ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 December: 20(4): 121~139 http://dx.doi.org/10.13088/jiis.2014.20.4.121

개선된 배깅 앙상블을 활용한 기업부도예측

민성환
한림대학교 경영학부
(E-mail : shmin@hallym.ac.kr)

기업의 부도 예측은 재무 및 회계 분야에서 매우 중요한 연구 주제이다. 기업의 부도로 인해 발생하는 비용이 매우 크기 때문에 부도 예측의 정확성은 금융기관으로서는 매우 중요한 일이다. 최근에는 여러 개의 모형을 결합하는 앙상블 모형을 부도 예측에 적용해 보려는 연구가 큰 관심을 끌고 있다. 앙상블 모형은 개별 모형보다 더 좋은 성과를 내기 위해 여러 개의 분류기를 결합하는 것이다. 이와 같은 앙상블 분류기는 분류기의 일반화 성능을 개선하는 데 매우 유용한 것으로 알려져 있다.

본 논문은 부도 예측 모형의 성과 개선에 관한 연구이다. 이를 위해 사례 선택(Instance Selection)을 활용한 배깅(Bagging) 모형을 제안하였다. 사례 선택은 원 데이터에서 가장 대표성 있고 관련성 높은 데이터를 선택하고 예측 모형에 악영향을 줄 수 있는 불필요한 데이터를 제거하는 것으로 이를 통해 예측 성과 개선도 기대할 수 있다. 배깅은 학습데이터에 변화를 줌으로써 기저 분류기들을 다양화시키는 앙상블 기법으로 단순하면서도 성과가 매우 좋은 것으로 알려져 있다. 사례 선택과 배깅은 각각 모형의 성과를 개선시킬 수 있는 잠재력이 있지만 이들 두 기법의 결합에 관한 연구는 아직까지 없는 것이 현실이다.

본 연구에서는 부도 예측 모형의 성과를 개선하기 위해 사례 선택과 배깅을 연결하는 새로운 모형을 제안하였다. 최적의 사례 선택을 위해 유전자 알고리즘이 사용되었으며, 이를 통해 최적의 사례 선택 조합을 찾고 이 결과를 배깅 앙상블 모형에 전달하여 새로운 형태의 배깅 앙상블 모형을 구성하게 된다. 본 연구에서 제안한 새로운 앙상블 모형의 성과를 검증하기 위해 ROC 커브, AUC, 예측정확도 등과 같은 성과지표를 사용해 다양한 모형과 비교 분석해 보았다. 실제 기업데이터를 사용해 실험한 결과 본 논문에서 제안한 새로운 형태의 모형이 가장 좋은 성과를 보임을 알 수 있었다.

주제어 : 배깅, 사례선택, 앙상블, 부도예측, 유전자 알고리즘

논문접수일 : 2014년 11월 12일 논문수정일 : 2014년 12월 17일 게재확정일 : 2014년 12월 18일
투고유형 : 국문급행 교신저자 : 민성환

1. 서론

기업의 부도 예측은 재무 및 회계 분야에서 매우 중요한 연구 주제이다. 기업의 부도로 인해 발생하는 비용이 매우 크기 때문에 부도 예측의 정확성은 금융기관으로서는 매우 중요한 일이다. 기업의 부도 예측 모형에 관한 초기의 연구는 주로 전통적인 통계적 모형에 기반을 둔 모형이 대부분이었다. 단일변량분석(univariate analysis) (Beaver, 1966), 다변량 판별분석(multiple discriminant analysis) (Altman, 1968), 다중회귀분석(multiple regression analysis) (Meyer and Pifer, 1970), 로지스틱 회귀분석(logistic regression) (Dimitras et al., 1996; Ohlson, 1980) 등의 통계 모형을 부도 예측

* 이 논문은 2013년도 한림대학교 교비학술연구비(HRF-201305-006)에 의해 연구되었음

에 적용해 보는 연구가 여기에 속한다. 그 뒤로 인공지능 기법을 부도 예측 문제에 적용해 보는 다양한 연구가 시작되었다. 사례기반추론(Case-based reasoning) (Buta, 1994; Bryant, 1997), 귀납적 학습방법(inductive learning) (Messier and Hansen, 1998; Shaw and Gentry, 1998), 인공신경망(Artificial neural networks) (Tam and Kiang, 1992; Zhang et al., 1999)과 같은 다양한 모형들을 부도 예측 문제에 적용한 연구가 여기에 속한다. 또한, 유전자 알고리즘(Genetic Algorithms)을 활용한 최적화 기법을 부도예측 문제에 적용하는 다양한 연구도 활발하게 진행 되고 있다. 부도 예측을 위한 최적의 입력변수 선정 문제(Hong and Shin, 2003), 최적의 입력 데이터 정규화 문제(Tai and Shin, 2010), 최적의 사례 선택 문제 (Kim, 2004), 최적의 분류기 선택 문제(Kim, 2010)에 유전자 알고리즘을 활용한 연구들이 여기에 속한다.

한편, 최근에는 여러 개의 모형을 결합하는 앙상블 모형을 부도 예측에 적용해 보려는 연구가 큰 관심을 끌고 있다. 앙상블 모형은 개별 모형보다 더 좋은 성과를 내기 위해 여러 개의 분류기를 결합하는 것이다. 이와 같은 앙상블 모형은 분류기의 일반화 성능을 개선하는 데 매우 유용한 것으로 알려져 있다. 그러나, 앙상블 모형의 성능을 향상시키기 위해서는 각각의 기저 분류기들의 성과가 좋아야 할 뿐만 아니라 기저 분류기들 간에 가능하면 다양성(diversity)을 가지고 있어야 한다. 앙상블 모형에서 분류기 사이의 다양성은 앙상블의 최종 성과에 많은 영향을 미치는 매우 중요한 요소이다. 만약에 분류기들 사이에 다양성이 존재하지 않는다면 앙상블 분류기의 일반화 성능의 개선은 기대할 수 없을 것이다. 극단적인 예로 만약 앙상블을 구성하고 있는 기저 분류기들이 모두 똑같다면, 이들의 결합으로 어떠한 성과개선도 기대할 수 없을 것이다. 그러므로, 앙상블 분류기를 통한 성과 개선을 기대하기 위해서는 기저 분류기들을 다양화시키는(diversify) 것이 필요하다. 앙상블 기법 중에서 배깅(bagging)과 부스팅(boosting)은 학습데이터에 변화를 줌으로써 기저 분류기들을 다양화시키는 앙상블 기법으로 가장 많이 사용되고 있으며 성과가 매우 좋은 것으로 알려져 있다.

최근에 앙상블 기법은 부도 예측 문제에 성공적으로 적용되고 있다. (Kim and Kim, 2007)은 여러 개의 후보 기저 분류기 중에서 평균 이상의 성과를 보인 기저 분류기만을 선택하여 결합하는 변형된 배깅 모형을 제안하였으며, 단일 모형보다 제안한 배깅 모형이 우수한 성과를 보였다. (Min, 2012)은 부도 예측문제에 배깅과 random subspace 기법을 각각 적용해 보았다. 또한 성과개선을 위해 배깅과 random subspace의 통합 모형을 제안하고 이를 부도 예측에 적용해 보았다. (Kim, 2009)은 기업의 부도 예측에 배깅과 부스팅을 적용해 보았다. 실험결과 의사결정 트리, 인공신경망을 기저 분류기로 했을 때의 단일 모형 보다 앙상블 모형이 성과가 좋음을 알 수 있었다. (Ok and Kim, 2009)은 로지스틱 회귀분석, 의사결정트리, 인공신경망, 사례기반추론의 최적 결합을 위해 유전자 알고리즘을 사용하였으며, 기업의 부도 예측 문제에 적용한 결과 제안한 모형이 기존의 단일 모형, 단순 결합 방법과 비교해 좋은 성과를 보였다. (Shin and Hong, 2011)은 SVM을 기저 분류기로 하는 AdaBoost 모형을 기업신용평가 문제에 적용해 보았으며, 실험 결과 제안한 모형이 오분류 문제를 줄 일 수 있음을 보였다. (Kim, 2012)은 기하평균개념을 부스팅 알고리즘에 적용한 새로운 형태의

GM-Boost알고리즘을 확장하여 다범주 문제인 회사채 등급평가 문제에 적용해 보았다.

본 논문에서는 기존의 앙상블 부도 예측 모형의 성과 개선을 위해 사례 선택(Instance selection)을 활용하여 배깅의 성능을 개선시키는 새로운 모형을 제안한다. 지금까지 많은 앙상블 모형 관련 연구가 있었지만 아직까지 사례 선택과 배깅을 동시에 고려하는 연구는 거의 없는 것이 현실이다. 사례 선택은 데이터 마이닝 분야에서 매우 효과적인 기법 중의 하나로 주어진 데이터에서 불필요한 데이터, 노이즈(noise) 데이터, 관련 없는 데이터 등을 제거하고 가장 대표성 있고 관련성 높은 핵심적인 사례(instances)를 선택하는 것을 말한다. 사례 선택을 적절하게 수행하였을 경우 불필요한 사례를 제거함으로써 데이터 사이즈를 줄일 수 있으며, 선택된 핵심적인 사례를 사용할 경우 원 데이터 모두를 사용하는 것과 비교해 비슷하거나 심지어는 더 좋은 성과를 기대할 수 있다 (García et al., 2012). 사례 선택과 배깅은 각각 모형의 성과를 개선시킬 수 있는 잠재력이 있지만 이들 두 기법의 결합에 대한 연구는 아직까지 없는 것이 현실이다.

본 논문에서 제안한 모형에서는 배깅의 입력 데이터로 사용될 최적의 사례를 선택하기 위해 유전자 알고리즘을 사용하였다. 본 논문에서 제안한 새로운 앙상블 모형의 성과를 검증하기 위해 기존의 단일 모형, 사례 선택을 활용한 모형, 단순 배깅 모형을 비교 모델로 사용하였으며, 실제 기업데이터를 이용해 다양한 성과지표를 가지고 비교 분석하였다.

본 논문은 다음과 같이 구성되었다. 먼저 2장에서 이론적 배경이 설명되고, 3장에서는 본 논문에서 제안한 모형에 대한 설명을 하였다. 4장과 5장에서는 실험 설계와 실험 결과에 대해 서술하였고 마지막 장에서는 요약 및 향후 연구과제에 대해 설명하였다.

2. 이론적 배경

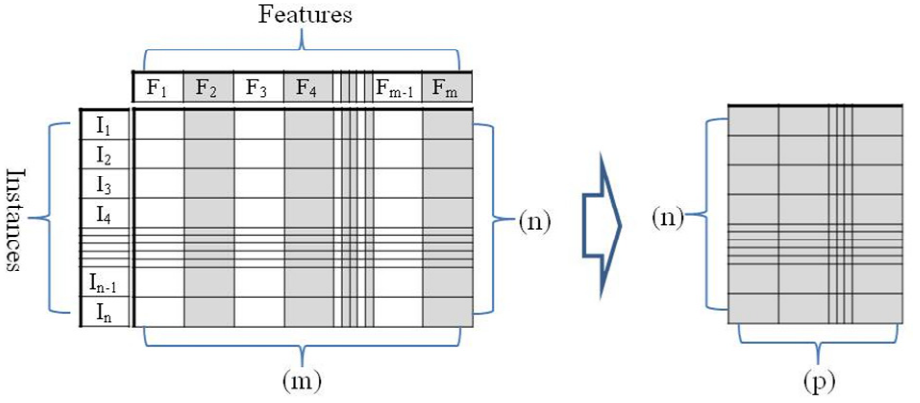
2.1 사례 선택(Instance Selection)

데이터 축소(data reduction) 기법은 데이터 전처리 기법 중의 하나로 대량의 데이터를 보다 더 쉽게 다룰 수 있도록 가장 대표성 있고 핵심적인 데이터를 선택함으로써 원 데이터를 감소시키는 것이다. 이와 같은 방법으로 데이터 저장 공간을 줄일 수 있으며 문제의 복잡성과 연산시간도 줄일 수 있고, 불필요하거나 관련 없는 데이터를 제거함으로써 모형의 성과 개선도 기대할 수 있다. 데이터 마이닝 분야에서 많이 활용되는 대표적인 데이터 축소 방법으로는 특징변수 선택(feature selection)과 사례 선택이 있다. 특징변수 선택은 데이터 셋(data set)에서 열(column)의 수를 줄이는 것이고 사례 선택은 데이터 셋에서 행(row)의 수를 줄이는 것이다. <Figure 1>은 사례 선택과 특징변수 선택을 예를 통해 비교해 주고 있다. 모형에 사용할 입력 데이터의 사례의 수가 n개이고 특징변수의 수가 m개라고 한다면 입력 데이터는 행렬로 나타낼 수 있다. 이때 특징변수 선택 문제는 m개의 열에서 p개의 열을 선택하는 문제라고 볼 수 있다 (단, $p \leq m$). 그리고, 사례 선택 문제는 n개의 행에서 k개의 행을 선택하는 문제라고 볼 수 있다 (단, $k \leq n$).

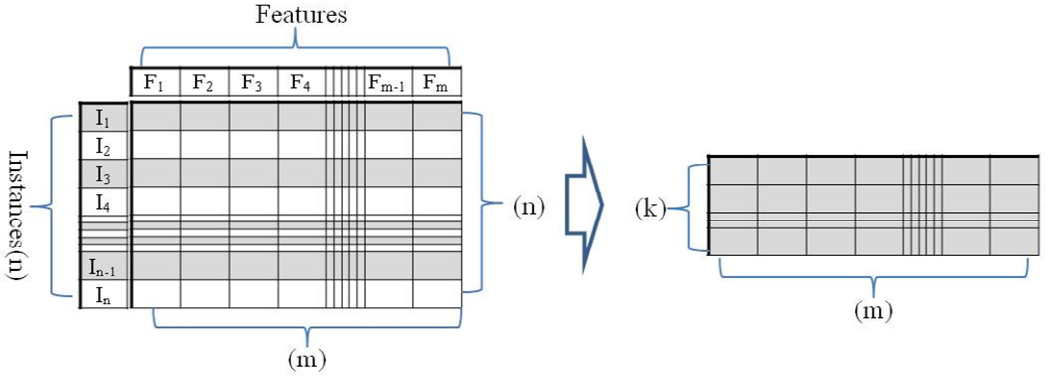
사례 선택은 데이터 마이닝 분야에서 매우 효과적인 기법 중의 하나로 원 데이터에서 불필요한 데이터, 관련 없는 데이터 또는 모형 개발에 오히려 해를 끼치는 노이즈와 같은 데이터를 제

거하고 모형 개발을 위해 핵심적이고 중요한 사례를 선택하는 것을 말한다. 사례 선택의 주목적은 원 데이터에서 가장 대표성 있고 관련성 높은 데이터를 선택하는 것이다. 사례 선택을 적절하게 수행하였을 경우 불필요한 사례를 제거함으로써 데이터 사이즈를 줄일 수 있으며, 선택된 핵심적인 사례를 사용할 경우 원 데이터 모두를 사용하는 것과 비교해 비슷하거나 심지어는 더 좋은 성과를 기대할 수 있다. (Kim, 2004; Ahn et al., 2005 ; Kim and Ahn, 2011)의 연구에 의하면 사례 선택을 적절하게 수행할 경우 저장 공간을 절약할 수 있고, 자료 처리 속도를 높일 수 있을 뿐만 아니라 예측 모형의 성과가 개선될 수 있다는 것을 알 수 있다.

최초의 사례 선택 기법 중의 하나는 Hart에 의한 Condensed Nearest Neighbor Rule(CNN) 이다 (Hart, 1968). 이후에 사례 선택에 관한 많은 연구가 있었고, 다양한 기법들이 제안되었다. 또한, 사례 선택에 대한 분류도 다양하게 제안되어 왔는데 (Derrac et al., 2012)는 사례 선택을 특징변수 선택에서 사용하는 것과 유사하게 사용하는 전략에 따라 크게 Wrapper 기법과 Filter기 법으



(a) Feature Selection: I(n,m) → I'(n,p) ($m \geq p$)



(b) Instance Selection: I(n,m) → I'(k,m) ($n \geq k$)

〈Figure 1〉 (a) Feature Selection (b) Instance Selection

로 분류하였다. Wrapper 접근방법은 분류기에 의해 얻어진 예측 결과에 기반을 둔 사례 선택을 수행하는 접근 방법을 의미하며 Filter 접근방법은 분류기와 관계없이 사례 선택이 수행되는 것을 의미한다. 이와는 달리 사용하는 모델에 따라 사례 선택은 크게 두 그룹으로 구분할 수 있다. Prototype Selection (PS) methods와 Training Set Selection(TSS) methods가 이에 속한다. PS 기법은 knn(k-nearest neighbor)과 같은 게으른 학습(lazy learning) 모형에 적용되는 기법이고, TSS는 인공신경망, 의사결정 트리 (decision tree)와 같은 일반적인 예측 모형에 적용되는 기법을 의미한다. 즉, PS 기법은 knn과 같이 prototype에 기초를 둔 알고리즘에 적용되는 사례 선택 기법을 의미하고, TSS 기법은 학습데이터를 사용하는 일반적인 학습 알고리즘에 적용되는 기법을 의미한다.

지금까지 사례 선택에 관한 많은 연구가 있어 왔으나, knn과 같은 게으른 학습 모형을 기반으로 하는 사례 선택에 관한 연구가 대부분이었다. 또한 사례 선택과 배깅의 결합에 관한 연구는 거의 없는 것이 현실이다. 본 연구는 SVM을 기저 분류기로 사용하는 앙상블의 성과개선에 관한 연구이다. 본 연구에서는 SVM 앙상블 모형의 성과 개선을 위해 사례 선택 기법과 배깅을 연결하는 새로운 모형을 제안하였다.

2.2 앙상블 모형

앙상블 모형은 최근 데이터 마이닝, 기계학습 분야에서 각광 받는 분야로 개별 분류기보다 더 좋은 성과를 내기 위해 여러 기저 분류기들의 결과를 결합하는 모형을 의미한다. 앙상블 모형은 개별적으로 학습된 일련의 분류기로 구성되며 각각의 분류기의 예측 결과는 다양한 방법을 통해 결합되어 최종 앙상블의 예측 결과가 나온다.

대부분의 경우에 앙상블 모형은 그것을 구성하고 있는 기저 분류기들보다 더 좋은 예측률을 보이는 것으로 알려져 있다 (Dietterich, 1997). 또한, 단순한 선형 분류기(simple linear classifier)도 앙상블 모형을 통해 결합이 되면 complex decision boundary를 만들어 낼 수 있으며 앙상블 모형은 단일 분류기보다 더 robust 하다고 알려져 있다 (Kuncheva , 2004).

앙상블 접근방법의 성과는 기저 분류기들의 정확성(accuracy), 다양성(diversity)과 관련이 있는 것으로 알려져 있다. 앙상블 모형이 기저 분류기들 보다 더 좋은 성과를 내려면, 앙상블 모형을 구성하고 있는 기저 분류기들의 성과가 가능하면 좋아야 하며, 이들 개별 기저 분류기들이 가능한 한 다양성을 갖는 것이 필요하다 (Bian and Wang , 2007; Kuncheva and Whitaker , 2003). 만약 앙상블을 구성하고 있는 개별 기저 분류기들의 성과가 너무 좋지 않다면 기저 분류기들의 예측 결과 정보에 기초를 둔 앙상블의 성과에도 좋지 않은 영향을 미칠 것이다. 또한 다양성이 없는 개별 분류기들의 결합은 모형의 복잡성만 증가할 뿐 성과 개선은 크지 않다. 만약에 앙상블을 구성하고 있는 기저 분류기들이 모두 완전하게 동일하다면 앙상블의 성과는 개별 분류기의 성과보다 더 좋아질 수 없을 것이다. 반대로 기저 분류기들이 서로 다르다면 좋은 결합 방법을 통해 앙상블의 성과 개선을 기대해 볼 수 있을 것이다. 여기서 기저 분류기들이 서로 다르다는 것은 다양성을 의미한다. 다양성과 정확성을 통해 소수의 개별 분류기가 잘못 예측하더라도 나머지 다수의 개별분류기가 정확하게 예측하면 앙상블 분류기는 정확하게 예측할 수 있을 것이

며 이와 같은 기저 분류기의 시너지 효과를 통해 앙상블 분류기는 단일 기저 분류기보다 좋은 성과를 기대할 수 있을 것이다. 그러므로, 앙상블 분류기를 통한 성과 개선을 기대하기 위해서는 기저 분류기들을 다양화시키는 것이 필요하다. 앙상블 모형은 다양한 분류기를 생성시키고 이들을 적절한 방법으로 결합함으로써 단일 모형보다 우수한 성과를 내는 것을 그 목표로 하고 있다. 분류기들을 다양화시키는 방법은 여러 가지 형태가 있을 수 있으며 대표적인 것들은 다음과 같다.

- 학습 데이터의 다양화: 각각의 기저 분류기들을 서로 다른 학습데이터를 사용하여 학습시킨다. 대표적인 방법으로는 배깅과 부스팅이 있다. 샘플링 방법을 이용해 서로 다른 학습데이터를 생성시키고 이를 이용해 다양한 기저 분류기를 생성한다.
- 분류기의 다양화: 학습데이터에 변화를 주기보다는 학습 알고리즘에 변화를 줌으로써 기저분류기를 다양화 시키는 방법이다. 동일한 알고리즘에 파라미터 값의 변화를 줌으로써 다양화를 시킬 수 있으며, 서로 다른 알고리즘을 사용해 기저 분류기들을 다양화 시킬 수 있다.
- 기타 방법: 위의 방법들 중 여러 개를 동시에 사용하는 경우 또는 새로운 기법과 위의 방법과 결합하는 방법이 여기에 속한다. 예를 들면, 학습데이터와 학습 알고리즘 파라미터를 동시에 변화를 줌으로써 기저 분류기를 다양화시키는 방법이 이에 속한다.

이 중에서 학습데이터에 변화를 주는 방법 중의 하나인 배깅과 부스팅 방법은 가장 많이 사용되는 앙상블 구성 기법이다. 이들은 학습데이터에 변화를 줌으로써 기저분류기의 다양성을 확보하는 접근방법이라고 볼 수 있다. 배깅은 bootstrap aggregating의 약자로 원 학습데이터(original training dataset)로부터 복원추출 방법에 의해 랜덤 샘플링을 함으로써 학습데이터의 부분집합(bootstrap)을 발생시킨다 (Breiman, 1996). 이와 같이 발생시킨 서로 다른 여러 개의 학습데이터를 이용해 각각의 기저 분류기를 학습시키고 이를 통해 다양성이 존재하는 기저 분류기를 생성시킬 수 있다. 부스팅 방법은 학습데이터에 변화를 줌으로써 기저 분류기를 다양화 시킨다는 측면에서는 배깅과 유사한 방법이나 배깅과 부스팅 방식의 가장 큰 차이는 부스팅은 이전 단계의 분류기의 예측 정확도 결과값에 의해 적응적으로(adaptively) 학습데이터에 변화를 준다는 것이다. 즉, 부스팅은 이전 단계에서 오분류된 데이터의 가중치를 증가시키고, 정확하게 분류된 데이터의 가중치는 감소시킴으로써 오분류 데이터에 더 집중하는 프로세스를 가진다는 것이 배깅과의 큰 차이점이다.

이와 같이 배깅이나 부스팅은 다양한 학습데이터를 생성시키고, 서로 다른 각각의 학습데이터를 기저 분류기 학습에 이용하게 된다. 이렇게 기저 분류기의 학습이 완료되면 이들 기저 분류기들의 예측 결과의 결과값(output)을 결합해야 한다. 본 연구에서는 앙상블 모형의 성과 개선을 위한 최적 사례 선택 및 최적 분류기 선택 쪽에 초점을 두었기 때문에 결합 방법은 가장 많이 사용되고 있는 방법 중의 하나인 다수결 투표(majority vote) 방법을 사용하여 실험하였다.

3. 연구 모형

본 논문은 부도 예측 모형의 성과 개선에 관한

연구이다. 이를 위해 사례 선택과 배깅을 연결하는 새로운 모형을 제안하였다. 사례 선택은 원 데이터에서 가장 대표성 있고 관련성 높은 데이터를 선택하는 것으로 이를 통해 입력 데이터의 크기를 줄일 수 있으며, 저장 공간의 절약, 데이터 처리 속도의 향상과 같은 효과를 기대할 수 있다. 뿐만 아니라, 예측 모형에 악영향을 줄 수 있는 불필요한 데이터, 노이즈 등의 제거로 인해 예측 성과 개선도 기대할 수 있다. 한편, 배깅을 통한 앙상블 모형은 단순하면서도 가장 많이 이용되는 기법 중의 하나로 이를 통해 단일 모형보다 개선된 성과를 기대할 수 있으며, 많은 과거 연구에 의하면 배깅을 통해 단일 모형보다 모형의 성과가 개선됨을 알 수 있다.

사례 선택과 배깅은 데이터 마이닝 분야에서 잘 알려져 있고 연구가 활발하게 진행되고 있는 분야이며 각각의 기법은 모형의 성과 개선에 기여할 수 있는 잠재력을 가지고 있다. 하지만, 이 두 가지 기법의 결합에 관한 연구는 거의 없는 것이 현실이다. 본 연구에서는 단일 모형의 예측 성과 개선에 매우 유용한 사례 선택과 배깅을 동시에 고려하는 새로운 모형에 대해 고찰해 보았다.

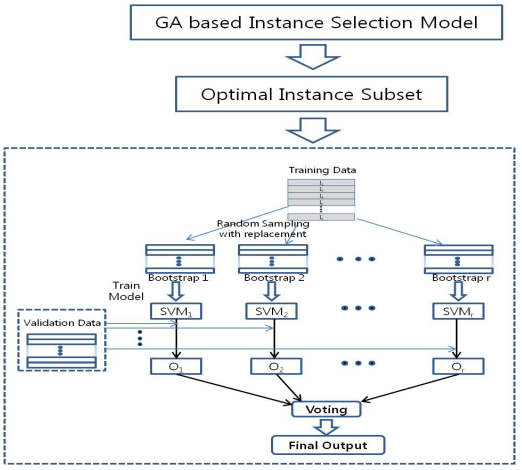
본 논문에서 제안한 앙상블 모형의 기저 분류기로는 SVM(Support Vector Machines)을 사용하였다. SVM은 (Vapnik, 1995)에 의해 소개된 이후 뛰어난 일반화 성능으로 인해 데이터 마이닝 분야에서 큰 관심을 끌고 있다. 인공신경망과 같은 모형은 경험적 위험을 최소화(empirical risk minimization)하는 원칙에 따른 모형인 반면 SVM은 구조적 위험을 최소화(structural risk minimization)하는 이론으로부터 개발되었다. SVM은 많은 응용분야에서 성공적으로 적용되어 왔다.

본 연구에서는 SVM 모형을 이용한 부도 예측 모형의 성과를 개선하기 위해 사례 선택과 배깅을 연결하는 새로운 방법을 제안하였다. 최적의 사례 선택을 위해 유전자 알고리즘이 사용되었으며, 이를 통해 최적의 사례 선택 조합을 찾고 이 결과를 배깅 앙상블 모형에 전달하여 새로운 형태의 배깅 앙상블 모형을 구성하게 된다.

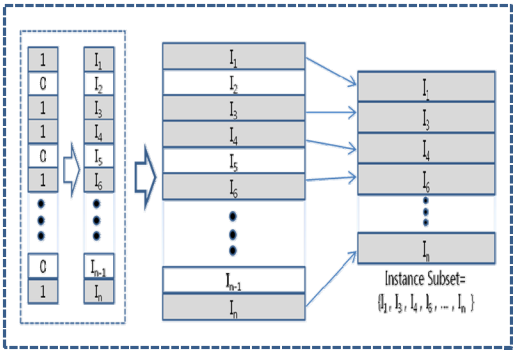
일반적인 사례 선택의 목적은 예측 정확도를 유지하면서 (또는 예측 정확도 향상을 기대하면서) 데이터의 크기를 줄이는 것이 목적이다. 데이터 사이즈를 줄임으로써 얻는 기대효과로 저장공간 절약, 데이터 처리 속도 개선 등을 들 수 있으며, knn과 같은 게으른 학습 모형의 경우 데이터 공간 절약은 매우 중요한 문제 중의 하나이다. 이와 같은 목적으로 설계된 사례 선택 기법은 목적함수로 예측정확도 유지나 개선뿐만 아니라 데이터크기 감소폭도 중요한 요소이다. 하지만 본 연구의 목적은 모형의 예측 정확도 개선이므로 데이터 감소 비율은 상대적으로 중요성이 떨어진다고 볼 수 있다. 그러므로 본 연구에서 사용한 유전자 알고리즘의 적합도 함수는 예측정확도 만을 포함시켰다. 본 연구의 궁극적인 목적은 사례 선택을 통해 예측 모형에 악영향을 주는 사례를 제거하고, 이를 통해 배깅 앙상블을 위한 기저 분류기의 성과 개선 및 앙상블의 성과를 개선하고자 하는 데 있다.

앞에서 살펴본 바와 같이 (Derrac et al., 2012)에 의하면 사례 선택은 wrapper와 filter의 두 가지 접근 방식이 있다. 이들 방법 중에 본 논문에서는 분류기에 의해 얻어진 예측 결과에 기반을 둔 wrapper 접근방법을 사용하였다. 본 논문에서는 유전자 알고리즘을 이용하여 SVM의 최적의 사례집합(optimal instance subset)을 찾고 이 결과는 배깅 앙상블의 성능 개선을 위해 사용되었다.

본 연구는 유전자 알고리즘을 기반으로 하는 사례 선택이 최종 목표가 아니고, 이 결과를 배깅 앙상블 모형의 성능 개선을 위해 사용하는 새로운 형태의 모형 개발이 주목적이다. 즉, 유전자 알고리즘을 이용해 최적(또는 근사 최적)의 사례가 선택되고, 이 결과가 배깅 모형의 입력 데이터로 들어가게 된다.



〈Figure 2〉 The overall architecture of the proposed model.



〈Figure 3〉 Encoding for GA

<Figure 2>는 본 논문에서 제안한 모형의 전반적인 절차(overall procedure)를 나타내고 있다. 본 논문에서 제안한 모형은 유전자 알고리즘을 이용한 사례 선택 모형과 배깅 모형으로 구성되어 있다. 유전자 알고리즘을 이용한 사례 선택 모형은 가장 먼저 랜덤하게 선택된 염색체(chromosomes)로부터 시작한다. 이 염색체는 SVM 모형의 입력 데이터를 위한 사례집합에 해당한다. 사례집합에 대한 염색체는 이진열(binary sting) 형태로 표현하였으며, 이는 원 학습데이터 셋 (original training set)의 부분집합인 사례 집합을 나타내게 된다. <Figure 3>은 최적의 사례 집합을 선택하기 위해 사용된 유전자 알고리즘의 염색체 구조를 나타내고 있다. <Figure 3>에서 I_1은 첫 번째 사례를 의미하고 I_2는 두 번째 사례를 의미한다. 그림에서 보는 바와 같이 원 학습데이터는 n개의 사례로 구성되어 있으며 유전자 알고리즘 염색체의 비트는 각각의 사례와 대응되도록 총 n 비트로 설계되었다. 각각의 비트는 해당되는 사례가 선택되었는지 아닌지를 나타내게 된다. 각각의 bit에서 1은 해당되는 사례가 선택되었다는 것을 뜻하고 0은 선택되지 않았다는 것을 뜻한다. 그림에서 I_1에 해당하는 염색체의 비트 값은 1로 이는 첫 번째 사례 I_1이 선택되었다는 것을 의미하고 I_2에 해당하는 염색체의 비트 값은 0으로 이는 두 번째 사례 I_2가 선택되지 않았다는 것을 의미한다. 이와 같은 방법으로 각각의 염색체는 선택된 각각의 사례집합을 의미하게 되며, 선택된 사례집합은 SVM 모형으로 보내져 적합도 함수에 의해 평가된다. SVM 모형은 성과 척도(performance measure)를 구하기 위해 입력변수로 선택된 사례집합을 사용하게 된다. 이 성과척도는 적합도 함수로 사용되고 유전자 알고리즘에 의해 진화하게 된다. 이와 같은 절차는 최적(또는 근사 최적)이 선택될 때까지 반복되게 된다. 본 논문에서는 적합도 함수로

예측 정확도를 사용하였으며 식 (1)과 같이 표현할 수 있다. 여기서 n은 과적합을 피하기 위해 사용한 데이터 셋인 테스트용 데이터(test data)의 크기를 의미하고, 테스트 용 데이터 중 i번째 데이터 (또는 사례)에서의 실제값과 예측값이 일치할 경우 H_i 값은 1의 값을 갖고, 그렇치 않을 경우 0의 값을 갖는다.

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

<Phase I>-GA based Instance Selection

1. Define the chromosome
   (The chromosome for the instance subset is encoded as a form of binary string)
2. Determine parameters of GA
3. Generate the initial population
4. Select the instance subset for each chromosome
5. Calculate the fitness values of different instance subsets
6. Perform GA operations and create a new generation
7. Repeat from step 4 to 6 until the termination criteria are satisfied.
8. Select the optimal instance subset (GAT)

<Phase II> Instance based Bagging

9. Use the optimal instance subset(GAT) as input data of bagging model
10. Generate a new training data set of size K' by randomly sampling with replacement
    (K'<K(size of GAT))
11. Repeat step10 to generate R new training data sets
    → $GAT(B)_1$, $GAT(B)_2$, … , $GAT(B)_R$
12. Train SMV model for each new training set
    (Different SVM models are generated) → SVM_1, … , SVM_R
13. Apply the validation data set to the SVM models generated in Step 12
    (R different output data) → O_1, … , O_R
14. Combine the output data (O_1, … , O_R) by a combining method
    (In this paper, we use majority voting scheme as a combining method)

〈Figure 4〉 Steps of Proposed Model

이와 같은 방법으로 유전자 알고리즘을 통해 최적의 사례집합이 선택되면, 이 결과는 배깅 앙상블 모형의 성능 개선을 위해 사용된다. 선택된 사례집합은 불필요한 데이터나 모형 개발에 해를 끼치는 노이즈와 같은 데이터를 제거한 핵심적인 사례로 구성된 최적의 사례집합으로 배깅 앙상블 모형에서는 기존의 학습 데이터 대신 이것을 입력데이터로 사용한다. 배깅 앙상블 모형에서는 선택된 사례로 구성된 입력데이터에서 랜덤 복원 추출방법을 통해 서로 다른 학습데이터를 생성하고, 이들 각각 학습데이터를 사용해 SVM을 학습하여 다양성을 갖는 여러 개의 SVM 모형을 생성하고 이들을 다양한 전략에 의해 결합하게 된다. <Figure 4>는 본 논문에서 제안한 모형의 전반적인 절차를 보여주고 있다.

4. 실험 설계

본 연구에서는 기존의 대표적인 앙상블 모형 중의 하나인 배깅 모형의 성과개선을 위해 배깅과 사례 선택기법을 연결하는 새로운 형태의 앙상블 모형을 제안하였다. 본 연구에서는 앙상블 모형을 위한 기저 분류기로 최근에 우수한 성과로 각광받고 있는 SVM을 사용하였다. SVM의 커널(kernel) 함수로는 가장 많이 사용되고 있는 linear 커널과 rbf 커널을 사용하여 실험하였으며 제안한 모형의 우수성을 검증하기 위해 단일 모형, 일반 배깅 모형, 사례선택 기법을 활용한 모형을 비교 모델로 사용하였다.

본 연구에서 사용한 데이터는 자산규모가 10억에서 70억 사이인 국내 비외감 기업의 데이터로 총 1832개로 구성되어 있다. 이 중 부도 기업의 데이터와 비부도 기업의 데이터는 같

은 수인 916개로 이루어져 있다. 데이터는 학습용 데이터(training data), 테스트용 데이터(test data), 그리고 검증용 데이터(validation data)로 나누어 실험을 하였다. 학습용 데이터는 모형의 학습을 위한 데이터로 사용되었으며, 테스트용 데이터는 유전자 알고리즘을 이용한 최적의 사례를 선택할 때 과적합(overfitting)을 피하기 위한 용도로 사용되었다. 검증용 데이터는 모형의 비교 검증을 위해 사용하였다.

기업의 부도 여부를 예측하기 위한 입력변수로는 재무비율을 사용하였다. 입력변수 선택을 위해 총 131개의 재무비율을 대상으로 1차적으로 단일표본 t검정(Independent-samples t-test)을 실시하였다. 이를 통해 p-value 값이 0.05보다 큰 변수는 제외하고, 나머지 변수를 대상으로 stepwise method를 이용한 로지스틱 회귀분석(Logistic Regression) 과 선행연구 결과 등을 종합적으로 고려해 최종변수를 선정하였다. <Table 1>은 최종 선정된 변수에 대한 설명과 동일 변수를 부도예측 모형에 사용한 선행연구를 보여주고 있다.

SVM의 성과에 중요한 영향을 미치는 파라미터 C와 의 경우 예비 실험을 통해 가장 성과가 좋은 값(Liner 커널: C=1; RBF 커널: C=1, =25)을 대표값으로 선택하여 이후의 실험에서 모두 같은 값을 가지고 실험하였다. 최적의 사례 선택을 위해 사용된 유전자 알고리즘의 실험에서 모집단(population)의 크기는 100으로 하고, 정지조건은 150세대로 설정하였다. 또한, 교배율(crossover rate)과 돌연변이 비율(mutation rate)은 각각 0.7과 0.1로 설정하여 실험하였다. 한편, 배깅을 이용한 앙상블 모형의 성과는 앙상블을 구성하는 기저 분류기의 총 수와 bootstrap의 크기에 따라 차이가 난다. 본 연구에서 배깅 실험에 사용한 기저 분류기의 총 수는 25로 고정하여 실험을 하였으며, bootstrap 크기는 예비실험을 통해 가장 성과가 좋은 값을 대표 값으로 사용하여 실험하였다.

〈Table 1〉 Input Variables

Variable	Description	Reference
X1	Sales coefficient of variation	Shin and Hong, 2011; Min, 2012
X2	Financial expenses/sales	Kim,2010; Kim,2012; Min, 2012
X3	(Capital surplus + retained earnings-dividend)/total assets	Kim et al.,2007; Min, 2012
X4	(Cash + cash equivalents)/ current liabilities	Shin and Hong,2011; Kim,2010; Kim,2012; Min, 2012
X5	Working capital change/sales	Min, 2012
X6	EBITDA/interest expenses	Kim,2010; Kim,2012
X7	Trade payable turnover period	Ok and Kim,2009; Min, 2012
X8	Borrowings/total assets	Ok and Kim,2009; Min, 2012

5. 실험 결과

본 논문에서 제안한 모형의 우수성을 검증하기 위해 기존의 모형과 다양한 비교를 하였으며, 비교를 위해 사용한 성과 지표로는 예측정확도, ROC 커브, AUC가 있다. 일반적으로 이진 분류 문제에서 가능한 두 개의 결과값을 positive class와 negative class라고 할 경우 분류기 예측을 통

해 가능한 모든 경우는 <Table 2>와 같이 정리할 수 있다. 표에서 보는 바와 같이 a는 실제 positive class에 속하는 사례 중에서 positive라고 옳게 예측한 사례의 총 수를 의미하며, b는 positive class 에 속하는 사례 중에서 negative라고 잘못 예측한 사례의 총수를 의미한다. c 는 실제 negative class에 속하는 사례 중에서 positive라고 잘못 예측한 사례의 총 수를 의미하며, d는 negative class에 속하는 사례 중에서 negative라고 옳게 예측한 사례의 총수를 뜻한다. 이와 같이 정의할 때 예측정확도, 예측오차, 민감도, 특이도 등과 같이 분류기의 성과를 평가하기 위한 다양한 지표값은 아래와 같은 산식에 의해 계산할 수 있다.

〈Table 2〉 Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	a	b
	Negative	c	d

Prediction Accuracy (예측정확도)= $\frac{(a+d)}{(a+b+c+d)}$

Error Rate(예측 오차)= $\frac{(b+c)}{(a+b+c+d)}$

Sensitivity(민감도) = $\frac{a}{(a+b)}$

Specificity(특이도) = $\frac{d}{(c+d)}$

True positive rate= $\frac{a}{(a+b)}$

False positive rate= $\frac{c}{(c+d)}$

True negative rate = $\frac{d}{(c+d)}$

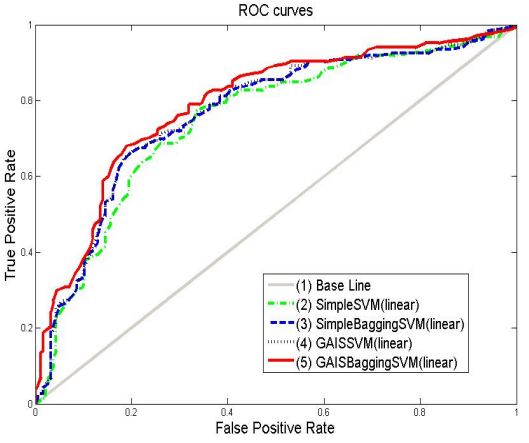
False negative rate= $\frac{b}{(a+b)}$

ROC(Receiver operating characteristics) 그래프는 X축에 false positive rate를 Y축에는 true positive rate를 표시한 2차원 그래프로 true positive와 false positive와의 상대적인 절충(tradeoff) 관계를 나타내 주고 있다 (Fawcett, 2006). ROC 분석은 분류기들의 성과를 도식화하는데 매우 유용한 방법으로 분류기들의 성과를 비교하고 보다 우수한 분류기를 선택하기 위해 이용된다. 의사결정 트리와 같이 결과값이 단지 하나의 class의 값을 나타내는 이산형 분류기(discrete classifier)는 ROC space 상에서 단지 한 점과 대응된다. 반면, Naive Bayes Classifier나 인공신경망(Artificial neural networks)과 같은 분류기는 결과값으로 단지 이산형 값이 나오는 것이 아니고 특정 사례가 어떤 class의 member인지의 정도를 나타내는 연속형 값이 산출물로 나오며, 이 값을 바탕으로 특정 임계치(threshold)를 기준으로 하여 분류를 수행하게 된다. 이와 같은 연속형 분류기는 분류를 수행할 때 기준이 되는 임계치 값의 변화를 줌으로써 ROC space 상에 서로 다른 점을 나타낼 수 있으며 이것을 연결한 것이 ROC 커브이다. 본 논문에서 기저 분류기로 사용한 SVM 모형은 결과값이 이진값인 이산형 분류기이므로 ROC space 상에 단지 한 점만 대응되지만, class boundary로부터의 거리 계산을 통해 특정 사례가 어떤 class의 member인지 정도를 알 수 있으며 이를 통해 연속형 분류기 형태로 변형이 가능하다. 이와 같이 SVM의 결과값을 변형한 형태로 사용하여 일반 연속형 분류기처럼 임계치에 변화를 줌으로써 ROC 커브를 완성할 수 있으며 본 연구에서는 이와 같은 방법을 이용하여 ROC 커

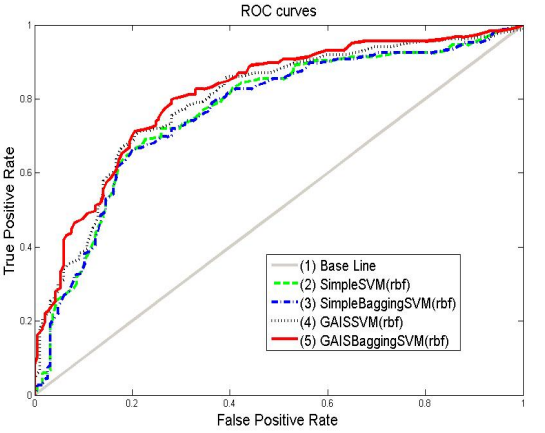
브를 구현하였다.

<Figure 5>와 <Figure 6>은 본 논문에서 제안한 모형과 다양한 비교 모형의 ROC 커브를 보여주고 있다. 여기에서 SimpleSVM는 SVM 단일 모형을 의미하며 SimpleBaggingSVM은 SVM을 기저분류기로 하는 기본 배깅 모형을 의미한다. GAISSVM은 유전자 알고리즘을 이용하여 최적의 사례를 선택한 SVM 모형을 의미하며 본 논문에서 제안한 모형인 유전자 알고리즘을 이용한 사례 선택과 배깅을 연결한 모형은 GAIS BaggingSVM으로 표기하였다. 각각의 모형은 SVM의 linear 커널과 rbf 커널을 사용하여 각각 실험을 하였으며 사용한 커널 함수는 괄호 안에 표기하였다. <Figure 5>는 linear 커널을 사용한 실험결과를 보여주고 있고 <Figure 6>은 rbf 커널을 사용한 실험결과를 보여주고 있다. 각각의 그림에서 Y=X를 의미하는 기준선(Base Line)은 임의 추측(randomly guessing)한 결과를 의미한다. 이 기준선으로부터 오른쪽 아래에 위치한 분류기는 임의 추측값보다 성과가 좋지 않다는 것을 의미하며, 왼쪽 위에 위치한 분류기는 임의 추측보다 성과가 좋은 분류기를 의미한다. 실험 결과 모든 모형의 ROC 커브가 기준선보다 왼쪽 위쪽에 위치하고 있음을 알 수 있다. 즉, 모든 모형이 임의 추측보다 좋은 결과를 나타내고 있다는 것을 알 수 있다. 그 중에서도 본 논문에서 제안한 모형인 GAISBaggingSVM모형이 Base Line을 기준으로 가장 바깥쪽에 위치해 있는 것을 확인 할 수 있다. 그러므로, ROC 커브를 기준으로 볼 때 제안한 모형이 가장 우수하다는 것을 알 수 있다.

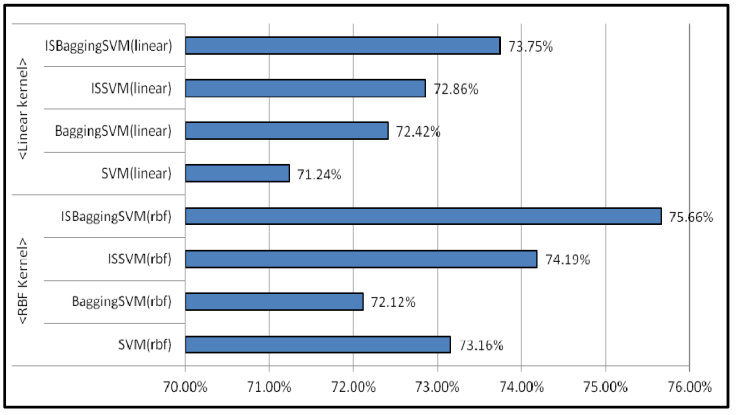
ROC 커브는 여러 분류기들의 성과를 도식화하고 비교하기에 유용한 방법이지만 정확한 정량값을 제공해 주지 못하는 단점이 있다. 이와 같은 단점을 극복하기 위해 많이 사용되고 있는 성과지표로 AUC 라는 것이 있다. AUC는 area under the ROC curve의 약자로 ROC 커브 아래의 면적을 의미한다. 그림에서 Base Line으로 표시된 임의 추측 모형의 경우 직선 아래의 면적이 0.5이므로 AUC값은 0.5가 되며 기준선 왼쪽 위에 ROC 커브가 위치할 경우 0.5보다 큰 값을 갖게 되며 반대의 경우는 0.5보다 작은 값을 갖게 된다. AUC 값은 0과 1사이의 값을 갖게 되며 값이 클수록 좋은 모형이라고 볼 수 있다. 완벽하



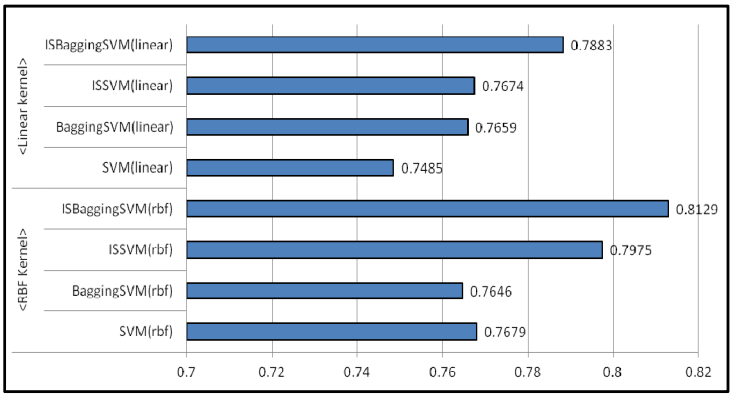
〈Figure 5〉 ROC curves (linear kernel)



〈Figure 6〉 ROC curves (rbf kernel)



〈Figure 7〉 Model Prediction Results (Accuracy)



〈Figure 8〉 Model Prediction Results (AUC)

게 분류를 하는 분류기(perfect classifier)의 경우 AUC 값이 1이 나온다.

<Figure 7>과 <Figure 8>은 각 모형 별 예측 정확도(Accuracy) 값과 AUC 값을 보여주고 있다. <Figure 7>에서 보는 바와 같이 예측 정확도를 기준으로 볼 때 본 논문에서 제안한 새로운 모형인 GAISBaggingSVM 모형이 각 커널에서 가장 좋은 성과를 보임을 알 수 있다. 또한 <Figure 8>에서도 본 논문에서 제안한 모형이 AUC를 기준으로 가장 좋은 값을 보임을 알 수 있다.

본 논문에서 제안한 모형과 비교 모형들의 성과 차이에 대한 통계적 유의성을 검토하기 위해 맥네마 검정(McNemar test)을 수행하였으며 그 결과는 <Table 3>와 같다. 표에서 ** 표시는 1% 수준에서 유의한 차이가 있다는 것을 의미한다. 맥네마 검정 분석 결과 Linear 커널 함수를 사용한 경우 제안한 모형이 단순 svm과 단순 배깅 모형보

〈Table 3〉 McNemar Test - p-value (l:linear kernel, r:rbf kernel)

`	BaggingSVM(l)	ISSVM(l)	ISBagging(l)	SVM(r)	BaggingSVM(r)	ISSVM(r)	ISBaggingSVM(r)
SVM(l)	0.134	0.035**	0.000**	0.002**	0.263	0.000**	0.000**
BaggingSVM(l)		0.453	0.049**	0.458	0.727	0.043**	0.000**
ISSVM(l)			0.238	0.856	0.267	0.122	0.001**
ISBaggingSVM(l)				0.572	0.027	0.629	0.004**
SVM(r)					0.265	0.337	0.016**
BaggingSVM(r)						0.020**	0.000**
ISSVM(r)							0.021**

다는 통계적으로 유의한 차이가 있는 것으로 나왔지만 유전자 알고리즘을 이용한 사례선택 모형과 비교할 때는 통계적으로 유의한 차이가 없는 것으로 나왔다. 하지만 rbf 커널 함수를 사용할 경우 제안한 모형이 다른 모든 비교 모형보다 통계적으로 유의한 차이가 있는 것으로 나와 본 논문에서 제안한 모형의 우수성을 알 수 있었다.

6. 결론

앙상블 모형은 개별 모형보다 더 좋은 성과를 내기 위해 여러 개의 분류기를 결합하는 것이다. 최근에는 여러 개의 모형을 결합하는 앙상블 모형을 부도 예측에 적용해 보려는 연구가 큰 관심을 끌고 있다. 이와 같은 앙상블 분류기는 분류기의 일반화 성능을 개선하는 데 매우 유용한 것으로 알려져 있다.

본 연구는 대표적인 앙상블 기법인 배깅의 성과 개선에 관한 연구이다. 배깅의 성과 개선을 위해 본 연구에서는 사례 선택기법을 활용한 배깅 모형을 제안하였다. 사례 선택은 데이터 마이닝 분야에서 매우 효과적인 기법 중의 하나로 원 데이터에서 불필요한 데이터, 관련 없는 데이터 또는 모형 개발에 오히려 해를 끼치는 노이즈와 같은 데이터를 제거하고 모형 개발을 위해 핵심적이고 중요한 사례를 선택하는 것을 말한다. 사례 선택을 적절하게 수행하였을 경우 불필요한 사례를 제거함으로써 데이터 사이즈를 줄일 수 있으며, 선택된 핵심적인 사례를 사용할 경우 원 데이터 모두를 사용하는 것과 비교해 비슷하거나 심지어는 더 좋은 성과를 기대할 수 있다. 사례 선택과 배깅은 데이터 마이닝 분야에서 잘 알려져 있는 기법으로 이들 각각은 모형의 성과를 개선시킬 수 있는 잠재력이 있지만 이들 두 기법의 결합에 대한 연구는 아직까지 없는 것이 현실이다.

본 연구는 앙상블 부도 예측 모형의 성과 개선을 위해 유전자 알고리즘을 이용한 사례 선택과 배깅의 연결에 관한 새로운 방법을 제안하였다. 유전자 알고리즘은 배깅의 입력데이터로 사용될 최적의 사례를 선택하기 위해 사용되었다. 이와 같이 유전자 알고리즘을 통해 선택된 최적의 사례들이 배깅의 최초 입력데이터로 사용되었다. 본 연구에서 제안한 새로운 앙상블 모형의 성과를 검증하기 위해 기존의 단일 모형, 사례 선택을 활용한 모형, 단순 배깅 모형을 비교 모델로 사용하였으며, ROC 커브, AUC, 예측정확도 등과 같은 다양한 성과지표를 사용해 비교 분석해 보았다. 실제 기업데이터를 사용해 실험한 결과 본 논문에서 제안한 새로운 형태의 모형이 가장 좋은 성과를 보임을 알 수 있었다.

본 연구의 한계와 향후 연구 방향을 정리하면 다음과 같다. 우선 본 연구에서 제안한 모형의 우수성을 검증하기 위해서는 보다 다양한 데이터에서의 검증이 필요할 것으로 보인다. 또한 앙상블 모형의 성과는 파라미터의 값에 따라 그 성과가 차이가 있으므로, 파라미터의 영향을 통제하기 위한 보다 다양한 실험이 추가로 필요할 것으로 여겨진다. 본 연구에서 제안한 모형은 부도 예측 문제가 아닌 다른 예측 문제에도 적용 가능할 것이다. 이에 대한 검증을 위해 추가적인 연구가 필요할 것으로 여겨진다.

참고문헌(References)

Ahn, H., K.-j. Kim, and I. Han, “Simultaneous Optimization Model of Case-Based Reasoning

for Effective Customer Relationship Management," *Journal of Intelligence and Information Systems,* Vol.11, No.2(2005),175~195.

Altman, E. I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, Vol.23, No.4(1968), 589~609.

Beaver, W. H., "Financial ratios as predictors of failure," *Journal of Accounting Research*, Vol.4(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.

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.

Derrac, J., C. Cornelis, S. García, and F. Herrera, "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection," *Information Sciences*, Vol.186, No.1(2012), 73~92.

Dietterich, T. G., "Machine-learning research: Four current directions," *AI Magazine*, Vol.18, No.4(1997), 97~136.

Dimitras, A. I., S. H. Zanakis, and C. Zopounidis, "A survey of business failure with an emphasis on prediction methods and industrial applications," *European Journal of Operational Research*, Vol.90, No.3(1996), 487~513.

Fawcett, T., "An Introduction to ROC Analysis," *Pattern Recognition Letters*, Vol.27, No.8(2006), 861~874.

García, V., A. I. Marqués, and J. S. Sánchez, "On the use of data filtering techniques for credit risk prediction with instance-based models," *Expert Systems with Applications,* Vol.39, No.18(2012), 13267~13276.

Hart, P. E., "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, Vol.14 (1968), 515~516.

Hong, S.-H., K.-S. Shin, "Using GA based Input Selection Method for Artificial Neural Network Modeling: Application to Bankruptcy Prediction," *Journal of Intelligence and Information Systems,* Vol.9, No.1(2003), 227~249.

Kim, D., S.-H. Min., I. Han, "Corporate Credit Rating using Partitioned Neural Network and Case-Based Reasoning," *Journal of Information Technology Applications and Management,* Vol.14, No.2(2007), 151~168.

Kim, K.-j., "Data Mining using Instance Selection in Artificial Neural Networks for Bankruptcy Prediction," *Journal of Intelligence and Information Systems,* Vol.10, No.1(2004), 109~123.

Kim, K.-j. and H. Ahn, "Optimization of Support Vector Machines for Financial Forecasting," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 241~254.

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

Kim, M., "Optimal Selection of Classifier Ensemble Using Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.16,

No.4 (2010), 99~112.

Kim, M.-J., "Ensemble Learning with Support Vector Machines for Bond Rating," *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012), 29~45.

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.

Kuncheva, L. I., *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.

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.

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

Meyer, P. A. and H. W. Pifer, "Prediction of bank failures," *The Journal of Finance,* Vol.25, No.4(1970), 853~868.

Min, S.-H., "Developing an Ensemble Classifier for Bankruptcy Prediction," *Journal of the Korea Society Industrial Information System*, Vol.17, No.7(2012), 139~148.

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

Ok, J.-k. and K.-j. Kim, "Integrated Corporate Bankruptcy Prediction Model Using Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.15, No.4(2009), 99~121.

Shaw, M. J. and J. A. Gentry, "Using and expert system with inductive learning to evaluate business loans," *Financial Management*, Vol.17, No.3(1988), 45~56.

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.

Tai, Q.-y. and K.-s. Shin, "GA-based Normalization Approach in Back-propagation Neural Network for Bankruptcy Prediction Modeling," *Journal of Intelligence and Information Systems*, Vol.16, No.3(2010), 1~14.

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), 926~947.

Vapnik, V. N., *The nature of statistical learning theory*, Springer, New York, 1995.

Abstract

Bankruptcy prediction using an improved bagging ensemble

Sung-Hwan Min*

Predicting corporate failure has been an important topic in accounting and finance. The costs associated with bankruptcy are high, so the accuracy of bankruptcy prediction is greatly important for financial institutions. Lots of researchers have dealt with the topic associated with bankruptcy prediction in the past three decades. The current research attempts to use ensemble models for improving the performance of bankruptcy prediction. Ensemble classification is to combine individually trained classifiers in order to gain more accurate prediction than individual models. Ensemble techniques are shown to be very useful for improving the generalization ability of the classifier.

Bagging is the most commonly used methods for constructing ensemble classifiers. 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. Instance selection is to select critical instances while deleting and removing irrelevant and harmful instances from the original set. Instance selection and bagging are quite well known in data mining. However, few studies have dealt with the integration of instance selection and bagging. This study proposes an improved bagging ensemble based on instance selection using genetic algorithms (GA) for improving the performance of SVM. GA is an efficient optimization procedure based on the theory of natural selection and evolution. GA uses the idea of survival of the fittest by progressively accepting better solutions to the problems. GA searches by maintaining a population of solutions from which better solutions are created rather than making incremental changes to a single solution to the problem. The initial solution population is generated randomly and evolves into the next generation by genetic operators such as selection, crossover and mutation. The solutions coded by strings are evaluated by the fitness function.

The proposed model consists of two phases: GA based Instance Selection and Instance based Bagging. In the first phase, GA is used to select optimal instance subset that is used as input data of

* Corresponding Author: Sung-Hwan Min
Department of Business Administration, Hallym University
39 Hallymdaehak-gil, Chuncheon Gangwon-do, 200-702, Korea
Tel: +82-33-248-1841, Fax: +82-33-256-3424 E-mail: shmin@hallym.ac.kr

bagging model. In this study, the chromosome is encoded as a form of binary string for the instance subset. In this phase, the population size was set to 100 while maximum number of generations was set to 150. We set the crossover rate and mutation rate to 0.7 and 0.1 respectively. We used the prediction accuracy of model as the fitness function of GA. SVM model is trained on training data set using the selected instance subset. The prediction accuracy of SVM model over test data set is used as fitness value in order to avoid overfitting. In the second phase, we used the optimal instance subset selected in the first phase as input data of bagging model. We used SVM model as base classifier for bagging ensemble. The majority voting scheme was used as a combining method in this study.

This study applies the proposed model to the bankruptcy prediction problem using a real data set from Korean companies. The research data used in this study contains 1832 externally non-audited firms which filed for bankruptcy (916 cases) and non-bankruptcy (916 cases). Financial ratios categorized as stability, profitability, growth, activity and cash flow were investigated through literature review and basic statistical methods and we selected 8 financial ratios as the final input variables. We separated the whole data into three subsets as training, test and validation data set. In this study, we compared the proposed model with several comparative models including the simple individual SVM model, the simple bagging model and the instance selection based SVM model. The McNemar tests were used to examine whether the proposed model significantly outperforms the other models. The experimental results show that the proposed model outperforms the other models.

Key Words : Bagging, Instance Selection, Ensemble, Bankruptcy Prediction, Genetic Algorithms

Received : November 12, 2014 Revised : December 17, 2014 Accepted : December 18, 2014
Type of Submission : Fast Track Corresponding Author : Sung-Hwan Min

저 자 소 개

민 성 환

현재 한림대학교 경영학부 교수로 재직 중이다. KAIST 테크노 경영대학원에서 경영정보시스템을 전공하여 박사를 취득하였다. 주요 관심 분야는 데이터 마이닝, 재무예측 모형 개발, 고객관계관리 등이다.