 C_1이 선택된다는 것을 의미하며, 두 번째 비트의 값 0은 대응되는 분류기 C_2가 선택되지 않았다는 것을 의미한다. 이와 같은 방식으로 <Figure 2>에서는 전체 분류기 풀 중에서 {C_1, C_4, C_5, …, C_n}의 분류기들이 선택되게 된다. 본 논문에서는 유전자 알고리즘을 통한 최적의 기저 분류기 조합을 찾기 위해 테스트용 데이터에서의 예측 정확도를 적합도 함수로 사용하였다. 본 논문에서 제안한 모형인 유전자 알고리즘을 이용한 배깅과 랜덤 서브스페이스 앙상블의 최적화 통합 모형의 전반적인 흐름은 <Table 3>에 나와 있다.

4. 실험 설계

본 연구에서 제안한 모형의 검증을 위해 부도 기업의 데이터 900개와 비부도 기업의 데이터 900개로 구성된 총 1,800개의 국내 비외감 기업의 데이터를 사용하였다. 실험에 사용된 데이터는 자산규모가 10억에서 70억 사이인 중공업 데이터로 1999년부터 2002년 사이의 재무 데이터로 구성되어 있다. 부도기업은 90일 이상 연체한 경우, 파산선고 또는 유사한 소송을 제기한 경우, 상당한 경제적 손실을 감수하고 신용채권을 매각한 경우 등에 해당하는 기업으로 구성하였으며 비부도 기업은 부실 사유에 해당하지 않는 기업으로 구성하였다.

데이터는 모형의 학습을 위한 학습 데이터와 과적합(overfitting)을 피하기 위해 사용한 테스트용 데이터, 그리고 모형의 검증을 위해 사용한 검증용 데이터로 분류하여 실험을 하였다.

본 연구에서는 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

〈Table 4〉 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 Owner's Equity
	Quick Asset to Current Liability
	Debt Ratio
	Current Liability to Total Asset
	Current Asset to Current Liability
	(Capital surplus+retained earnings-dividend)/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

활용하였다. 앙상블 모형의 경우 10회 반복하여 실험을 수행하였으며, 10회 실험 결과의 평균값을 대푯값으로 사용하였다. 앙상블 모형의 기저 분류기로는 의사결정 트리(Decision tree) 모형을 사용하였으며 각각의 앙상블 모형에서 기저 분류기의 총 수는 100으로 고정하고 실험을 하였다.

본 논문에서는 기업의 부도 여부 예측을 위해 재무비율을 입력 변수로 사용하였다. 수익성, 안정성, 성장성, 활동성 및 현금 흐름으로 분류된 총 131개의 재무비율을 대상으로 단일 표본 t검정(independent-samples t-test)과 후진 선택법(backward selection)을 이용한 로지스틱 회귀분석을 통해 최종 변수를 선정하였으며 그 결과는 <Table 4>와 같다.

5. 실험 결과

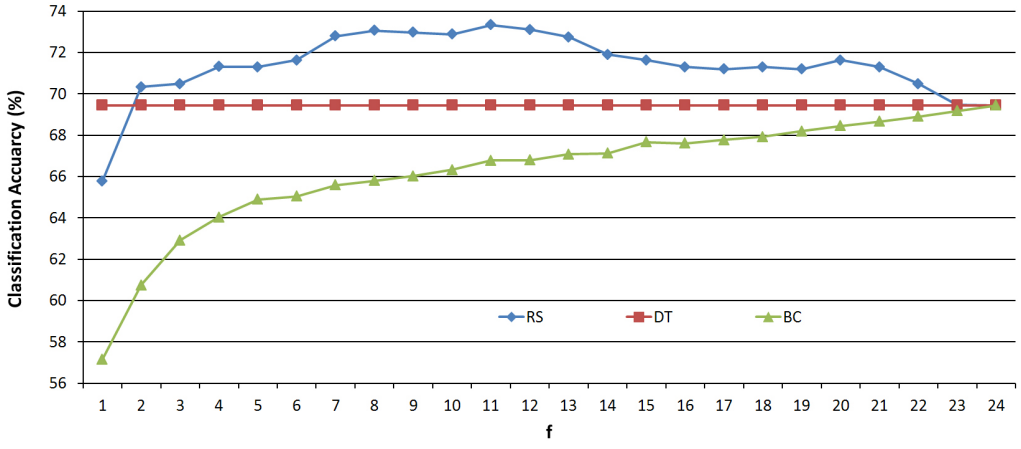
본 연구에서 제안한 모형의 검증을 위해 검증용 데이터에서의 예측 정확도를 비교해 보았다. 본 연구에서는 랜덤 서브스페이스 앙상블 모형과 배깅 앙상블 모형의 다양한 통합 모형과 통합 모형의 최적화 모형을 제안하였다. 제안한 모형의 성과 비교에 앞서 각각의 개별 앙상블 모형에 대한 탐색적 연구를 수행하였으며 그 결과는 <Figure 3>과 <Figure 4>에 나와 있다.

<Figure 3>은 전체 입력변수의 수(F = 24) 중에서 랜덤하게 선택한 입력변수의 수(f)에 따른 랜덤 서브스페이스 앙상블의 성과를 보여주고 있다. 여기서 DT는 의사결정 트리 단일 모형을 의미하며, RS는 의사결정 트리를 기저 분류기로 사용하는 랜덤 서브스페이스 앙상블 모형을 의미한다. BC는 랜덤 서브스페이스 앙상블 모형을 구성하고 있는 기저 분류기들을 의미한다. <Figure 3>에서의 값은 각 모형에서의 예측률을 의미하며, BC에서의 값은 랜덤 스페이스 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률을 의미한다. 의사결정 트리 단일 모형의 경우 <Table 4>에 있는 모든 입력 변수를 사용하여 모형을 구성하였으며, 랜덤 서브스페이스 모형의 경우 전체 입력 변수 중에서 랜덤하게 f개 만큼 선택하여 기저 분류기를 구성한 후 이들을 다수결 투표방식에 의해 결합하였다. 본 실험에서는 비복원 추출 방식을 이용하여 랜덤하게 입력 변수를 선택하였다. 실험 결과에서 알 수 있듯이 랜덤 서브스페이스 앙상블의 예측률은 f의 값이 커질수록 증가하다가 f의 값이 11일 때 가장 좋은 성과를 보인 후 그 이후로 점차로 감소함을 알 수 있다. <Figure 3>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 있

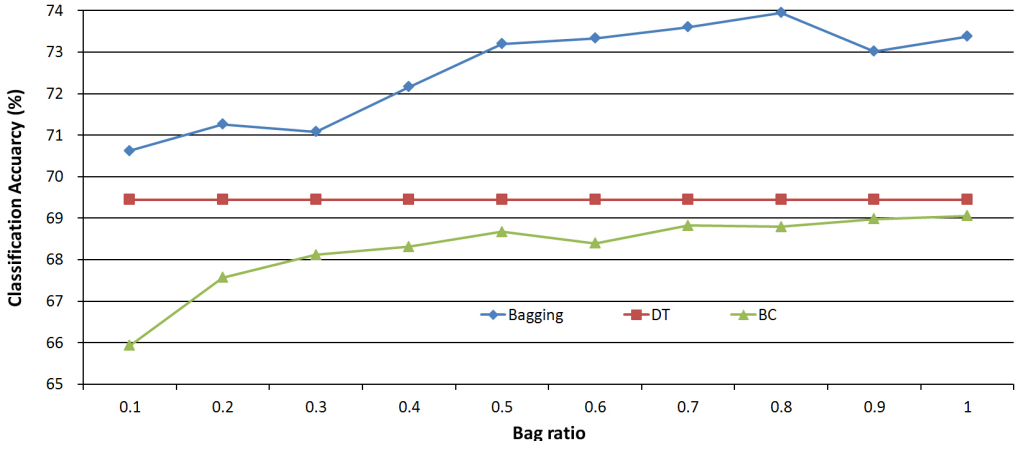
음을 알 수 있다. f = 1인 경우, 즉 입력 변수 한 개만을 가지고 기저 분류기를 구성하는 경우 앙상블 모형의 예측률이 65.78%로 단일 모형(DT)보다 성과가 좋지 않음을 알 수 있다. 이는 한 개의 입력 변수만을 사용한 기저 분류기들의 평균 예측률이 57.16%로 매우 낮기 때문에 이들의 결합을 통한 앙상블 모형의 성과 개선에 한계가 있다는 것을 의미한다.

본 연구에서 랜덤 서브스페이스 앙상블 모형에서 비복원 추출로 기저 분류기를 구성했으므로 f = 24인 경우는 모든 변수를 사용하여 기저 분류기들을 구성한 것을 의미한다. 즉, f = 24일 때 앙상블을 구성하고 있는 100개의 기저 분류기는 모두 단일 모형 DT와 동일한 모형이다. 이와 같은 이유로 f = 24일 때 RS, BC, DT는 같은 값을 갖는 다는 것을 알 수 있다. 즉, 모두 동일한 기저 분류기들을 결합할 경우 성과 개선이 전혀 없음을 알 수 있다. 이와 같이, 앙상블 모형의 성과는 기저 분류기들의 평균 예측률과 다양성이 중요함을 알 수 있다.

<Figure 3>에서 보는 바와 같이 f의 값이 2 이상인 경우부터 랜덤 서브스페이스 앙상블 모형의 성과가 단일 모형보다 더 좋아 지는 것을 알 수 있다. Ho[1998]의 실험에 의하면 f의 값



〈Figure 3〉 Sensitivity Analysis of Random Subspace Ensemble



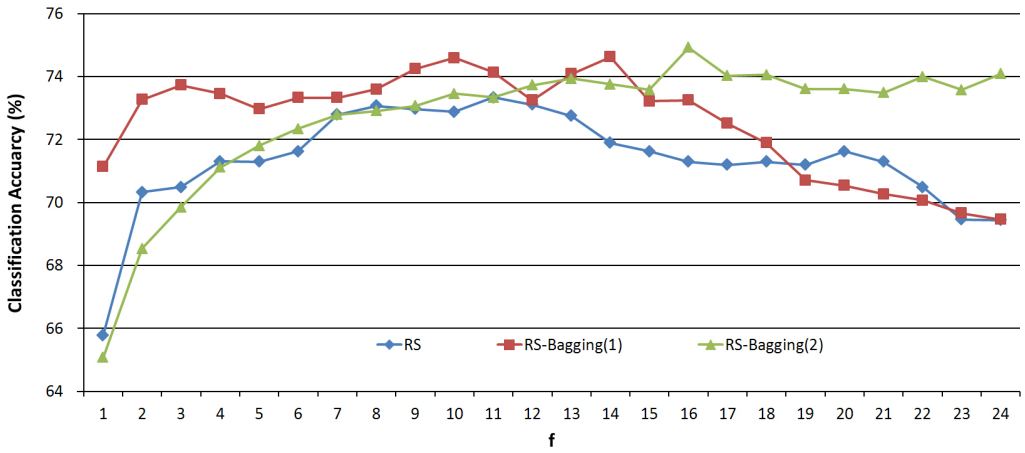
〈Figure 4〉 Sensitivity Analysis of Bagging Ensemble

이 전체 입력 변수 총수의 0.5에 해당할 경우 가장 좋은 결과 또는 가장 좋은 결과와 근사한 결과를 냈다고 보고하고 있다. 본 실험에서는 <Figure 3>에서 보는 바와 같이 전체 입력 변수의 총 수(F = 24)의 반인 f = 12에서 두 번째로 좋은 결과를 보였으며, f = 11에서 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.

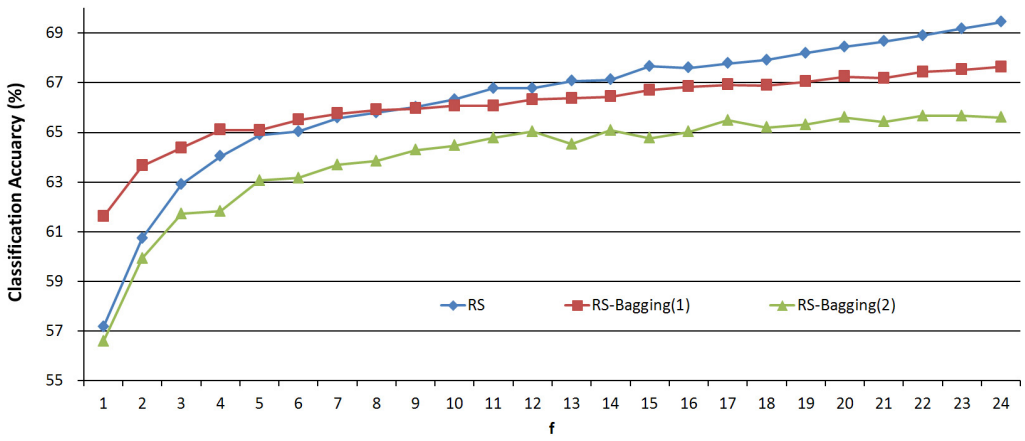
배깅 앙상블 모형의 성과에 영향을 주는 대표적인 파라미터로는 앙상블을 구성하는 기저 분류기의 총 수와 원 데이터로부터 랜덤하게 샘플링한 데이터의 수가 있다. 본 연구에서는 원 학습데이터의 크기 중에서 복원 추출방식으로 랜덤하게 선택한 부트스트랩 표본(bootstrap sample)의 크기의 비율을 Bag ratio로 정의하고 이에 따른 성과를 비교 분석하였으며 그 결과는 <Figure 4>와 같다. 그림에서 보는 바와 같이 배깅 앙상블은 모든 Bag ratio에서 단일 분류기 모형보다 좋은 성과를 보임을 알 수 있다. 배깅 앙상블 모형의 성과는 Bag ratio 값이 커짐에 따라 증가하다가 Bag ratio의 값이 0.8일 때 가장 좋은 값을 보임을 알 수 있다. Bag ratio가 증가할수록 배깅 앙상블을 구성하고 있는 분류기들의 평균 예측률 값이 증가하는 것을 알 수 있으며, 이는 랜덤 서브스페이스 앙상블 모형의 경우와 같은 이유로 해석할 수 있다. 즉, 더 많은 데이터를 사용할수록 기저 분류기들의 예측 성과가 향상됨을 알 수 있다.

〈Table 5〉 Experimental Results

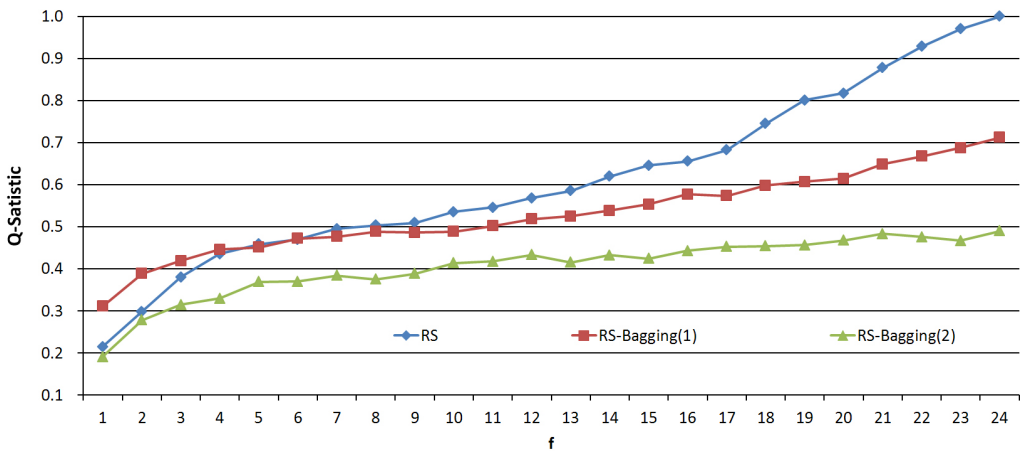
f	DT	RS			RS-Bagging(1)			RS-Bagging(2)		
		EOC	BC	Q	EOC	BC	Q	EOC	BC	Q
1	69.44	65.78	57.16	0.21	71.14	61.60	0.31	65.08	56.58	0.19
2		70.32	60.74	0.30	73.27	63.65	0.39	68.53	59.93	0.28
3		70.49	62.91	0.38	73.73	64.37	0.42	69.85	61.72	0.31
4		71.31	64.03	0.44	73.46	65.10	0.45	71.11	61.82	0.33
5		71.30	64.89	0.46	72.97	65.08	0.45	71.80	63.06	0.37
6		71.62	65.03	0.47	73.32	65.50	0.47	72.34	63.17	0.37
7		72.79	65.58	0.50	73.32	65.76	0.48	72.79	63.70	0.38
8		73.06	65.79	0.50	73.59	65.90	0.49	72.91	63.84	0.38
9		72.97	66.01	0.51	74.24	65.96	0.49	73.06	64.29	0.39
10		72.88	66.32	0.54	74.59	66.07	0.49	73.45	64.47	0.41
11		**73.33**	**66.78**	**0.55**	74.14	66.08	0.50	73.33	64.78	0.42
12		73.11	66.78	0.57	73.24	66.32	0.52	73.72	65.04	0.43
13		72.75	67.07	0.59	74.08	66.38	0.53	73.93	64.53	0.42
14		71.89	67.12	0.62	**74.62**	**66.43**	**0.54**	73.75	65.08	0.43
15		71.62	67.66	0.65	73.22	66.70	0.55	73.57	64.77	0.42
16		71.30	67.60	0.66	73.24	66.84	0.58	**74.92**	**65.02**	**0.44**
17		71.19	67.77	0.68	72.51	66.91	0.57	74.02	65.49	0.45
18		71.30	67.92	0.74	71.89	66.89	0.60	74.05	65.20	0.45
19		71.19	68.19	0.80	70.70	67.05	0.61	73.60	65.31	0.46
20		71.62	68.44	0.82	70.54	67.24	0.61	73.60	65.61	0.47
21		71.30	68.66	0.88	70.27	67.19	0.65	73.48	65.42	0.48
22		70.49	68.90	0.93	70.07	67.44	0.67	73.99	65.66	0.48
23		69.46	69.17	0.97	69.66	67.53	0.69	73.57	65.66	0.47
24		69.44	69.44	1.00	69.46	67.64	0.71	74.08	65.60	0.49
Mean	69.44	71.35	66.25	0.61	72.55	66.07	0.53	72.69	63.99	0.41
Maximum	69.44	73.33	69.44	1.00	74.62	67.64	0.71	74.92	65.66	0.49



〈Figure 5〉 Ensemble Classification Performance



〈Figure 6〉 Average Classification Accuracy of base Classifiers in Each Ensemble Model



〈Figure 7〉 Average Q-Statistic Value of base Classifiers in Each Ensemble Model

본 연구에서는 단일 앙상블 모형뿐만 아니라 두 앙상블 모형의 통합에 관한 다양한 실험을 수행하였다. 그 결과는 <Table 5>, <Figure 5>, <Figure 6>, <Figure 7>에 나와 있다.

<Table 5>는 f의 변화에 따른 다양한 통합 앙상블 모형과 단순 앙상블 모형의 결과를 보여주고 있다. 표에서 DT는 단일 모형을 의미하고 RS는 단순 랜덤 서브스페이스 앙상블 모형을 의미한다. 또한 RS-Bagging(1)과 RS-Bagging(2)는 랜덤 서브스페이스와 배깅을 통합한 통합 앙상블 모형을 의미한다. RS-Bagging(1)은 랜덤 서브스페이스와 배깅 모형을 통해 각각 서로 다른 분류기를 생성한 후 이들을 통합한 모형을 의미하며, RS-Bagging(2) 모형은 먼저 랜덤 서브스페이스 기법을 통해 원 데이터로부터 랜덤하게 입력 변수를 선택한 후, 선택된 데이터 셋을 가지고 배깅 기법을 적용하여 기저 분류기를 생성한 통합 모형을 의미한다. 표에서 EOC는 분류기 앙상블의 예측성과를 나타내고, BC는 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률을 의미한다. Q는 대표적인 다양성 지표인 Q-통계량을 의미한다.

앙상블의 성과는 앙상블을 구성하고 있는 기저 분류기의 평균 예측률(BC)과 앙상블을 구성하고 있는 기저 분류기 간의 다양성 지수가 중요한 영향을 미친다. 그러므로, 앙상블의 모형의 성과를 이해하기 위해서는 이들을 함께 분석하는 것이 매우 중요하다. 지금까지 기저 분류기들 간의 다양성을 측정하기 위한 많은 방법들이 제안되어 왔으며, 본 연구에서는 가장 대표적인 다양성 척도 중의 하나인 Q-통계량을 살펴보았다. 두 개의 분류기 C_i와 C_j가 있다고 가정하고 이들간의 예측 오차에 대한 일치 정도가 <Table 6>과 같다고 한다면 이들 간의 Q-통계량은 식 (2)와 같이 구할 수 있으며 <Table 5>에서의 Q는 앙상블을 구성하고 있는 각 분류기의 Q-통계량의 평균값을 의미한다[Kuncheva and Whitaker, 2003].

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

표에서 보는 바와 같이 평균값을 기준으로 RS-Bagging(2)이 가장 좋은 예측성과를 보였으며, 기저 분류기의 평균 예측률은 RS가 가장 높다는 것을 알 수 있다. 또한, Q 통계량 값은 RS-Bagging(2) 모형이 가장 낮은 값인 0.41을 나타내고 있다. 이는 RS-Bagging(2) 앙상블을 구성하고 있는 기저 분류기들의 다양성 지표 값이 가장 좋다는 것을 뜻한다.

<Figure 5>는 각 앙상블 모형의 f값에 따른 예측률 변화를 보여주고 있다. RS와 RS-Bagging(1)의 경우 f = 12 전후로 가장 좋은 예측성과를 내다가 그 이후로 점차 감소함을 알 수 있다. 반면에 RS-Bagging(2)의 경우 f = 12 이후에도 예측성과의 감소가 없음을 알 수 있다. 이는 f = 12이후에도 다른 앙상블 모형과 달리 다양성 지수 측면에서 좋은 결과를 보이고 있는 것과 관련이 있는 것으로 분석할 수 있을 것이다.

<Figure 6>에서 보는 바와 같이 각각의 앙상블 모형은 f의 값이 증가함에 따라 기저 분류기들의 평균 예측률 값이 좋아 짐을 알 수 있다. 반대로 <Figure 7>에서 보는 바와 같이 각각 앙상블 모형의 Q 통계량의 값은 f의 값이 증가함에 따라 커짐을 알 수 있으며, 이는 f의 값이 증가함에 따라 각 기저 분류기 간의 다양성이 감소하고 있다는 것을 의미한다. 모두 동일한 기저 분류기로 구성된 경우인 f = 24일 경우에는 Q-통계량의 값이 1이 됨을 알 수 있다. <Table 5>에서 보는 바와 같이 Q-통계량이 커짐에 따라 기저 분류기를 결합하여 앙상블 모형을 구성할 경우의 성과 개선의 폭이 줄어드는 것 알 수 있다.

본 논문에서는 <Table 5>의 실험에 사용된 통합 앙상블 모형의 성능 개선을 위해 유전자 알고리즘을 사용한 최적화 모형을 제안하였다. 최종 비교를 위해 사용한 비교 모형의 값은 <Table 5>에서 가장 좋은 성과를 보였을 때의 값을 각 모형의 대푯값으로 사용하였다. 또한, 본 연구에서 제안한 모형에서 Bag ratio 값과 f 값은 단일 앙상블 모형에서 가장 좋은 성과를 보인 값을 사용하여 실험하였다. 각 모형 별 최종 예측성과는 <Table 7>과 같다.

<Table 7>에서 보는 바와 같이 다양한 통합 앙상블 모형의 기저 분류기 중에서 유전자 알고리즘을 이용하여 찾아낸 최적의 분류기 조합으로 이루어진 앙상블 모형인 GARSBagging의 값이 가장 좋은 예측 정확도를 보임을 알 수 있다. 앞에서 살펴본 바와 같이 본 연구에서 제안한 GARSBagging 모형은 기존의 통합 앙상블 모형보다 더 다양한 형태의 기저 분류기를 발생시킬 수 있다는 장점이 있으며 이들 다양한 기저 분류기들 중에서 최적의 조합을 선택하여 앙상블을 구성한다는 장점이 있다. 즉, 생성된 모든 기저 분류기를 결합하는 것이 아니고, 결합하였을 경우 앙상블 모형의 성과 측면에서 가장 좋은 성과를 낼 수 있는 기저 분류기들의 조합을 유전자 알고리즘을 이용하여 탐색하고 이를 통해 최종 모형을 구성한다는 점에서 기존의 단순한 통합 앙상블 모형보다 성과 측면에서 장점이 있다고 볼 수 있다. <Table 7>에서 보는 바와 같이 GARSBagging 모형을 구성하고 있는 기저 분류들의 평균 예측률 값이 가장 좋지는 않지만 Q-통계량의 값은 가장 작은 것을 알 수 있다. 이는 유전자 알고리즘이 여러 기저 분류기들 중에서 평균 예측률은 유지하면서 서로 다른 패턴을 보이는 기저 분류기들의 조합을 잘 선택하였으며 이를 통해 제안한 모형의 성과가 개선되었다는 것을 유추해 볼 수 있을 것이다.

각 모형간의 성과 차이에 대한 통계적 유의성을 검토하기 위해 t검정을 이용하였으며 <Table 8>은 검정 결과를 보여주고 있다. <Table 8>에서 보는 바와 같이 본 연구에서 제안한 모형이 기존의 단일 모형, 단일 랜덤 서브스페이스 모형, 배깅 모형뿐만 아니라 기존의 통합 모형 보다 통계적으로 유의한 차이가 있는 것으로 나왔다. 즉, 본 연구에서 제안한 모형이 기존의 모형보다 효과적임을 알 수 있었다.

〈Table 6〉 Coincident Errors between Classifier C_i and C_j

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

〈Table 7〉 Ensemble Model Performance

	EOC(%)	BC(%)	Q
RS	73.33	66.78	0.55
Bagging	73.95	68.79	0.43
RS-Bagging(1)	74.62	66.43	0.54
RS-Bagging(2)	74.92	65.02	0.44
GARSBagging	77.11	65.48	0.39

〈Table 8〉 t-test(p-value)

	RS	Bagging	RS-Bagging(1)	RS-Bagging(2)	GARSBagging
DT	0.000	0.000	0.000	0.000	0.000
RS		0.424	0.122	0.061	0.000
BAG			0.541	0.345	0.000
RS-Bagging(1)				0.845	0.001
RS-Bagging(2)					0.019

6. 결 론

앙상블 학습이란 동일한 문제를 풀기 위해 다수의 분류기를 활용하는 기법으로, 일반적으로 하나의 분류기를 사용할 때보다 좋은 성과를 낼 수 있는 것으로 알려져 있으며 이로 인해 최근에 데이터 마이닝 분야에서 많은 관심을 끌고 있다.

앙상블 분류기의 좋은 예측 성과로 부도 예측 문제에도 이를 적용하기 위한 다양한 연구가 진행되고 있지만 아직까지 여러 앙상블 기법을 통합하는 연구는 많지 않은 것이 현실이다. 또한, 기존의 앙상블 통합 모형도 단순히 두 개의 모형을 결합하는 모형이 대부분이다. 이에 본 논문에서는 기존의 앙상블 모형의 성능 개선을 위해 새로운 형태의 통합 앙상블 최적화 모형을 제안하였다. 제안한 모형은 기존의 단순한 앙상블 결합 모형이 아닌 유전자 알고리즘을 이용한 다양한 앙상블 모형의 최적 결합 모형이다.

본 연구에서 제안한 모형의 우수성을 실제 기업 데이터를 가지고 실험을 수행하였으며, 실험 결과 제안한 앙상블 모형이 기존의 단일 모형, 단일 앙상블 모형뿐만 아니라 다른 통합 모형보다 우수한 성과를 보임을 알 수 있었다.

본 연구에서 제안한 모형은 기존의 앙상블 모형의 성능을 개선하는 데 효과적이었으며, 이는 부도 예측 문제뿐만 아니라 다양한 분류 문제에도 적용될 수 있을 것으로 기대된다. 본 논문에서는 각 모형의 성과 비교를 위해 예측 정확도만을 사용하였으나 향후 ROC 커브, AUC와 같은 다양한 성과 지표도 함께 비교해 볼 필요가 있을 것이다. 또한, 향후 보다 다양한 데이터를 가지고 제안한 모형이 다른 분류 문제에도 효과적인지에 대한 추가적인 연구가 필요할 것으로 보인다. 또한, 본 논문에서는 배깅과 랜덤 서브스페이스의 최적 통합 방안에 대해 살펴보았지만, 향후 다른 앙상블 기법들의 통합에 대한 연구가 필요할 것으로 생각된다.

References

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

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

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

[4] Buta, P., “Mining for financial knowledge with CBR”, *AI Expert*, Vol. 9, No. 10, 1994, pp. 34-41.

[5] Choi, H. N. and Lim, D. H., “Bankruptcy prediction using ensemble SVM model”, *Journal of the Korean Data and Information Sciences Society*, Vol. 24, No. 6, 2013, pp. 1113-1125.

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

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

[8] Hansen, L. and Salamon, P., “Neural network ensembles”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 12, No. 10, 1990, pp. 993-1001.

[9] Ho, T., “The random subspace method for construction decision forests”, *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence, 1998, pp. 832-844.

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

[11] 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,* Vol. 42, No. 3, 2015, pp. 1074-1082.

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

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

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

[15] Marques, A. I., Garcia, V., and Sanchez, J. S., "Two-Level Classifier Ensembles for Credit Risk Assessment", *Expert Systems with Applications*, Vol. 39, No. 12, 2012, pp. 10916-10922.

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

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

[18] Min, S., "Developing an Ensemble Classifier for Bankruptcy Prediction", *Journal of the Korea Industrial Information Systems Research*, Vol. 17, No. 7, 2012, pp. 139-148.

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

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

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

[22] 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*, Vol. 116, 1999, pp. 16-32.

저자소개

민 성 환

KAIST 테크노 경영대학원에서 경영정보시스템을 전공하여 박사를 취득하였으며 현재 한림대학교 경영학과에 재직 중이다. 주요 관심분야는 데이터 마이닝, 재무예측 모형 개발, 고객관계관리 등이다.