

## 기업신용등급 예측을 위한 랜덤 포레스트의 응용<sup>1</sup>

김 성 진<sup>2</sup> · 안 현 철<sup>3</sup>

### 논문초록

지금까지 금융기관의 신용위험관리를 효과적이고 효율적으로 수행하기 위한 지능형 기술에 대한 연구가 활발하게 이루어져 왔다. 특히 기업의 부실 예측이나 신용등급 예측 문제를 해결하는데 있어, 다양한 기계학습 분류 알고리즘들이 제안되고 적용되어 왔는데, 이들은 효과적인 입력변수 선정에 대한 해답을 제공해 주지 못하고, 과적합 및 잡음·이상치에 취약할 수 있는 위험이 있으며, 각종 모수에 대한 조정 작업이 요구되는 한계점이 있었다. 이러한 기존 기법들의 한계를 극복하기 위한 대안으로 본 연구는 랜덤 포레스트(Random Forests)의 응용을 제안한다. 미국의 통계학자 Breiman(2001)에 의해 제안된 랜덤 포레스트는 배깅(bagging)과 무작위 입력변수 선택(random input selection)을 통해 단일 의사결정 나무를 다수의 의사결정 나무로 확장시킨 기계학습 기법이다. 랜덤 포레스트는 과적합에서 자유로우며, 잡음이나 이상치에 받는 영향이 적고, 높은 정확도를 얻을 수 있는 특징을 지닌다. 본 연구에서는 기업신용등급 예측에 랜덤 포레스트 기법을 적용해 보고, 그 성과를 전통적으로 사용되어 온 다른 기법들과 비교하여 예측성과의 개선이 있는지 확인해 보고자 하였다. 이를 위해 본 연구에서는 1,295개 국내 상장 기업을 대상으로 하는 기업신용등급 평가모형 구축에 랜덤 포레스트를 적용해 보았다. 제안된 기법의 성과를 비교평가하기 위한 전통적인 분류 기법으로는 다중판별분석, 인공신경망, 다분류 SVM 모형을 사용하였다. 실증분석 결과, 제안 기법이 전통적인 기법들과 비교해, 보다 정확한 예측결과를 산출함을 확인할 수 있었다.

주제어: 랜덤 포레스트, 신용위험관리, 기업신용등급평가, 다분류 분석, 데이터 마이닝

1 이 논문 또는 저서는 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2014S1A5A2A01013060).

2 국민대학교 비즈니스IT전문대학원 석사과정. 136-702 서울 성북구 정릉로 77 tel: 02-910-4018, e-mail: scott\_kim@kookmin.ac.kr (주저자)

3 국민대학교 비즈니스IT전문대학원 부교수. 136-702 서울 성북구 정릉로 77 tel: 02-910-4577, e-mail: hcahn@kookmin.ac.kr (교신저자)

## I. 서 론

오늘날 금융기관들에게 자사가 보유하고 있는 다양한 형태의 신용 위험 자산에 대하여 그 위험을 측정, 평가하는 것은 조직의 번영과 생존을 결정짓는 매우 중요한 의사 결정 문제가 되고 있다. 특히 지난 97년 외환위기와 2008년 글로벌 금융위기 이후, 재무적 자산에 대한 보다 조심스러운 관리가 금융 시장 전반에 요구되면서, 기업부도예측 모형과 신용평가시스템으로 대표되는 신용위험관리(Credit Risk Management)에 대한 중요성이 최근 수년간 대단히 높아진 상황이다(최진배, 2002). 여기에 새로운 바젤(BASEL) 협약에서는 발생할지도 모르는 미래 위험, 즉 기대 손실에 대해서도 충당금을 쌓도록 요구하는데, 은행이 자체적인 실증 모형을 개발해 신용 위험에 따른 충당금을 산출할 경우, 표준 방식을 적용할 때보다 충당금에 대한 부담이 적어지게 된다(De Andres et al., 2012). 심지어 금융기관의 위험관리 개선은 해당 기관이 속한 국가의 경제성장에도 도움이 될 수 있다(이규복·하준경, 2012). 이런 이유들로 인해 데이터 기반 신용위험관리 모형에 대한 산업계의 관심이 더욱 폭발적으로 높아지고 있는 상황이다.

학계에서도 지능형 정보처리기술을 활용하여 신용위험관리를 효과적으로 수행하고자 하는 연구를 지난 수십여년간 활발하게 수행하여 왔다(Kim & Ahn, 2012; Kim et al., 1993; Kumar & Ravi, 2007). 특히 최근 들어 이른바 빅데이터 시대가 도래하면서, 금융기관들은 자사가 보유하고 있는 방대한 데이터를 활용해 신용위험관리를 보다 과학적이고 체계적으로 수행하는데 큰 관심을 가지고 있다. 금융 및 보험 산업은 빅데이터의 잠재적 활용 가치가 가장 높은 산업군으로 기대되고 있는데, 이러한 금융 산업 내 여러 분야들 중에서도 가장 직접적이면서도 큰 파급효과를 만들어 낼 수 있는 분야가 바로 이 신용위험관리 분야라고 할 수 있다(Manyika et al., 2011).

방법론 관점에서 금융기관의 신용위험관리 문제는 크게 2가지 유형으로 구분된다. 첫 번째 유형은 대출 혹은 투자대상 기업(혹은 개인)의 부실 여부를 판별하는 부실 예측(bankruptcy prediction) 문제이다. 부실 예측 문제를 해결하기 위해서는 0(정상) 혹은 1(부실)로 구분되는 종속변수를 예측할 수 있도록 설계된 이분류(binary classification) 방법론이 요구된다. 신용 위험 관리 문제의 두 번째 유형은 대출 혹은 투자대상 기업(혹은 개인)의 신용 등급 수준을 판별하는 신용등급평가(credit rating)

문제이다. 일반적으로 기업 신용등급평가의 경우, 적게는 5등급, 많게는 10개 이상의 등급으로 구분되어 있으므로, 이러한 문제를 해결하기 위해서는 다분류(multiclass classification) 방법론이 적용되어야 한다. 이 두 가지 유형 중 본 연구는 후자인 신용등급평가의 효과적 수행을 위한 대안으로 ‘랜덤 포레스트(Random Forests, 이하 RF)’라는 새로운 기법의 응용을 제안한다.

지금까지 학계에서는 기업신용등급평가 예측을 위한 지능형 기법으로 다항 로지스틱 회귀분석(Multinomial LOGIT, MLOGIT)이나 다중판별분석(Multiple Discriminant Analysis, MDA)와 같은 통계 기법, 그리고 인공신경망(Artificial Neural Network, ANN), 사례기반추론(Case-based Reasoning, CBR), 의사결정나무(Decision Tree) 및 다분류 문제를 해결할 수 있도록 변형된 다분류 SVM(Multiclass SVM)과 같은 기계학습 기법들을 주로 적용하여 왔다(Kim and Ahn, 2012; Kim et al., 1993; Shin and Han, 2001 등). 이처럼 다양한 기계학습 기법들이 제안되고 적용되어 왔지만, 기존의 분류 기법들은 종종 효과적인 입력변수 선정에 대한 해답을 제공해 주지 못하고, 과적합화(overfitting)이나 잡음(noise), 이상치(outlier)에 취약할 수 있는 위험을 수반하고 있고, 많은 모수(parameter)에 대한 조정 작업이 요구된다는 문제를 지적받아 왔다.

이에 본 연구는 기업 신용위험관리 분야에서 기존 알고리즘을 대체할 수 있는 새로운 기법으로 RF의 응용을 제안한다. RF는 CART(Classification and Regression Trees)의 창시자로 유명한 Leo Breiman이 지난 2001년에 처음 제안한 분류기법으로서, 여러 개의 의사결정나무들을 숲으로 확장하여 예측의 정확도와 안정성 향상을 도모하는 기계학습 방법이다(Breiman, 2001). RF는 금융 분야의 빅데이터를 활용하여 신용위험 관리를 구현하는데 적합한 많은 특징점을 가지고 있음에도 불구하고, 지금까지 국내·외 경영 분야의 연구에서 거의 다루어지지 않았다. 이에 본 연구는 RF 기법을 기업신용등급 평가 예측에 적용해 보고, 그 결과를 전통적인 다분류 기법들과 비교하여 제안기법의 적용가능성을 확인해 보고자 한다.

## II. 이론적 배경

### 2.1 기업신용등급 평가

기업신용등급은 기업의 재무상태나 경영성과를 나타내는 회계정보 뿐만 아니라 경제 및 산업 동향이 기업가치에 미치는 영향을 고려하여 평가되는데, 주주나 채권자 같은 투자자들이 그들의 투자의사결정 시 중요하게 고려하는 요인 중 하나라는 점에서 중요하다(김문헌, 2013). 기업신용등급 평가 분야의 초창기 연구들은 OLS(Ordinary Least Squares), 다중판별분석 등과 같은 통계 기법에 기반한 신용등급평가모형을 연구해 왔으나, 낮은 예측정확도 문제로 인해 1980년 이후부터는 인공신경망, 귀납적 학습방법, 사례기반추론, 유전자 알고리즘, 다분류 SVM 등 인공지능기법이 주로 적용되고 있다(안현철·김경재·한인구, 2006). 이 중에서 최근까지도 가장 많이 연구되고 있고, 현장에서 활용되고 있는 방법으로 인공신경망과 다분류 SVM이 있다.

인공신경망 기법을 기업신용등급 예측에 적용한 연구들은 주로 비교모형으로 다중판별분석, 다항 로지스틱 회귀분석, RBF(Radial Basis Function), LVQ(Learning Vector Quantization) 등을 인공신경망과 비교하였다. 해당 연구들에서 인공신경망의 기업신용등급평가 예측력은 비교모형들을 뛰어넘는 성과를 보여주었다(Chaveesuk et al., 1993; Kwon et al., 1997; Lee et al., 1996; Maher & Sen, 1997). 이처럼 많은 연구들을 통해 증명된 인공신경망의 탁월한 성능에도 불구하고, 인공신경망은 기법 자체에 여러 한계점을 지니고 있다. 인공신경망은 예측 결과의 원인을 설명하기 어렵고, 기법 적용시 데이터가 소규모일 때 나타나는 일반화의 어려움도 한계점으로 작용된다. 또한 모형 구축시 발생할 수 있는 과적합 문제와 신경망 설계에 투자되는 시간과 노력도 한계점으로 나타난다(Altman et al., 1994; Jo & Han, 1996).

이러한 문제점을 해결하기 위한 대안으로 SVM, 그 중에서도 다분류 문제를 해결하기 위해 특별하게 변형된 다분류 SVM이 최근 채권등급예측평가 기법으로 많이 사용되고 있다(안현철·김경재·한인구, 2006; Kim & Ahn, 2012). 많은 연구들에서 다분류 SVM은 인공신경망이 갖는 다양한 문제점들은 회피하면서, 예측 정확도 측면에서 인공신경망과 유사하거나 혹은 더 나은 수준의 성과를 보여주고 있다고 보고되고 있다

(안현철·김경재·한인구, 2006; Huang et al., 2004; Kim & Ahn, 2012). 하지만, 다분류 SVM 역시 효과적인 입력변수 선정에 대한 과정이 알고리즘 내부에 포함되어 있지 않다는 점, 많지는 않지만 커널함수 및 커널파라미터 등과 같은 직관에 의해 설정되어야 할 모수들이 일부 있다는 점에서 한계가 있다.

## 2.2 Random Forests

RF는 Breiman(2001)에 의해 개발된 분류기법으로서, 전통적인 의사결정나무 기법을 하나가 아닌 여러 개의 나무로 확장시킨 의사결정나무의 메타학습(meta-learning) 형태를 갖고 있는 기계학습 기법이다. RF를 구성하는 각 의사결정나무는 무작위로 선택된 학습 데이터와 입력변수들에 의해 형성되는데, 이 경우 각 개별 의사결정나무의 정밀도는 떨어질 수 있으나, 이들을 종합하여 예측을 수행하게 되는 숲(즉, RF)의 정확도와 안정성은 높아지게 된다.

RF의 경우, 대수의 법칙에 의해 숲의 크기(나무의 수)가 커질수록 일반화 오류가 특정 값으로 수렴하게 되어 과적합화를 피할 수 있으며, 각 개별 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음이나 이상치로부터 크게 영향을 받지 않는다(한은정, 2004). 무엇보다 RF가 갖는 가장 큰 장점은 모형의 설계자가 입력변수 선정으로부터 자유로울 수 있다는 점이다. 때문에, 수백여개에 이르는 많은 수의 독립변수와 방대한 양의 학습 사례로부터 분류·예측을 수행하여야 하는 기업 신용위험관리 분야에 매우 적합한 기법이 될 수 있다. 또한 최근 발표된 한 연구(Brown & Mues, 2012)에 따르면 RF는 빈도가 불균형한(imbalanced) 이항분류의 예측에 있어 가장 우수한 예측력을 보인 것으로 보고되고 있다.

이처럼 많은 장점이 있는 기법임에도 불구하고, RF를 경영 분야의 의사결정, 그 중에서도 특히 금융 분야의 기업 신용 위험 관리에 분류·예측 알고리즘으로 적용한 연구는 최근 발표된 소수의 사례(Brown & Mues, 2012; Hajek & Michalak, 2013; Kalsyte & Verikas, 2013)를 제외하고 그간 충분히 이루어지지 못했다. 또한, 국내에서도 의료·보건 분야(한은정, 2004; 윤태균·이관수, 2008)나 공학 분야(이현주 외, 2011; 최혁진·최성욱·한경숙, 2012; 홍준혁·고병철·남재열, 2013)에서 일부 연구가 발표되었을 뿐, 경영 분야의 지능형 의사결정 문제에 적용된 사례는 거의 찾아보기 어렵다. 이러한 배경에서 본 연구는 기존 이공학 분야에서 주로 적용되어 온 RF를 국내 신용위험 관리, 그

중에서도 기업신용등급평가 예측에 적용해 보고, 그 우수성을 검증해 보고자 하였다.

### III. 연구모형

RF는 단일의 의사결정나무를 만드는 대신, 여러 개의 나무로 확장시킴으로서 분류의 정확도를 높이려 개발된 분류 알고리즘이다. 본 알고리즘은 Breiman(1996)에 의해 제안된 배깅(bagging) 방법과 Amit & Geman(1997)에 의해 처음 제안된 무작위 입력 변수 선택(random input selection)에 기반한다.

“Bootstrap aggregating”을 줄여서 표현한 배깅은 말 그대로 다수의 부트스트랩들을 결합하여, 예측을 수행하는 기법을 말한다. 여기서 배깅의 기반을 이루는 부트스트랩을 생성하는 부트스트래핑(bootstrapping)은 Efron(1979)에 소개된 비모수 통계기법으로서 기존 모수 통계기법이 가진 모집단 추론의 한계를 극복하고자 개발되었다. 부트스트래핑에서는 실제 모집단 특성에 가깝도록 원표본 크기와 동일한 표본을 복원(replacement) 과정을 통해 반복 생성하는데, 이 과정에서 동일한 관측치가 중복적으로 포함될 수도 있다(심준섭, 2004; Efron & Tibshirani, 1993). 이런 과정을 적용하여 부트스트랩 표본이 다수 생성되고 나면, 배깅은 각 부트스트랩 표본에 대한 각각의 분류 모형을 평행하게(parallel) 생성한 다음, 각 모형별 예측 결과를 산출하게끔 하여 분류 문제시 각 모형별 예측결과를 투표(voting)방식으로 결합하고, 수치예측 문제시 평균(averaging)을 통하여 결합한다(Witten et al., 2011).

이러한 배깅에 무작위 입력변수 선택을 결합해 다수의 의사결정나무를 생성한 후, 이들의 예측결과를 결합하도록 설계한 기법이 바로 RF이다. 이렇게 특징변수(feature)와 학습용 표본(instance) 모두에 대해 무작위성을 포함하고 있어, RF는 이론적으로 과적합으로부터 자유로울 수 있고, 잡음이나 이상치에 영향을 크게 받지 않으며, 일반화 오류를 낮춰 높은 정확도를 얻을 수 있다.

다음의 <그림 1>은 RF 알고리즘의 일반적인 수행 과정을 나타내는데, 이 과정을 상세히 설명하면 다음과 같다(Breiman, 2001).

단계1: 우선 RF에 적용될 기본 모수에 대한 값을 정한다. RF 알고리즘을 학습하기 위해서는 ntree와 mtry를 사전에 결정해야 한다. 이 중, ntree는 RF를 구성할 전체 의사결정나무의 개수이며, mtry는 RF를 구성하는 의사결정나무에서 사용될 특징변수의

개수를 의미한다. 보통  $ntree$ 는 대수의 효과를 얻을 수 있도록 큰 상수값으로 설정되는데, 기본값으로 500 정도가 설정된다.  $mtry$ 의 경우에는 전체 특징변수의 개수가  $M$ 이라고 할 때, 분류를 위한 의사결정나무(classification tree)에는  $mtry = \sqrt{M}$ , 예측을 위한 의사결정나무(regression tree)에는  $mtry = M/3$ 이 적용된다(권안나, 2013).

단계2: 부트스트랩 표본  $X_i$ ( $i$ =부트스트랩 반복 횟수)를 추출한다. 이 때,  $X_i$ 는 전체 학습용 데이터( $X$ )에서 단순 무작위 표본 추출하되, 복원 추출(sampling with replacement) 방식으로 추출한다. 이렇게 되면 부트스트랩 표본  $X_i$ 는 평균적으로 원본 학습용 데이터  $X$  원소들의 약 2/3를 포함하게 되는데, 여기에 포함되지 않은 약 1/3가량의  $X$  원소들은 따로 OOB(Out-Of-Bag) 데이터로 인식되어, 추후 참조된다.



<그림 1> Random Forests 알고리즘 수행과정

단계3: 부트스트랩 표본  $X_i$ 를 이용해 이진 의사결정나무를 성장시킨다. 여기서, 가지를 분할하기 위한 최적 분할을 결정할 때, 전체 특징변수를 사용하지 않고 무작위로

선택된  $mtry$ 개의 특징변수만 사용한다. 최적 분할을 위한 분류 기준은 Gini 계수를 사용하며, 나무는 가지치기를 하지 않고, 최대한 성장시킨다.

단계4: 단계2에서 추출된 OOB 데이터들에 앞서 구축된 의사결정나무에 적용하여, 예측결과를 산출한다. 이 예측결과가 실제결과와 얼마나 일치하는지를 확인하여, OOB 데이터에 대한 오류율(ERROOB)을 산출한다.

단계5: 단계2~단계4의 활동들을  $ntree$ 회 반복하여,  $ntree$ 개의 개별 의사결정나무로 구성된 RF를 최종적으로 구성한다. 이 때, RF의 최종 예측값은 종속변수가 연속형인 경우 평균, 범주형인 경우 투표를 사용해 개별 의사결정나무의 예측값을 종합하여 산출한다. 아울러 각 나무들로부터 산출된 ERROOB를 종합하면, 기준으로 설정된  $ntree$ 와  $mtry$  값이 얼마나 적절한 지를 점검하는 지표로 사용할 수 있다.

## IV. 실증분석

### 4.1 실험데이터

본 연구에서 제안하는 RF 알고리즘이 실제로 우수한 성능을 보이는지 검증해보기 위해 국내 한 신용평가회사의 기업신용등급평가 예측모형 개발에 제안모형을 적용해 보았다. 연구에서 사용되는 데이터는 KOSPI또는 KOSDAQ에 상장되어 있는 실제 1,295개의 제조업으로 분류되는 기업들의 데이터이다. 해당 기업들의 신용등급을 목적 변수로 설정하여 예측실험을 수행하였다. 신용등급은 한국신용정보에서 2002년에 공시한 자료를 기준으로 실험에 사용하였다. 여기서 기업의 신용등급은 총 5가지로 A1, A2, A3, B, C로 나뉘어진다.

실험에 앞서 연구의 편의성을 위해 신용등급을 A1은 1, A2는 2, A3은 3, 그리고 B와 C는 통합하여 4로 각각 표기하였다. B와 C등급을 통합하여 표기한 이유는 데이터 속에 실제 C 등급에 해당하는 기업들의 수가 적어 등급구분에 어려움이 있었기 때문이다. 또한 신용평가회사들이 기업에 대한 신용등급 구분시 'B등급 이하'를 사실상 투자부적격 채권, 즉 정크 본드(junk bond)로 해석하기 때문이다(안현철·김경재, 2009; 안현철·김경재·한인구, 2006). 다음의 <표 1>은 본 연구의 실증분석에 사용된 실험데이터의 각 등급별 빈도 및 구성비를 나타낸다.



&lt;표 1&gt; 실험데이터의 등급별 빈도 및 구성비

등 급	빈도수(건)	구성비(%)
1(A1)	106	8.19
2(A2)	552	42.63
3(A3)	332	25.64
4(B&C)	305	23.55
합 계	1,295	100.00

연구모형 구축을 위하여 각 등급별 데이터의 80%는 학습용(training)데이터로 선정하였다. 나머지 20%는 차후 연구모형을 검증(validation)하기 위한 데이터로 남겨두었다. 또한 분석에 사용될 데이터가 다소 부족하다는 한계점을 극복하기 위한 방편으로 5-집단 교차검증을 적용하였다.

후보 입력변수로 활용될 기업신용등급평가에 활용되는 요인변수로는 재무제표 등으로부터 추출된 총 39개의 재무변수를 사용하였다. 이들 변수들은 기존 관련 문헌들에서 기업신용등급평가에 유의한 영향을 미치는 것으로 제시되었던 지표들이다(박기남·이훈영·박상국, 2000; 이진창·한인구·김명중, 1996; Huang et al., 2004; Shin & Han, 1999). 이 39개 후보변수들 중에서, 제안 모형에 사용될 독립변수로 일원배치 분산분석(One-way ANOVA)과 다중판별분석의 단계별 선택을 통해 기업신용등급과 통계적으로 가장 유의한 연관이 있는 것으로 파악된 총 14개의 변수를 선정하였다(안현철·김경재, 2009; 안현철·김경재·한인구, 2006). 다음의 <표 2>는 전체 39개 후보 입력변수들의 내역과 함께, 이 중 어떤 변수들이 최종 14개 독립변수로 선정되었는지 나타내고 있다.

데이터 전처리와 관련해서는 우선 최대-최소 정규화(Max-Min normalization)를 통해 입력변수의 모든 값들이 0~1사이의 값을 갖도록 조정하였다. 또한 이상치 관리를 위하여 각 변수별 상위, 하위 각각 5%에 해당되는 값을 기준값으로 한 다음, 그 이상 혹은 이하에 해당되는 값을 기준값으로 대체하는 95% 윈저화 방법(Winsorization)을 적용하였다.

<표 2> 후보 입력변수 및 최종 선정 결과

구 분	후보변수	선정여부
규모 지표	총자산	
	유형자산	
	고정자산	
	자기자본	O
	매출액	O
	부가가치	
	총부채	O
	감가상각비	
	영업이익	
	당기순이익	
수익성 지표	업력	O
	주당순이익	O
	유보액대총자산비율	O
	금융비용부담률	O
	금융비용대부채비율	
	금융비용대총비용비율	O
	감가상각비대총비용비율	
	이자보상배율	
	총자본순이익률	
	자기자본순이익률	
	자본금영업이익률	
	자본금경상이익률	
	매출액총이익률	
	매출액경상이익률	
안정성 지표	고정자산구성비율	O
	차입금의존도	
	자기자본구성비율	
	재고자산대유동자산비율	O
	단기차입금대총차입금비율	O
	고정부채대자본금비율	
	부채비율	
	(자본+고정부채)/고정자산	

현금흐름 지표	현금흐름대부채비율	
	현금흐름대고정부채비율	
	현금흐름대총자본비율	O
	(영업활동으로인한현금흐름-현금배당)/(고정자산+운전자본)	O
생산성 지표	1인당매출액	O
	총자본사업이익률	
	총자본투자효율	

## 4.2 비교모형 설정

본 연구에서 제안하는 연구모형 RF의 성과와 비교하기 위해 다중판별분석, 인공신경망, 다분류 SVM을 본 연구의 비교모형으로 설정하였다. 이 중, 다중판별분석은 등간 혹은 비율척도로 정해진 독립변수와 명목척도로 정해진 종속변수 사이에서 두 개 이상의 집단을 판별하기 위해 사용되는 분석 기법이다. 전술했듯이 본 연구에서는 단계별 선택을 적용해 총 14개의 독립변수를 선택하였으며, 이 변수들로 구성된 판별모형을 SPSS 통계 프로그램을 사용하여 구축하였다.

인공신경망은 모형 설계에 따른 일반적인 원칙이 규정되어 있지 않기 때문에, 모형 설계에 있어서 선행 연구들을 참고하였다. 우선 인공신경망 모형의 입력변수로는 앞의 다중판별분석을 통해 도출된 14개 독립변수를 사용하였다. 인공신경망 성능에 영향을 주는 요인은 여러 가지가 존재하지만, 크게 은닉층의 수, 은닉층 노드의 수, 학습 횟수 3가지로 알려져 있다. 은닉층의 수는 1개로 적용하였고, 은닉층 노드의 수는 7, 14, 21, 28을 각각 실험한 뒤, 가장 우수한 성과를 보이는 것을 선택하였다. 또한 과도적합의 문제를 예방하기 위하여 학습용 데이터를 다시 3:1 비율로 학습용과 테스트용으로 구분하고, 실제 테스트용 데이터로 인공신경망 모델을 검증한다. 학습 횟수는 과도하거나 과소하면 학습데이터에 과도적합 되거나 학습이 제대로 이루어지지 않는 특성을 지닌다. 그러므로 779건 학습용 데이터 개수의 약 50배에 해당하는 38,950회가 지나면 학습을 중지시킨다. 추가적으로 인공신경망의 학습율(learning rate)은 0.1, 모멘텀(momentum)은 0.1로 지정하였다. 이상의 인공신경망 실험은 상용 소프트웨어인 Neuroshell2 R4.0을 이용해 진행하였다.

다분류 SVM 모형의 경우도 앞의 두 비교모형과 동일하게 14개 독립변수를 활용해 구축하였다. 다분류 SVM 모형으로는 OAO(One-Against-One), WW(Method by

Weston and Watkins), CS(Method by Crammer and Singer)의 3가지를 사용하였으며, 커널함수로는 선형과 다항식, RBF의 3가지를 사용하였다. 다분류 SVM의 경우, 커널 함수 모수의 설정값에 따라 그 성과가 달라질 수 있어, 최적의 커널함수 모수값을 찾기 위해 그리드 탐색(grid search) 방법을 적용하였다. 다분류 SVM 실험은 LIBSVM v2.80(Chang and Lin, 2011)과 BSVM v2.08(Hsu and Lin, 2012)를 사용하여 진행하였다.

#### 4.3 실험결과

본 연구에서는 제안하는 RF 알고리즘과 비교 모형으로서 사용된 다중판별분석, 인공 신경망, 다분류 SVM 모형의 성과를 전통적인 기업신용등급평가 연구에서 널리 사용되어 온 예측 정확도(hit ratio)의 관점에서 비교, 분석해 보았다(안현철·김경재, 2009; 안현철·김경재·한인구, 2006; 이견창·한인구·김명중, 1996). 이 때, 예측 정확도 HR(%)은 다음 식에 의해 산출된다.

$$HR(\%) = \frac{\sum_{i=1}^m \sum_{j=1}^{N_i} H_{ij}}{\sum_{i=1}^m N_i} \times 100 \quad \text{where } H_{ij} = 1 \text{ if } PO_{ij} = AO_{ij}, H_{ij} = 0 \text{ otherwise}$$

상기 식에서  $i$ 는 신용등급을 나타내는 식별자이며,  $PO_{ij}$ 는  $i$ 등급  $j$ 번째 기업의 예측된 신용등급,  $AO_{ij}$ 는  $i$ 등급  $j$ 번째 기업의 실제 신용등급,  $m$ 은 분석 대상이 되는 전체 신용등급의 수, 그리고  $N_i$ 는  $i$ 등급에 속한 기업의 수를 의미한다.

RF의 경우, 오픈소스 통계 프로그램인 R에서 제공되는 randomForest 패키지를 사용해 실험을 수행하였는데, 이 때 3장에서 설명했듯이 ntree와 mtry라고 하는 설계모수의 값을 적절하게 지정하여야 한다. 우선 RF를 구성할 전체 의사결정나무의 개수를 나타내는 ntree의 경우, R 패키지에서 기본값으로 제공하는 500을 적용하였다. 반면 RF를 구성하는 의사결정나무에서 사용될 특징변수의 개수를 나타내는 mtry의 경우, 권안나(2013)에서 추천한 방식을 통해 최적값을 추정해 보면  $\sqrt{14} \approx 3.74$ 가 되는데, 이에 본 연구에서는 mtry 값을 2에서 6까지 모두 적용해 보고, 이 중에서 OOB 오류율이 가장 작게 나타난 값을 선정하는 방식을 취하였다. 다음의 <표 3>은 mtry에 따른 OOB 오류율의 변화를 나타내고 있는데, 이 결과에 의거하여 본 연구에서 적용할 최적

의 mtry는 가장 낮은 평균 OOB 오류율(28.04%)을 보인 6으로 선정되었다.

<표 3> mtry 값에 따른 OOB 오류율

데이터셋	OOB 오류율(%)				
	mtry=2	mtry=3	mtry=4	mtry=5	mtry=6
#1	31.34	30.09	29.41	29.12	29.22
#2	28.93	28.45	27.87	27.87	27.29
#3	28.74	28.54	27.68	28.35	27.68
#4	28.64	27.97	27.48	27.00	26.90
#5	29.89	29.32	28.64	29.89	29.12
평 균	29.51	28.87	28.22	28.45	28.04

다음의 <표 4>는 ntree로 500을, mtry로 6을 적용했을 때, 각 데이터셋별 예측 정확도를 나타내고 있다. <표 4>에 제시된 바와 같이, RF는 검증용 데이터셋을 기준으로 평균 72.79%의 정확도로 기업신용등급을 예측할 수 있음을 확인할 수 있었다.

<표 4> Random Forests 제안 모형 결과

데이터셋	Random Forests	
	학습용	검증용
#1	70.78%	75.19%
#2	72.71%	70.93%
#3	72.32%	73.64%
#4	73.10%	70.54%
#5	70.88%	73.64%
평 균	71.96%	72.79%

본 연구에서는 RF의 성능을 검증하기 위해, 다중판별분석, 인공신경망, 그리고 다분류 SVM을 비교모형으로 선정하고, 이들의 성능도 함께 살펴보고자 하였다. 이 중 우선, 다중판별분석의 결과가 <표 5>에 제시되어 있다. 기업신용등급평가에 있어서 다중

관별분석은 학습 및 검증 모두에서 RF는 물론 이후 소개될 다른 비교모형들에 비해 상대적으로 낮은 성과를 보여준다.

<표 5> 다중판별분석 결과

데이터셋	다중판별분석	
	학습용	검증용
#1	65.68%	59.30%
#2	64.90%	62.02%
#3	64.51%	69.38%
#4	64.80%	65.50%
#5	65.09%	59.30%
평 균	65.03%	63.10%

인공신경망 결과는 <표 6>과 같이 나타난다. 인공신경망의 성과는 은닉층의 수에 따라 달라질 수 있어, 각 데이터셋을 기준으로 4가지의 경우를 모두 적용하여 최적의 은닉층 수를 확인하였다. 인공신경망의 결과는 다중판별분석에 비해서 개선되었지만, 제안모형인 RF에 비해서는 많이 정확도가 떨어짐을 확인할 수 있다.

<표 6> 인공신경망 결과

데이터셋	인공신경망			
	학습용	테스트용	검증용	최적 은닉층 노드의 수
#1	70.22%	64.73%	64.73%	h=14
#2	71.76%	65.50%	65.12%	h=7
#3	69.45%	63.18%	66.67%	h=21
#4	69.19%	64.73%	67.44%	h=28
#5	67.14%	68.60%	64.34%	h=21
평균	69.55%	65.35%	65.66%	

끝으로 다분류 SVM의 실험결과가 <표 7>에 제시되어 있다. 다분류 SVM은 인공신

경망보다 월등히 개선된 예측력을 보여주었는데, 그럼에도 불구하고 RF와 비교해서는 크게 못미치는 성능을 보이고 있음을 확인할 수 있다.

<표 7> 다분류 SVM 실험결과

데이터셋	다분류 SVM		
	학습용	검증용	최적 모수값
#1	67.12%	65.89%	WW/LINEAR, C=10
#2	84.09%	65.50%	OAO/RBF, C=78, $\sigma^2=1$
#3	67.21%	69.38%	OAO/LINEAR, C=100
#4	81.20%	68.99%	WW/RBF, C=55, $\sigma^2=1$
#5	75.31%	68.22%	OAO/RBF, C=10, $\sigma^2=1$
평균	75.06%	67.60%	

(\*) OAO: One-Against-One, WW: Method by Weston and Watkins, LINEAR: 선형커널, RBF: RBF 커널

<표 8>은 검증용 데이터셋을 기준으로 각 모형별 예측 정확도를 정리하여 제시하고 있다. 이 표를 통해 알 수 있듯이, RF는 다른 그 어떤 모형과 비교해도 검증용 데이터셋 기준으로 월등히 우수한 예측 정확도를 보여주었다. 모든 비교모형과 본 연구에서 제안하는 모형의 성과를 정리해보면, 예측력은 RF가 가장 높았고, 다분류 SVM, 인공신경망, 다중판별분석 순이었다. 그리고 각 데이터셋의 예측력을 비교해 보았을 때 모든 데이터셋에서 RF의 성과가 다른 비교모형에 비해서 뛰어났다. 또한 전체적인 평균 예측력을 비교해보면 상대 모형들에 비해 최소 5.19%에서 최대 9.69%까지 성과 차이가 발생함을 알 수 있었다.

<표 8> 검증용 데이터셋에 대한 모형별 예측 정확도

데이터셋	다중판별분석	인공신경망	다분류SVM	RF
#1	59.30%	64.73%	65.89%	75.19%
#2	62.02%	65.12%	65.50%	70.93%
#3	69.38%	66.67%	69.38%	73.64%
#4	65.50%	67.44%	68.99%	70.54%
#5	59.30%	64.34%	68.22%	73.64%
평 균	63.10%	65.66%	67.60%	72.79%

앞서 제시된 RF와 비교 모형 간 성과 차이가 통계적으로 유의한 지 검증하기 위해 비모수 통계기법인 맥네마 검정(McNemar test)을 적용하였다(안현철·김경재·한인구, 2006). 다음의 <표 9>가 그 결과를 나타내고 있는데, RF가 모든 비교 모형과 99% 신뢰수준 하에서 통계적으로 유의한 차이를 보이고 있음을 확인할 수 있다.

<표 9> 맥네마 검정 수행 결과(카이제곱 통계량 및 유의수준)

	인공신경망	다분류 SVM	RF
다중판별분석	5.044**	12.893***	48.813***
인공신경망		3.016*	30.004***
다분류 SVM			15.284***

\* 90% 신뢰수준에서 유의, \*\* 95% 신뢰수준에서 유의, \*\*\* 99% 신뢰수준에서 유의

## V. 결 론

RF가 기업 채권등급평가에서 적용 가능하다는 점을 확인해 본 본 연구는 2가지 측면에서 연구의 의의를 찾을 수 있다. 첫 번째는 경영분야 문제, 특히 기업신용위험 관리에 있어서의 RF 알고리즘의 적용 가능성을 확인했다는 점이다. RF를 경영분야에 접목하려는 시도 자체가 전세계적으로 많지 않았고, 2000년대 후반기부터 연구 성과가 조금씩 발표되고 있음을 감안할 때, 아직 초기 단계에 있다고 할 수 있다. 특히 기업 신용등급평가라는 주제에 RF가 접목된 사례가 국내에서 전무했다는 점은 본 연구가



높은 학술적 가치를 지닌다는 것을 증명한다.

RF가 기업신용등급 예측에 적용되어 온 기존 분류 알고리즘이 내재한 문제들에서 상대적으로 자유롭다는 점도 주목해 볼 필요가 있다. 인공신경망의 경우 예측 정확성이 높지만 모형이 제시하는 결과에 대한 해석이 어렵고, 결과에 대한 원인을 분석하기 쉽지 않다. 또한 모형의 학습이 복잡하고, 과적합 문제에 노출된다는 점도 인공신경망을 선택하기 쉽지 않게 만드는 한계점이다. 다분류 SVM의 경우도 커널함수를 통한 고차원으로의 사상(mapping)과 같은 복잡한 수리 연산을 포함하고 있기 때문에, 학습이 완료된 모형을 실제로 응용하는데 있어 구현이 어렵다는 한계를 지닌다. 하지만 RF의 경우는 특징변수와 학습사례의 잡음이나 이상치로부터 거의 영향을 받지 않으면서도 빠른 시간 안에 학습이 가능한 대단히 비용 효율적인 알고리즘이다. 또한 RF의 최종 학습결과는 일련의 규칙들로 표현 가능한 의사결정나무의 집합으로 구성되어 있어, 구현도 쉽고 적용도 빠르다는 장점이 있다.

특히 RF는 그 속성 상 빅데이터 분석 환경에 매우 잘 맞는다는 장점을 가지고 있다. 하둡(hadoop)으로 대표되는 빅데이터 분석 환경에 기계학습 기법을 적용하려고 할 때, 가장 큰 난제 중 하나가 분산 처리를 위해 모형 구축에 필요한 데이터와 모형을 잘게 나누어야 한다는 점이다. 하지만, RF의 경우 본래 알고리즘이 데이터와 입력변수를 적절하게 나누어 다수의 의사결정나무를 학습한 뒤 이를 나중에 결합하도록 설계되어 있어 빅데이터 환경과 그 궁합이 매우 잘 맞는다고 할 수 있다.

이렇듯 많은 특징점을 가진 기법이기에 향후 국내 학계에서 RF에 대해 더욱 관심을 갖고 후속 연구를 수행해야 할 것으로 생각된다. 우선 본 연구는 기업신용등급 예측에 RF를 적용하고 그 성능을 살펴본 초기단계의 연구이므로, 앞으로 경영분야의 다른 예측문제에 RF를 적용해 보는 후속 연구가 뒤따라야 할 것으로 보인다. 또한 앞서 빅데이터 환경과 RF가 그 속성 상 잘 맞는다는 점을 지적했는데, 빅데이터 환경에서의 효과적인 기계학습 기법으로서 RF가 얼마나 효과적인지 실제적으로 검증해 보기 위해 하둡 환경에서 RF 학습 및 구현을 수행하는 연구도 필요할 것으로 보인다.

본 연구는 기업신용위험관리 분야에 대한 RF의 적용 가능성을 검증하고, 앞서 설명한대로 많은 특징점을 가진 RF를 국내 경영학 분야의 연구자들에게 선도적으로 소개한다는 측면에서 의의가 있지만, 한계점 역시 다수 가지고 있다. 첫째, RF의 성능을 개선하기 위한 다양한 접근들이 추후 더 고려될 필요가 있다. 특히 RF의 성능은  $n_{tree}$ 나  $m_{try}$ 와 같은 설계모수 값들을 어떻게 최적화 할 것인가와 관련될 수 있는데, 이 중

ntree에 대한 모수값 최적화를 본 연구에서는 전혀 고려하지 못하였다. 향후 수행될 연구에서는 이와 같은 RF의 모수들에 대한 민감도 분석 및 최적화 시도가 고려될 필요가 있다.

둘째, 실증분석 과정이 보다 견고하게 재설계되어야 할 필요가 있다. 본 논문에서 수행된 실증분석에서는 RF의 입력변수로 다중판별분석의 단계별 선택을 통해 선정된 14개의 재무지표를 사용하였는데, 이는 비교모형과 동일한 입력변수를 제안모형에도 적용함으로써 모형 간 성능을 공정한 조건에서 비교한다는 차원에서는 장점이 될 수 있으나, 입력변수의 무작위 선택이라는 RF의 특징을 최대한 활용하는데 있어서는 치명적인 제약이 되고 있다. 또한 다수의 입력변수들을 고려할 경우, 특히 통계모형을 적용할 때에는 다중공선성 문제에 대해 좀 더 조심스럽게 접근할 필요가 있는데, 본 연구에서는 이와 같은 부분들을 고려하지 못하였다. 때문에 후속 연구에서 수행될 실증분석에서는 이러한 문제들이 충분히 개선되어야 할 것으로 생각된다.

셋째, 인공신경망이나 SVM과 같은 비교모형에 일반화 학습(generalized learning)의 원리를 적용해 보고, 이 때 산출된 성과를 이번 연구에서 도출된 RF의 성능과 비교해 볼 필요가 있다. RF는 의사결정나무 기법을 기반으로 한 일반화 학습의 원리가 적용되어 있는데, RF가 누리는 많은 특징점들이 사실상 이러한 일반화 학습에 기인한 것일 수도 있다. 따라서, 일반화 학습을 적용하지 않은 비교모형들의 성능을 RF와 비교하는 것은 애초에 공정한 비교가 아닐 수 있으며, 다른 분류모형에 일반화 학습의 원리를 적용했을 때 RF보다 더 우수한 성능을 보일 가능성을 배제할 수 없다. 때문에 후속 연구를 통해, 이러한 부분들을 보다 심도있게 검증하고 확인해 볼 필요가 있다.

넷째, 실증분석에 사용된 데이터가 다소 시간이 지난 데이터(2002년도)라는 점도 본 연구의 또 다른 한계점이다. 이는 우리가 확보 가능했던 신용등급 데이터가 해당 년도에 국한되어 있었기에 발생한 한계점이다. 추후 여건이 된다면, 가급적 기업들의 최근 신용등급 데이터와 재무지표 데이터를 확보하여, 최신의 패턴이 반영된 기업신용등급 예측 모형 구축에 RF가 얼마나 효과적으로 적용될 수 있는지 다시 한 번 검증할 필요가 있다.

다섯째, 보다 효과적인 기업신용등급 예측을 위해서는 분류 모형 자체의 성능을 개선하는 노력도 중요하지만, 그 밖에도 여러 추가적인 요인들이 함께 종합적으로 검토되어야 한다. 특히 기업신용등급 예측의 최근 연구동향을 살펴보면, 단일 등급 분류(one class classification), 다양성 및 정확성에 기반한 분류모형 선택(diversity and

accuracy-based classifier selection), 데이터 불균형(data imbalance) 문제 해소 등의 주제들이 부상하고 있는데, 본 연구에서는 이러한 이슈들이 고려되지 못했다. 또한 최근에 발표된 한 연구에서는 불균형한 분류 문제의 성능을 보다 합리적으로 평가하기 위한 지표로 등급별 평균 예측 정확도에 기하평균 정확도를 추가로 산출하는 방식을 제안하기도 하였다(Kim et al., 2015). 이에 향후 수행될 연구에서는 본 연구를 통해 제안된 RF 기법과 앞서 제시한 이슈들이 종합적으로 다루어지는 한층 더 고도화된 연구의 수행이 요구된다 하겠다.

## 참 고 문 헌

- 김문현(2013), “신용등급의 주가관련성 연구,” 산업혁신연구, 제29권, 제4호, pp. 171-192.
- 권안나(2013), “랜덤포레스트를 이용한 변수 선택,” 인하대학교 대학원 통계학과 석사학위 논문.
- 박기남, 이훈영, 박상국(2000), “리프집합을 이용한 통합형 채권등급 평가모형 구축에 관한 연구,” 한국경영과학회지, 제25권, 제3호, pp. 125-135.
- 심준섭(2004), “부스트래핑(bootstrapping) 기법을 활용한 회귀분석,” 정책분석평가학회보, 제14권, 제2호, pp. 167-183.
- 안현철, 김경재(2009), “Corporate Bond Rating using Various Multiclass Support Vector Machines,” *Asia Pacific Journal of Information Systems*, 제19권, 제2호, pp. 157-178.
- 안현철, 김경재, 한인구(2006), “다분류 Support Vector Machine을 이용한 한국 기업의 지능형 기업채권평가모형,” 경영학연구, 제35권, 제5호, pp. 1479-1496.
- 윤태균, 이관수(2008), “의료진단 및 중요 검사 항목 결정 지원 시스템을 위한 랜덤 포레스트 알고리즘 적용,” 전기학회논문지, 제57권, 제6호, pp. 1058-1062.
- 이건창, 한인구, 김명중(1996), “통계적 모형과 인공지능 모형을 결합한 기업신용평가 모형에 관한 연구,” 한국경영과학회지, 제21권, 제1호, pp. 81-100.
- 이규복, 하준경(2012), “한국의 경제성장과 금융의 역할,” 산업혁신연구, 제28권, 제2호, pp. 1-34.

산업혁신연구 (제32권 1호), 2016. 3. pp. 187~211.

이현주, 신동규, 박희원, 김수한, 신동일(2011), “부정맥 증상을 자동으로 판별하는 Random Forests 분류기의 정확도 향상을 위한 수정 알고리즘에 관한 연구,” 정보처리학회논문지 B, 제18-B권, 제6호, pp. 341-348.

최진배(2002), “외환 위기와 기업구조조정 - 현황과 문제점,” 산업혁신연구, 제18권, 제1호, pp. 21-49.

최혁진, 최성욱, 한경숙(2012), “Random forest를 이용한 단백질에서의 DNA 결합 부위 예측,” 정보과학회논문지: 소프트웨어 및 응용, 제39권, 제7호, pp. 515-522.

한은정(2004), “건강검진 자료에서 Random Forests를 이용한 백내장 발생 위험군 예측 모형,” 연세대학교 대학원 의학통계학전공 석사학위논문.

홍준혁, 고병철, 남재열(2013), “가중치 기반 Bag-of-Feature와 앙상블 결정 트리를 이용한 정지 영상에서의 인간 행동 인식,” 한국통신학회논문지, 제38A권, 제1호, pp. 1-9.

Altman, E. I., G. Marco, and F. Varetto(1994), “Corporate Distress Diagnosis Comparisons Using Linear Discriminant Analysis and Neural Networks,” *Journal of Banking and Finance*, Vol. 18, No. 3, pp. 505-529.

Amit, Y. and D. Geman(1997), “Shape Quantization and Recognition with Randomized Trees,” *Neural Computation*, Vol. 9, No. 7, pp. 1545-1588.

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

Breiman, L.(2001), “Random Forests,” *Machine Learning*, Vol. 45, pp. 5-32.

Brown, I. and C. Mues(2012), “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” *Expert Systems with Applications*, Vol. 39, No. 3, pp. 3446-3453.

Chaveesuk, R., C. Srivaree-Ratana, and A. E. Smith(1993), “Alternative neural network approaches to corporate bond rating,” *Journal of Engineering Valuation and Cost Analysis*, Vol. 2, No. 2, pp. 117-131.

Chang, C.-C. and C.-J. Lin(2011), “LIBSVM : a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27:1-27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (downloaded on 2015.07.14.).

De Andres, J., P. Lorca, F. Sanchez-Lasheras, and F. J. De Cos-Juez(2012), “Bankruptcy Prediction and Credit Scoring: A Review of Recent Developments,” *Recent*

- Patents on Computer Science*, Vol. 5, pp. 11-20.
- Efron, B.(1979), "Bootstrap methods: Another look at the Jackknife," *The Annals of Statistics*, Vol. 7, No. 1, pp. 1-26.
- Efron, B. and R. Tibshirani(1993), *An Introduction to the Bootstrap*, New York, Chapman and Hall.
- Hajek, P. and K. Michalak(2013), "Feature selection in corporate credit rating prediction," *Knowledge-Based Systems*, Vol. 51, pp. 72-84.
- Huang, C.C. and T.L. Tseng(2004), "Rough set approach to case-based reasoning application," *Expert Systems with Applications*, Vol. 26, No. 3, pp. 369-385.
- Huang, Z., H. Chen, C-J. Hsu, W-H. Chen, and S. Wu(2004), "Credit rating analysis with support vector machines and neural networks: a market comparative study," *Decision Support Systems*, Vol. 37, pp. 543-558.
- Jo. H. and I. Han(1996), "Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction," *Expert Systems with Applications*, Vol. 11, pp. 415-422.
- Kalsyte, Z. and A. Verikas(2013), "A novel approach to exploring company's financial soundness: Investor's perspective," *Expert Systems with Applications*, Vol. 40, No. 13, pp. 5085-5092.
- Kim, K.-j. and H. Ahn(2012), "A Corporate Credit Rating Model using Multi-class Support Vector Machines with an Ordinal Pairwise Partitioning Approach," *Computers & Operations Research*, Vol. 39, No. 8, pp. 1800-1811.
- Kim, M.-J., D.-K. Kang, and H. B. Kim(2015), "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Systems with Applications*, Vol. 42, No. 3, pp. 1074-1082.
- Kim, J. W., H. R. Weistroffer, and R. T. Redmond(1993), "Expert systems for bond rating: a comparative analysis of statistical rule-based and neural network systems," *Expert Systems*, Vol. 10, No. 3, pp. 167-171.
- Kumar, P. R. and V. Ravi(2007), "Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review," *European Journal of Operational Research*, Vol. 180, No. 1, pp. 1-28.

산업혁신연구 (제32권 1호), 2016. 3. pp. 187~211.

- Kwon. Y. S., I. G. Han, and K. C. Lee(1997), “Ordinal Pairwise Partitioning(OPP) approach to neural networks training in bond rating,” *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, pp. 23-40.
- Lee. K., I. Han, and Y. Kwon(1996), “Hybrid neural networks for bankruptcy predictions,” *Decision Support Systems*, Vol. 18, pp. 63-72.
- Maher, J. J. and T. K. Sen(1997), “Predicting bond ratings using neural networks: a comparison with logistic regression,” *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, pp. 59-72.
- Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers(2011), *Big Data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute.
- Shin, K.-S. and I. Han(1999), “Case-based reasoning supported by genetic algorithms for corporate bond rating,” *Expert Systems with Applications*, Vol. 16, No. 2, pp. 85-95.
- Shin, K.-S. and I. Han(2001), “Case-based approach using inductive indexing for corporate bond rating,” *Decision Support Systems*, Vol. 32, No. 1, pp. 41-52.
- Witten, I. H., E. Frank, and M. A. Hall(2011), *Data Mining: Practical Machine Learning Tools and Techniques(3rd Ed)*, San Francisco, Morgan Kaufmann Publishers Inc.

05월/04일/2015 접수, 07월/20일/2015 수정, 11월/26일/2015 게재확정

## Application of Random Forests to Corporate Credit Rating Prediction

*Seong-jin, Kim<sup>1</sup> · Hyun-chul, Ahn<sup>2</sup>*

### Abstract

There has been a lot of research on intelligent information processing techniques to effectively and efficiently manage credit risk in the financial industry. To predict corporate credit ratings better, various algorithms of machine learning have been proposed and applied. However, the algorithms have some limitations. First of all, they are unable to give a solution for which variables in the dataset should be considered as input variables to effectively predict a dependent variable. Also, they are easily exposed to overfitting and risk being affected by outliers and noise. In addition, they are required to tune their numerous parameters.

Under this background, this study proposes the application of Random Forests(RF) to overcome the weakness of the traditional techniques. RF were designed by Leo Breiman, an American statistician and professor, and are an ensemble learning technique. This algorithm has both features originating from bagging algorithm, which is to extend a single decision tree to various decision trees, and from random input selection. The benefit of our proposed algorithm is free from an overfitting issue, and both noise and outliers in the dataset cannot have an effect on it. Thus, our model can yield high accuracy in forecasting corporate credit ratings.

To validate the effectiveness of RF in predicting corporate credit ratings, we

---

1 Master's Candidate. Graduate School of Business IT, Kookmin University. 77, Jeongneung-ro, Seongbuk-gu, Seoul, 136-702, South Korea tel: 02-910-4018, e-mail: scott\_kim@kookmin.ac.kr (First Author)

2 Associate Professor. Graduate School of Business IT, Kookmin University. 77, Jeongneung-ro, Seongbuk-gu, Seoul, 136-702, South Korea tel: 02-910-4577, e-mail: hcahn@kookmin.ac.kr (Corresponding Author)

산업혁신연구 (제32권 1호), 2016. 3. pp. 187~211.

applied it to a real world dataset of corporate credit rating in Korea. The data that we used for this research consisted of various financial ratios of 1,295 corporations and their corporate credit ratings. To evaluate and compare our model with traditional methods, we used the following methods: MDA(Multiple Discriminant Analysis), ANN(Artificial Neural Network) and MSVM(Multiclass SVM). Each model's performance was calculated by predictive accuracy based on training and validation. Empirical results showed that RF outperforms other traditional methods from the perspective of prediction accuracy.

Keywords: Random Forests, Credit Risk Management, Corporate Credit Ratings, Multiclass Classification, Data Mining.



저자소개



김성진

최종학력: 국민대학교 정보미디어경영 이학석사  
관심분야: 데이터마이닝, 경영정보시스템, 정보미디어



안현철

현재 국민대학교 비즈니스IT전문대학원 부교수  
최종학력: KAIST 테크노경영대학원 경영공학 박사  
관심분야: 지능형 정보기술, 고객관계관리, 정보시스템 수용 등

