지능정보연구 제16권 제4호 2010년 12월(pp.99~112)

유전자 알고리즘을 이용한 분류자 앙상블의 최적 선택

김명종
부산대학교 경영학과
(mjongkim@pusan.ac.kr)

앙상블 학습은 분류 및 예측 알고리즘의 성과개선을 위하여 제안된 기계학습 기법이다. 그러나 앙상블 학습은 기저 분류자의 다양성이 부족한 경우 다중공선성 문제로 인하여 성과개선 효과가 미약하고 심지어는 성과가 악화될 수 있다는 문제점이 제기되었다. 본 연구에서는 기저 분류자의 다양성을 확보하고 앙상블 학습의 성과개선 효과를 제고하기 위하여 유전자 알고리즘 기반의 범위 최적화 기법을 제안하고자 한다. 본 연구에서 제안된 최적화 기법을 기업부실예측 인공신경망 앙상블에 적용한 결과 기저 분류자의 다양성이 확보되고 인공신경망 앙상블의 성과가 유의적으로 개선되었음을 보여주었다.

논문접수일 : 2010년 11월 22일 게재확정일 : 2010년 12월 07일 교신저자 : 김명종

1. 서 론

앙상블 학습은 분류 및 예측 알고리즘의 성과개선을 위하여 제안된 기계학습 방법이다. 앙상블 학습은 기저 분류자(base classifiers) 집합인 앙상블을 구성하고, 앙상블에서 추론된 복수의 학습결과는 단일 강분류자(a single strong classifier)의 결합과정을 통하여 하나의 최종 결과를 도출한다. 즉, 앙상블 학습은 학습 데이터를 정확하게 표현할 수 있는 하나의 가설을 선택하는 것이 아니라 가설들의 집합을 구성하고 새로운 데이터를 예측할 때에는 그 가설들의 예측을 결합하여 최종 결정을 내리게 된다. 앙상블 학습은 탁월한 성과개선 효과와 더불어 의사결정트리(Decision Tree : DT), 인공신경망(Neural Networks : NN) 및 Support Vector Machines(SVM) 등 다양한 학습알고리즘과 유연한 결합이 가능하다는 장점으로 인하여 기계학습 및 인공지능분야에서 큰 관심을 받아 왔다.

특히 이론 및 실증 연구들은 앙상블 학습이 CART (Classification and Regression Tree)와 C4.5 등 DT 기반의 학습 알고리즘의 예측력을 크게 개선할 수 있음을 보여주었다(Freund and Schapire, 1997 Drucker and Cortes, 1993; Quinlan, 1996; Bauer and Kohavi, 1999; Macline and Optiz, 1997). 최근 기업부실 예측문제를 대상으로 DT 앙상블 학습을 적용한 결과 DT 앙상블 학습이 NN에 비하여 오분류율을 약 30% 감소시키는 등 탁월한 예측력을 보유하고 있음을 보여주었다(Alfrao, Gámez, and García, 2007; Alfrao, García, Gámez, and Elizondo, 2008).

* 이 논문은 2010년도 부산대학교 특성화분야 육성사업의 연구비를 지원받아 연구되었음.

현재, DT 앙상블의 탁월한성과를 기초로 최근 NN 앙상블 학습 및 SVM 앙상블 학습 등에 대한 연구가 활발하게 진행되고 있다. 많은 실증 연구에서 앙상블 학습은 NN 및 SVM의 성과개선에 효과적임을 보여주었다. 최근 NN 앙상블 기법을 이용하여 기업부실예측에 적용한 연구에서도 NN 앙상블 학습전략이 부실예측모형의 정확성을 유의적으로 향상시킬 수 있음을 보여주었다(Kim and Kang, 2010). 그러나 NN 앙상블 학습 및 SVM 앙상블 학습은 DT 앙상블 학습과 비교하여 성과개선 효과가 미약하고 심지어 일부 실증연구에서는 앙상블 학습이 단일 분류자와 비교하여 예측력이 떨어지는 문제점이 보고되었다(Valentini et al., 2003; Buciu et al., 2001; Evgeniou et al., 2000; Dong and Han, 2004). 이러한 성과 저하는 기저 분류자의 다양성이 부족하고 분류자 사이에 높은 상관관계가 존재하는 경우에 나타나는 다중공선성 문제에 기인하고 있다. 따라서 분류자의 다양성을 확보하고 앙상블 학습의 성과를 개선하기 위한 차별화된 학습전략에 대한 필요성이 제기되어 왔다(Hansen and Salamon, 1990; Brieman, 1996; Kim, 2009).

본 연구에서는 인공신경망 앙상블을 대상으로 분류자의 다양성을 확보하고 앙상블 학습의 성과개선을 위하여 범위 최적화 기법(Coverage Optimization of NN Ensemble : CO-NN)을 제안하고자 한다. 범위 최적화는 앙상블 학습의 결과로서 생성된 기저 분류자 앙상블 집합에서 다양하고 성과개선에 효과적인 최적 또는 최적에 가까운 하위 앙상블을 탐색하는 문제로 정의할 수 있다. CO-NN은 이러한 하위 앙상블을 탐색하기 위하여 유전자 알고리즘(Genetic Algorithms : GA)을 활용하고 있다. 본 연구에서는 CO-NN 기법을 기업부실 예측문제에 적용하여 성과를 검증하고자 한다.

본 연구는 다음과 같이 구성되어 있다. 제 2장에서는 앙상블 학습알고리즘 및 효과적인 앙상블구성전략에 대하여 설명하고자한다. 제 3장에서는 본 연구에서 제안하고 있는 CO-NN의 알고리즘에 대하여 설명하고자 한다. 제 4장에서는 제안 모형의 유용성을 확인하기 위한 기업부실예측 데이터 및 변수 선정에 대하여 설명하고자 한다. 제 5장에서는 CO-NN의 실험결과를 종합적으로 정리하여 제시하고자 한다. 마지막 장에서는 결론과 함께 향후 연구 과제를 제시하고자 한다.

2. 앙상블 학습

본 장에서는 앙상블 학습의 대표적 학습알고리즘인 배깅 알고리즘과 AdaBoost 알고리즘에 대하여 설명하고 효과적인 앙상블 구성전략에 대하여 기술하고자 한다.

2.1 앙상블 학습

앙상블 학습은 복수의 기저 분류자로 구성된 앙상블 집합을 생성하여 단일 분류자의 정확성을 개선하기 위한 목적으로 제안되었다. 앙상블 학습을 하는 이유는 단일 분류자만으로는 학습이 어려운 복잡한 패턴들을 여러 개의 하위 패턴으로 나누어 효과적인 학습을 진행할 수 있으며 여러 하위 패턴에 대한 학습 결과를 결합하여 단일 분류자보다 정확한 예측 결과를 제공하기 때문이다.

앙상블 학습 중 가장 보편적으로 사용되고 있는 알고리즘은 배깅 알고리즘(Breiman, 1996)과 AdaBoost 알고리즘(Freund and Schapire, 1997)이다. 배깅 알고리즘은 n개의 원시 데이터에서 복원추출 방법으로 K개의 붓스트랩(bootstrap) 분석용 데이터를 생성하고 각 붓스트랩 분석용 데이터에 분류

알고리즘을 적용하여 K개의 분류자를 구성한다. i번째 관측치에 대한 최종 결과는 강분류자인 배깅 분류자의 결합함수 α_k를 이용하여 다음과 같이 산출된다. 여기에서 x_i는 i번째 관측치의 예측변수 벡터이며 $C_k(x_i)$는 예측변수 벡터 x_i에 대한 k번째 분류자의 학습결과이며 α_k는 k번째 분류자에 부여되는 결합 가중치로 분류자의 신뢰도(또는 정확성)를 의미한다.

$$C(x_i) = sign\left(\sum_{k=1}^{k} \alpha_k C_k(x_i)\right)$$

AdaBoost 알고리즘은 가장 보편적으로 사용되는 부스팅 알고리즘으로 Freund and Schapire (1997)에 의하여 제안되었다. n개의 학습 표본과 K개의 기저 분류자로 구성된 앙상블 $C=\{C_1, C_2, \cdots, C_k\}$을 가정하면 k번째 분류자의 오류율(e_k)은 다음과 같이 단순평균으로 계산된다.

$$e_k \frac{1}{n}\sum_{i=1}^{n} L(C_k(x_i), y_i)$$

$$L(C_k(x_i), y_i) = \begin{cases} 1 & C_k(x_i) \neq y_i \\ 0 & C_k(x_i) = y_i \end{cases}$$

여기에서 x_i는 i번째 관측치의 예측변수 벡터이고 y_i는 i번째 관측치의 범주를 나타내며 $C_k(x_i)$는 예측변수 벡터 x_i에 대한 k번째 분류자의 분류결과이다. $k+1$번째 분류자에서 i번째 관측치에 부여되는 가중치는 $w_{k+1}(i) = w_k(i)\exp(\alpha_k L(C_k(x_i), y_i)$로 조정되어 오분류된 관측치에 더 높은 가중치가 부여된다. 여기에서 k는 $\alpha_k = \ln((1-e_k)/e_k)$로 계산된다. 이러한 과정을 통하여 오분류된 관측치에 높은 가중치가 부여되며 $k+1$번째 분류자의 학습 표본을 구성할 때 가중치가 높은 오분류 관측치가 포함될 가능성이 높기 때문에 부스팅 알고리즘은 오분류 관측치에 초점을 맞춘 학습을 진행할 수 있게 된다. i번째 관측치의 최종 결과는 강분류자인 부스팅 분류자를 통하여 다음과 같이 계산된다. 여기에서 α_k는 배깅 알고리즘과 마찬가지로 k번째 분류자에 부여되는 결합 가중치로 분류자의 신뢰도(또는 정확성)를 의미한다.

$$C(x_i) = sign\left(\sum_{k=1}^{K} \alpha_k C_k(x_i)\right)$$

2.2 앙상블 구성전략

많은 실증연구에서 앙상블 학습은 CART(Classification and Regression Tree)와 C4.5 등 다양한 종류의 DT 알고리즘과 결합하여 DT와 같은 불안정한 학습 알고리즘의 예측력을 크게 개선시킬 수 있음을 보여주었다(Freund and Schapire, 1997; Drucker and Cortes, 1993; Quinlan, 1996; Bauer and Kohavi, 1999; Macline and Optiz, 1997). 이러한 연구를 기초로 앙상블 학습과 관련된 대부분의 연구는 DT를 기저 분류자로 활용한 성과개선에 초점을 맞추고 있다.

한편, NN 앙상블 학습 및 SVM 앙상블 학습과 관련된 많은 연구에서도 앙상블 학습은 NN 및 SVM의 성과개선에 효과적임을 보여주고 있다. 그러나 NN 및 SVM 앙상블 학습은 DT 앙상블 학습과 비교하여 성과개선 효과가 미약하고 오히려 앙상블 학습의 성과가 단일 분류자의 성과보다 떨어진다는 연구 결과도 보고되고 있다(Valentini et al., 2003; Buciu et al., 2001; Evgeniou et al., 2000; Dong and Han, 2004). 이러한 결과는 분류자 사이에 높은 상관관계가 존재하는 경우 다중공선성 문제가 발생하게 되며 결과적으로 앙상블 학습의 판별함수를 왜곡시킴으로써 성과가 저하되기 때문

이다(Hansen and Salamon, 1990; Brieman, 1996; Kim, 2009).

Hansen and Salamon(1990)은 앙상블 학습이 단일 분류자와 비교하여 높은 성과를 나타내는 이유를 분석하였다. 앙상블의 분류 결과가 상호 독립적이고 정확성 수준을 p라 하면 K개의 분류자가 구성된 앙상블의 분류 정확성은 다음과 같이 계산된다.

$$\sum_{k>K/2}^{K}\binom{K}{k}p^k(1-p)^{K-k}$$

이진분류 문제에 있어 만일 p가 0.5보다 크다면 K가 증가함에 따라 앙상블의 분류 정확도가 높아지며 결과적으로 단일 분류자의 정확도보다 높아지게 된다. 이와 같은 방법으로 앙상블 학습은 분류자의 정확성 향상과 더불어 일반화 성능도 개선할 수 있다

Hansen and Salamon(1990)은 성과 개선을 위한 앙상블의 필요충분조건으로 분류자들이 임의 추측보다 정확해야 하고 다양하게 구성되어야 함을 주장하였다. 이는 분류자의 정확도가 50% 이상이며 오분류 패턴이 서로 다른 다양한 분류자로 앙상블을 구성하였을 때 성과가 크게 개선될 수 있음을 의미한다. 예를 들어, 다양한 분류자로 구성된 앙상블은 특정 분류자가 부정확한 예측을 하더라도 다른 분류자들이 정확히 예측할 수 있다면 최종 결과에서 정확하게 분류될 수도 있다. 그러나 분류자의 다양성이 부족한 경우, 특정 분류자가 부정확하게 예측한 관측치에 대하여 나머지 분류자들도 부정확하게 예측할 가능성이 높기 때문에 최종 결과 역시 부정확하게 된다. 결과적으로 Hansen and Salamon은 성과저하의 원인을 분류자의 다양성이 부족하고 분류자 사이에 높은 상관관계가 존재하는 경우에 나타나는 것으로 설명하고 있다.

Brieman(1996)은 배깅과 같은 앙상블 학습은 불안정한 학습 알고리즘의 성능을 크게 향상시킬 수 있는 반면에 안정적인 학습 알고리즘에 대해서는 뚜렷한 성능 향상이 관찰되지 않는다고 주장하였다. 앙상블 학습은 분류자 풀(classifier pool)의 다양성을 확보하기 위한 학습 전략으로 배깅, 부스팅 및 random forest 등 다양한 샘플링 방법을 이용하고 있다. 의사결정트리와 같은 불안정한 학습 알고리즘은 학습 데이터의 변화에 민감하게 반응하여, 작은 데이터의 변화에도 의사결정트리 구조가 변경되는 등 생성되는 분류자에도 큰 변화가 나타나게 되며 결과적으로 분류자의 다양성을 확보할 수 있다.

반면, NN 및 SVM과 같은 안정적인 학습 알고리즘은 학습 데이터가 변하더라도 유사한 분류자를 반복적으로 생성하므로 분류자 사이의 상관관계가 높아지게 된다. 분류자 사이의 높은 상관관계로 인하여 다중공선성 문제가 발생되며, 결과적으로 앙상블 학습의 성과가 저하될 수 있다고 분석하였다.

최근 국내기업의 부실예측문제를 대상으로 학습 알고리즘에 대한 성과비교를 수행한 Kim (2009)의 연구에서는 단일 분류자의 경우 NN 및 SVM이 DT보다 유의적으로 높은 성과를 보여주었다. 하지만, 앙상블 학습의 경우 DT 앙상블의 예측력이 NN 앙상블 및 SVM 앙상블의 예측력과 비교하여 유의적으로 높은 것으로 나타났다. 추가적으로 수행된 분산팽창요인(Variance Inflation Factior : VIF) 분석을 통하여 앙상블 학습의 성과저하의 원인이 다중공선성에 기인하고 있으며 이를 해결하기 위한 차별화된 학습전략, 즉 최적화 전략이 필요하다고 주장하였다(Kim, 2009).

3. 유전자 알고리즘 기반의 범위 최적화

Ho(2002)는 앙상블이 구성되었음을 전제로 분

류자의 선택 문제와 분류자 결과의 결합 문제를 각각 범위 최적화(Coverage optimization)와 결정 최적화(decision optimization)로 정의하였다. 범위 최적화에서 선택된 앙상블을 대상으로 결정 최적화가 적용되기 때문에 범위 최적화는 최적화 문제의 가장 근간이 되는 문제로 간주되고 있다. 하지만 아쉽게도 현재 결정 최적화 문제와 관련하여 다양한 연구가 진행되고 있는 반면, 범위 최적화에 대한 연구는 상대적으로 적게 보고되고 있다.

범위 최적화 문제는 K개의 분류자를 갖는 앙상블에서 $d(K \geq d)$개의 분류자를 갖는 하위 앙상블을 선택하는 문제로 성과개선을 위한 최적의 하위 앙상블을 구성하는 것을 목적으로 한다. 이 때 가능한 탐색공간의 크기는 ${}_{K}C_{d}$로 지수적 탐색 공간(exponential search space)을 가지게 된다. 최근 최적해 탐색을 위하여 주로 활용되는 기법은 GA이다. GA는 교배(cross-over)나 변이 연산(mutation)을 통해 국부 최적해에 빠지는 것을 방지할 수 있으며, 방대하고 복잡한 공간을 탐색하면서 최적 혹은 최적에 가장 가까운 결과를 찾아주는 확률적 검색 방법을 이용하여 빠른 탐색이 가능하다는 장점으로 인하여 방대한 탐색공간에서 최적해 탐색기법으로 자주 활용되고 있다.

Zhou et al.(2002)의 연구는 앙상블에 포함된 모든 분류자를 사용하는 것 보다 크기가 작은 하위 앙상블을 선택하는 것이 성능 측면에서 우수할 수 있다는 점을 이론적으로 증명하였다. 또한 수많은 분류자를 가진 신경망 앙상블을 생성한 후 하위 앙상블을 선택하는 문제에 GA를 적용하여 그들의 이론을 실험적으로 증명하였다. Oliverira et al.(2003)의 연구에서도 다수의 분류자로 구성된 앙상블에서 하위 분류자 앙상블을 GA를 이용하여 선택하였다. Kim and Oh(2008)는 단순 유전자 알고리즘과 혼합형유전자 알고리즘을 제시하고 혼합형 유전자 알고리즘이 탐색 능력에서 뛰어남을 보여주었다.

그러나 이러한 기존 접근방법은 성과개선의 측면에서예측력(정확성)만을 고려하였을 뿐 정확성과 다양성을 동시에 고려하지 못했다는 문제점이 있다. 결과적으로 본 연구에서 제안하는 범위 최적화 기법은 성과개선의 효과와 더불어 분류자의 다양성을 동시에 확보할 수 있다는 장점이 있다. 본 연구에서 제안한 CO-NN은 다음과 같은 3단계 과정을 통하여 구성된다.

1단계 : 모집단 설정

범위 최적화 문제의 해결을 위해서 GA가 인식할 수 있도록 염색체(Chromosome) 형태로 코드화되어야 한다. 본 연구와 같은 범위 최적화 문제에서 탐색공간은 앙상블 분류자 수인 K가 되며 최적화 대상인 각 분류자에게 부여되는 가중치(d_k)는 각각 0 또는 1의 이진열(binary string)로 코드화할 수 있다. 여기에서 이진값 0은 기저 분류자가 하위 앙상블에서 제외됨을 의미하며 이진값 1은 기저 분류자가 하위 앙상블에 포함되었음을 의미한다. <그림 1>은 이진열로 표현된 염색체의 예를 보여주고 있다. <그림 1>의 경우 이진열 (1010100011)의 10비트 코드는 기저분류자 앙상블 집합이 총 10개의 분류자로 구성되었음을 나타낸다. 또한 10개의 분류자 중 이진값 1로 코드화된 1번, 3번, 5번, 9번, 10번의 5개 분류자가 범위 최적화를 통하여 하위 앙상블에 포함되었음을 나타내고 있다. 이러한 이진열 코딩 방법을 이용하여 표현된 모집단(population)은 사전에 설정된 개체수 만큼 무작위

Classifier	1	2	3	4	5	6	7	8	9	10
Coding	1	0	1	0	1	0	0	0	1	1

<그림 1> Binary String Coding의 예

로 난수(Random number)를 발생시켜 생성한다.

2단계 : 정확성과 다양성을 고려한 적합도 함수

i번째 관측치에 대한 최종 결과는 범위 최적화를 통하여 선택된 하위 앙상블의 결과를 단일 강 분류자를 통하여 다수결 투표(Marjority Voting)에 의하여 1차적으로 결정되나, 동일 투표수인 경우 가장 높은 신뢰도(정확성) α_k를 가진 분류자의 결과를 최종 결과로 선택하였다. <그림 2>는 이러한 다수결 투표방법을 이용한 최종 결론 도출과정을 보여주고 있다. <그림 2>의 경우 3개 분류자(1번, 3번, 10번)의 결과는 1(정상기업)이며 2개 분류자(5번, 9번) 분류자의 결과는 0(부실기업)으로 다수결투표의 결과로서 최종 결론은 클래스 1(정상기업)로 결정된다.

Classifier	1	2	3	4	5	6	7	8	9	10
Coding	1	0	1	0	1	0	0	0	1	1
Output	1		1		0				0	1

<그림 2> 최종 결론 도출 과정의 예

범위 최적화의 목적은 예측오류가 낮은(또는 예측성과가 높은) 하위 앙상블 집합을 선택하는 것이므로 본 연구에서는 학습용 데이터에 대한 평균 오류율을 적합도 함수로 설정하였다. 앙상블 학습에서 도출된 최종 결론의 오류율(E)은 다음과 같이 계산된다.

$$E=\frac{1}{n}\sum_{i=1}^{n}L(C(x_i),y_i$$

$$L(C(x_i),y_i=\begin{cases}1 & L(C(x_i)\neq y_i\\ 0 & L(C(x_i)=y_i\end{cases}$$

하위 앙상블을 구성 후 평균 오류율은 위의 평균 오류율 계산과 동일한 방법으로 계산되며 이를 E_S라고 할 때 E-E_s 0은 하위 앙상블의 오류율이 보다 낮음을 의미하며 이는 하위 앙상블의 예측 정확성이 개선되었음을 의미한다. 결과적으로 본 연구에서는 E-E_s를 최대화하는 것으로서 적합도 함수를 설정하였다.

성과측면의 적합도 함수 설정과 더불어 범위 최적화의 또 다른 목적인 분류자의 다양성을 확보하기 위하여 다중공선성 측정과 관련된 제약조건을 추가하였다. 다중공선성 측정을 위하여 자주 활용되는 측정 방법은 VIF 분석이다. VIF는 회귀계수 추정치의 분산이 이상적인 경우(서로 직교하는 경우)에 비해 얼마만큼 계수 추정치들의 분산이 팽창하는가를 알려주는 척도이며, k번째 분류자의 VIF는 다음과 같이 계산된다.

$$VIF(k)\frac{1}{1-R_k^2}$$

여기서 R_k^2 는 k번째 분류자의 결과를 종속변수로, 나머지 분류자의 결과를 독립변수로 가정한 회귀분석의 결정계수(R^2)이다. k번째 분류자가 다른 분류자들과 밀접한 관계에 있으면 R_k^2는 1에 가깝게 되고 VIF(k)의 값이 커지게 된다. 만일 5 < VIF(k) < 10이면 k번째 분류자의 다중공선성 가능성이 있으며, VIF(k) < 10이면 k번째 분류자의 다중공선성이 심각함을 의미한다. 따라서 하위 앙상블에 포함되는 분류자 사이의 다중공선성 문제가 없을 경우 VIF 값은 5보다 작아야 한다. 이러한 다중공선성 제약조건을 포함한 적합도 함수는 다음과 같은 수식으로 나타낼 수 있다.

$$\text{Fitness}: Max(E-E_s)$$
$$\text{Subject to}: VIF_k<5$$

3단계 : 유전자 연산

이 단계에서, GA는 초기 설정된 염색체에 대해 교배(Crossover), 돌연변이(mutation) 등 다양한 탐색과정을 적용하여 새로운 후보들을 생성한다. 각 후보들의 적합도 점수를 산출하여 적합도가 우수한 후보들은 적합도가 낮은 모집단의 개체(organisms)들을 대체한다. 이러한 작업은 중지조건(stopping condition)이 만족될 때까지 반복적으로 수행된다.

4. 자료 분석 및 가설검증연구 설계

4.1 표본의 특성

부실기업은 2002~2005년 중 은행연합회 신용정보등록, 당좌부도발생, 회사정리절차개시, 기업구조조정절차개시 사유에 해당하는 600개 외부감사 제조 기업으로 구성하였으며 부실기업에 대응되는 정상 기업은 외부감사 제조기업 중 2005년 말 기준으로 부실 사유에 해당하지 않는 600개의 기업으로 구성하였다.

부실 예측에 사용되는 재무비율은 일차적으로 기존의 기업부실 예측연구에 사용된 비율 및 실무에서 부실예측의 지표로 유용하게 활용되는 비율을 중심으로 31개의 재무비율을 수집하였다. 수집된 재무비율을 수익성, 부채상환능력, 레버지리, 자본구조, 유동성, 활동성 및 규모의 7개 재무비율군으로 재분류하였으며 7개 군별로 최종변수는 ROC 분석을 이용하여 선정하였다. ROC 분석은 분류자 결과 값의 서열화 순위에 따라 수평축에는 1-특이도, 수직 축에는 민감도를 표시하여 연결한 ROC 곡선에 기초하여 분류자의 정확성을 분석하는 방법이다. 여기에서 민감도와 특이도는 <표 1>의 오분류표(Confusion matrix)에서 각각 TP/(TP+FN))과 TN/(FP+TN)으로 측정된다. 분류자의 정확도는 ROC 곡선의 면적 (Area Under ROC : AUROC)로 계산된다. 완벽한 모형의 AUROC는 1이 되며, 임의추측 모형의 AUROC는 0.5가 된다. 대부분의 모형은 일반적으로 0.5보다 크고 1보다 작은 AUROC를 가지며 AUROC가 1에 근접할수록 정확도가 높은 모형으로 평가된다(Fawcett, 2006).

<표 1> 오분류표

		예측 범주	
		부실 (positive)	정상 (Negative)
실제 범주	부실 (Positive)	True Positive(TP)	False Negative(FN)
	정상 (Negative)	False Positive(FP)	True Negative(TN)

ROC 곡선에 의하여 산출된 AUROC를 이용하여 각 분류군별로 AUROC가 높은 7개 재무비율을 선정하였으며 ROC 곡선을 이용한 AUROC는 <표 2>에 제시되어 있다.

7개 재무비율 사이의 다중공선성을 확인하기 위하여 VIF 분석을 실시한 결과는 <표 3>에 제시되어 있으며 최종 선정된 7개 재무비율 간에는 다중공선성이 실질적으로 존재하지 않음을 확인하였다.

4.2 실험설계

본 연구에서는 범위 최적화 성과를 검증하기 기업부실예측을 위한 DT 앙상블 모형과 NN 앙상블 모형을 구성하였다. DT 앙상블의 기저 분류자로 C4.5를 활용하였으며 NN 앙상블의 기저 분류자로 다층 퍼셉트론(Multi-Layer Perceptron : MLP) NN을 활용하였다.

CO-NN의 구성을 위한 GA의 파라미터들로서 모집단(population)은 20개체로 설정하였고, 교배 및 돌연변이 비율은 각각 0.5~0.7 및 0.06~0.1의

범위로 설정하였다. 또한 중지조건으로 1000회를 설정하여 50세대 만큼 하위 앙상블 탐색과 결합함수 탐색을 반복하도록 설정하였다.

<표 2> 재무비율의 AUROC(* 최종 7개 재무비율)

분류군	재무비율	AUROC
수익성	총자산경상이익율*	52.5
	총자산순이익율	45.9
	금융비용/매출액	49.7
	금융비용/총부채	48.9
	순금융비용/매출액	50.8
	매출액경상이익율	45.9
	매출액순이익율	49.9
	자기자본경상이익율	48.8
	자기자본순이익율	48.1
부채상환	EBITDA/이자비용*	53.7
	EBIT/이자비용	40.1
	영업현금흐름/이자비용	48.9
	영업현금흐름/총부채	48.8
	잉여현금흐름/이자비용	52.3
	잉여현금흐름/총부채	53.1
	부채상환계수	51.7
	차입금/이자비용	53.4
레버리지	자기자본비율*	51.9
	유동자산/총자산	50.9
자본구조	이익잉여금/총자산*	53.5
	이익잉여금/총부채	52.7
	이익잉여금/유동자산	51.1
유동성	현금비율*	46.5
	당좌비율	45.5
	유동비율	43.2
활동성	재고자산회전율*	30.8
	유동부채회전율	29.2
	매출채권회전율	27.7
규모	총자산*	24.8
	매출액	22.4
	고정자산	22.6

<표 3> 분산팽창요인 분석의 결과

재무비율	VIF
총자산경상이익율	1.64
EBITDA/이자비용	2.34
자기자본비율	1.95
이익잉여금/총자산	2.77
현금비율	1.54
재고자산회전율	1.73
총자산	1.67

5. 연구 결과

분류자의 성과검증을 위하여 전체 표본의 70%를 학습용 표본으로 30%를 검증용 표본으로 분류하du 실험을 진행하였다. <표 4>는 DT 앙상블, NN 앙상블, CO-NN에 대한 VIF 분석 결과로 실질적인 다중공선성 이 존재하는 분류자 (VIF > 5)를 보여주고 있다. DT 앙상블은 분류자 사이에 다중공선성 문제가 나타나지 않았지만, NN 부스팅 분류자 및 NN 배깅 분류자의 경우 각각 4개 및 5개의 분류자 사이에 다중공선성이 존재하고 있음을 보여준다. 한편 범위 최적화를 통하여 부스팅 CO-NN의 경우 9개의 하위 앙상블이 선택되었고 배깅 CO-NN의 경우 12개의 하위 앙상블이 선택되었다. CO-NN에 대한 VIF 분석 결과는 다중공선성 문제가 존재하는 분류자가 하위 앙상블에서 제외되었으며 결과적으로 CO-NN이 다양한 분류자로 구성되었음을 보여주고 있다.

<표 4> 분류자 사이의 다중공선성 분석

분류자	DT		NN		CO-NN	
	부스팅	배깅	부스팅	배깅	부스팅	배깅
총 분류자	15	20	14	20	9	12
VIF > 5	-	-	4	5		

<표 5>는 각 분류자의 검증용 표본에 대한 예측성과로 모든 분류자에서 앙상블 학습을 통하여 단일 분류자의 성과가 개선되었음을 보여주고 있다. DT 앙상블의 경우 5.28%~5.56%, NN 앙상블의 경우 2.22%~2.5%의 성과개선 효과가 나타났다. CO-NN의 경우 단일 NN 분류자와 비교하여 6.11~6.39%로 성과개선 효과가 가장 크게 나타났다. 단일 분류자와 앙상블 분류자의 성과에 유의적인 차이가 있는지 검증하기 위하여 카크란의 Q 검

정(Cochran's Q test)을 수행하였다. 분석 결과 DT 앙상블은 단일 DT 분류자보다 1% 수준이하에서 유의적인 성과 차이가 나타났고 NN 앙상블은 단일 NN 분류자와 비교하여 5% 수준이하에서 유의적인 성과 차이를 가진 것으로 분석되었다. 결과적으로 NN 앙상블 학습을 통하여 단일 NN 분류자의 성과를 개선하였으나, 성과개선 효과가 DT 앙상블에 비하여 미약한 것으로 나타났다. 하지만 CO-NN의 경우 단일 분류자와 비교하여 1% 수준이하에서 가장 유의적인 성과차이를 보여주고 있다.

<표 5> 단일 분류자와 앙상블 학습의 예측 성과비교(%)

분류자	단일	부스팅	배깅
DT	70.00	75.28*	75.56*
NN	71.39	73.61**	73.89**
CO-NN		77.50*	77.78*

* 1% 유의수준 ** 5% 유의수준

각 학습알고리즘에 따른 성과차이가 있는지 분석하기 위하여 맥네마 검정(McNemar test)을 수행하였으며 그 분석 결과는 <표 6>에 제시되어 있다. 단일 분류자의 경우 DT와 NN 사이에는 유의적인 성과 차이가 없는 것으로 분석되었다. 앙상블 학습의 경우 DT 앙상블의 학습성과가 NN 앙상블의 학습성과와 비교하여 5% 수준에서 유의적인 차이가 있는 것으로 분석되었으며, DT 앙상블과 CO-NN의 성과도 5% 수준에서 유의적인 차이가 있는 것으로 나타났다. 한편 CO-NN과 NN 앙상블 학습의 경우 1% 수준에서 유의적인 차이가 발생하여 범위 최적화는 성과개선 효과가 탁월한 것으로 나타났다.

<표 6> 예측성과에 대한 맥네마 검증 결과 (카이제곱값)

단일		부스팅			배깅		
	NN		NN	CO-NN		NN	CO-NN
DT	87.69	DT	-126.52**	129.51**	DT	-125.25**	129.09**
		NN		139.62*	NN		133.57*

* 1% 유의수준 ** 5% 유의수준.

이상의 실험 결과를 정리하면, 첫째, CO-NN은 GA를 이용한 하위 앙상블의 분류자 선택과정에서 다중공선성의 원인이 되는 분류자를 제거함으로써 분류자의 다양성을 확보할 수 있다. 둘째, 범위 최적화를 통하여 성과개선에 효과적인 분류자를 선택함으로써 안정적인 성과개선 효과가 나타났다. 결론적으로 본 연구에서 제안한 CO-NN은 다양한 분류자의 선택을 통한 앙상블 구성과 앙상블 학습의 안정적인 성과 개선에 효과적으로 사용될 수 있음을 보여주었다.

6. 결론 및 향후 연구 방향

본 연구는 NN 앙상블의 안정적인 성과 개선을 위하여 CO-NN을 제안하였다. CO-NN은 범위 최적화 과정에서 분류자의 다양성 확보를 VIF 분석 결과를 제약요인으로 활용하였고 최적화된 성과개선을 위하여 GA를 이용하였다. 기업부실화 예측문제에 적용 결과, CO-NN은 상관관계를 고려한 분류자 선택과 최적의 앙상블 결합을 통하여 NN 앙상블의 다양성 확보와 안정적인 성과 개선에 효과적으로 사용될 수 있음을 보여주었다.

그러나 본 연구의 한계점을 해결하기 위한 다음과 같은 후속연구가 수행되기를 기대한다.

첫째, 본 연구는 범위 최적화에 중점을 두고 연구를 진행하였으나, 최적화의 또 다른 문제는 선택

된 분류자 사이의 최적의 결합함수를 탐색하는 결정 최적화 문제이다. 본 연구의 후속연구로 결정 최적화에 대한 연구를 진행하고자 한다.

둘째, 앙상블 학습의 또 다른 문제는 데이터 노이즈(noise) 문제이다. 학습표본의 노이즈는 학습 알고리즘에 있어서 분류경계를 왜곡시키고 학습 성과를 저하시킨다. 특히, 오분류 관측치의 학습에 초점을 맞추고 있는 부스팅 앙상블의 경우 새로운 분류자 생성시에도 반복적으로 영향을 미치게 된다. 이상치 처리를 위한 Probabilistic Roulette Selection, KKT Condition-based Heuristic Selection, Automatic Feature Selection 등의 SVM 앙상블 기법이 제안되었다(Maia et al., 2008). 이러한 연구 결과와 접목된 발전적인 후속연구가 수행되기를 기대한다.

참고문헌

Alfaro, E., M. Gámez and N. García, "Multiclass corporate failure prediction by AdaBoost.M1", *AdvancedEconomic Research*, Vol.13(2007), 301~312.

Alfaro, E., N. García, M. Gámez and D. Elizondo, "Bankruptcy forecasting : an empirical comparison of AdaBooost and neural networks", *Decision Support Systems*, Vol.45 (2008), 110~122.

Bauer, E. and R. Kohavi, "An empiricalcomparison of voting classification algorithms : Bagging, boosting, and variants", *Machine Learning*, Vol.36(1999), 105~139.

Breiman, L., "Bagging predictors", *Machine learning*, Vol.24, No.2(1996), 123~140.

Buciu, I., C. Kotrooulos and I. Pitas, "Combining support vector machines for accuracy face detection", Proc. ICIP, (2001), 1054~1057.

Dong, Y. S. and K. S. Han, "A comparison of several ensemble methods for text categorization", *IEEE International Conference on Service Computing*, 2004.

Drucker, H. and C. Cortes, "Boosting decision trees", *Advanced Neural Information Processing Systems*, Vol.8(1996).

Evgeniou, T., L. Perez-Breva, M. Pontil and T. Poggio, "Bound on the generalization performance of kernel machine ensembles", *Proc. ICMI*, (2000), 271~278.

Fawcett, T., "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol.27(2006), 861~874.

Freund, Y. and R. E. Schapire, "A decision theoretic generalization of online learning and an application to boosting", *Journal of Computer and System Science*, Vol.55, No. 1(1997), 119~139.

Hansen, L. and P. Salamon, "Neural network ensembles", *IEEE Trans, PAMI*, Vol.12(1990), 993~1001.

Ho, T. K., "Multiple classifier combination: lessons and next steps, in Hybrid Methods in Pattern Recognition(Ed. By H. Bubke and A. kandel)", World Scientific, 2002.

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

Kim, M. J. and D. G. Kang, "An Ensemble with neural networks for bankruptcy prediction", *Expert Systems with applications*, Vol.37 (2010), 3373~3379.

Kim, Y. W., I. S. Oh, "Classifier ensemble selection using hybrid genetic algorithms", *Pattern Recognition Letters*, Vol.29, No.6

(2008), 796~802.

Maclin, R. and D. Opitz, "An empirical evaluation of bagging and boosting", *Proceedings of the Fourteenth National Conferenceon Artificial Intelligence*, (1997), 546~551.

Maia, T. T., A. P. Braga and A. F. Carvalho, "Hybrid classification algorithms based on boosting and support vector machines", *Kybernetes*, Vol.37, No.9(2008), 1469~1491.

Oliveira, L. S., R. Sabourin, F. Bortolozzi and C. Y. Suen, "Feature selection for ensembles : a hierarchical multi-objective genetic algorithm approach", *ICDAR*, 2003.

Quinlan, J. R., "Bagging, boosting and C4.5. Machine Learning", *Proceedings of the Fourteenth International Conference*", (1996), 725~730.

Valentini, G., M. Muselli and F. Ruffino, "Bagged ensembles of SVMs or gene expression data analysis", *The IEEE-INNS-ENNS International Joint Conference on Neural Networks*, (2003), 1844~1849.

Zhou, Z. H., J. X. Wu, and W. Tang, "Ensembling neural networks: many could better than all", *Artificial Intelligence*, Vol.137 (2002), 239~263.

Abstract

Optimal Selection of Classifier Ensemble Using Genetic Algorithms

Myungjong Kim

Ensemble learning is a method for improving the performance of classification and prediction algorithms. It is a method for finding a highly accurateclassifier on the training set by constructing and combining an ensemble of weak classifiers, each of which needs only to be moderately accurate on the training set. Ensemble learning has received considerable attention from machine learning and artificial intelligence fields because of its remarkable performance improvement and flexible integration with the traditional learning algorithms such as decision tree (DT), neural networks (NN), and SVM, etc.

In those researches, all of DT ensemble studies have demonstrated impressive improvements in the generalization behavior of DT, while NN and SVM ensemble studies have not shown remarkable performance as shown in DT ensembles. Recently, several works have reported that the performance of ensemble can be degraded where multiple classifiers of an ensemble are highly correlated with, and thereby result in multicollinearity problem, which leads to performance degradation of the ensemble. They have also proposed the differentiated learning strategies to cope with performance degradation problem.

Hansen and Salamon (1990) insisted that it is necessary and sufficient for the performance enhancement of an ensemble that the ensemble should contain diverse classifiers. Breiman (1996) explored that ensemble learning can increase the performance of unstable learning algorithms, but does not show remarkable performance improvement on stable learning algorithms. Unstable learning algorithms such as decision tree learners are sensitive to the change of the training data, and thus small changes in the training data can yield large changes in the generated classifiers. Therefore, ensemble with unstable learning algorithms can guarantee some diversity among the classifiers. To the contrary, stable learning algorithms such as NN and SVM generate similar classifiers in spite of small changes of the training data, and thus the correlation among the resulting classifiers is very high. This high correlation results in multicollinearity problem, which leads to performance degradation of the ensemble.

Kim,s work (2009) showedthe performance comparison in bankruptcy prediction on Korea firms

using tradition prediction algorithms such as NN, DT, and SVM. It reports that stable learning algorithms such as NN and SVM have higher predictability than the unstable DT. Meanwhile, with respect to their ensemble learning, DT ensemble shows the more improved performance than NN and SVM ensemble. Further analysis with variance inflation factor (VIF) analysis empirically proves that performance degradation of ensemble is due to multicollinearity problem. It also proposes that optimization of ensemble is needed to cope with such a problem.

This paper proposes a hybrid system for coverage optimization of NN ensemble (CO-NN) in order to improve the performance of NN ensemble. Coverage optimization is a technique of choosing a sub-ensemble from an original ensemble to guarantee the diversity of classifiers in coverage optimization process. CO-NN uses GA which has been widely used for various optimization problems to deal with the coverage optimization problem.

The GA chromosomes for the coverage optimization are encoded into binary strings, each bit of which indicates individual classifier. The fitness function is defined as maximization of error reduction and a constraint of variance inflation factor (VIF), which is one of the generally used methods to measure multicollinearity, is added to insure the diversity of classifiers by removing high correlation among the classifiers. We use Microsoft Excel and the GAs software package called Evolver.

Experiments on company failure prediction have shown that CO-NN is effectively applied in the stable performance enhancement of NNensembles through the choice of classifiers by considering the correlations of the ensemble. The classifiers which have the potential multicollinearity problem are removed by the coverage optimization process of CO-NN and thereby CO-NN has shown higher performance than a single NN classifier and NN ensemble at 1% significance level, and DT ensemble at 5% significance level.

However, there remain further research issues. First, decision optimization process to find optimal combination function should be considered in further research. Secondly, various learning strategies to deal with data noise should be introduced in more advanced further researches in the future.

Key Words : Neural Networks, Ensemble, Genetic Algorithms, Coverage Optimization

저 자 소 개

김명종

성균관대학교 회계학과, 동대학원에서 경영학 석사학위 취득 후 한국과학기술원에서 경영공학박사 학위를 취득하였다. 현재 부산대학교에 경영학과 교수로 재직하고 있으며 주요 연구관심 분야는 회계, 재무, 데이터마이닝, 지식공학 등의 결합메커니즘이다.