지능정보연구 제15권 제4호 2009년 12월(pp.99~121)

유전자 알고리즘 기반의 기업부실예측 통합모형

옥중경
동국대학교-서울 일반대학원 경영정보학과
(joongkok@naver.com)

김경재
동국대학교-서울 경영대학 경영정보학과
(kjkim@dongguk.edu)

최근 데이터마이닝 기법을 이용하여 기업의 부실을 예측하고자 하는 연구가 많이 이루어져 왔다. 여러 연구자들에 의해 다양한 데이터마이닝 기법이 연구되었으나 각 방법론이 장단점을 가지고 있기에 이를 보완적으로 사용하고자 하는 결합기법에 대한 연구도 꾸준하게 발표되고 있다. 본 연구에서는 데이터마이닝 기법을 각 기법의 특성을 바탕으로 4가지 형태로 구분하고 각 형태의 대표적인 기법을 선택하여 이를 유전자알고리즘을 통하여 통합하는 기법을 제안한다. 유전자알고리즘은 전역최적화기법으로 다양한 기법의 결과를 유기적으로 통합하여 최적해 또는 유사최적해를 찾게 해 줄 것이다. 본 연구에서는 기업부실예측에서 유용한 모형을 찾기 위하여 단일모형, 기존의 통합모형과 본 연구에서 제안하는 유전자알고리즘 통합기법의 결과를 비교한다.

논문접수일 : 2009년 11월 12일 게재확정일 : 2009년 11월 29일 교신저자 : 김경재

1. 서론

기업은 수많은 이해관계자들과 직・간접적으로 관계를 맺고 있다. 따라서 여러 기업의 부실은 국가적인 생산력 약화나 관련 기업의 연쇄도산, 국가신용도 하락 등을 초래하고 다수의 이해관계자에게 경제적 손실을 발생시키는 등 사회문제를 야기시킨다. 또한 기업규모의 확대로 기업이 국민경제에 미치는 영향이 크므로, 최근의 몇몇 중견 건설업체의 부도와 이에 따른 금융기관들의 동반 부실화 및 금융기관을 정상화시키기 위한 공적 자금의 투입은 국민 모두를 기업의 이해관계자로 만들고 있다. 이와 같이 국민경제에 큰 영향을 미치는 기업의 부실을 조기에 예측할 수 있다면 이해관계자들은 조기에 필요한 조치를 취할 수 있으며 손실을 최소화할 수 있을 것이다.

또한, 정확한 기업부실 예측은 시장 원리에 따른 금융자원의 효율적 배분을 도모하고 금융기관 자산의 건전화를 촉진할 수 있는 기반이다. 과거 국내 금융기관들의 대출관행은 모기업 및 관계기업의 규모나 담보설정 여부 등을 위주로 기업평가를 해 왔다. 그러나 외환위기 이후 기업의 부도와 경영환경의 급변으로 인해 대부분의 국내 금융기관들은 상당한 규모의 부실채권을 부담하게 되었고 이러한 영업 환경의 변화는 각 금융기관들로 하여금 기업부실예측의 중요성을 인식하게 하는 원인이 되고 있다.

이러한 인식에 따라 기업의 부실을 예측하기 위한 모형에 대한 연구가 활발하게 진행되어 왔다. 기존의 연구에서는 전통적인 통계 기법을 사용하여 기업 부실을 예측하기 위한 문제를 해결하려고 하였다. 그러나 이러한 연구들에는 통계적 가정이

만족되어야 한다는 방법론적인 한계가 존재하고 있다(배재권과 김진화, 2006). 따라서, 여러 가지 가정에 상대적으로 유연하고 예측성능이 좋은 것으로 알려진 인공신경망, 규칙유도기법 등과 같은 인공지능 기법을 기업 부실 예측 문제에 적용하는 연구가 최근에 많이 발표되고 있다(배재권과 김진화, 2006; 이건창 등, 1994; Barniv et al., 1997).

본 연구에서는 선행연구에서 기업부실예측에 유용한 것으로 알려진 여러 가지 인공지능 기법의 모형들을 유기적으로 결합하여 예측성능이 개선된 통합모형을 제안하고자 한다. 기존의 연구들에서도 다양한 모형을 결합한 통합모형을 제안하여 왔으나, 기존 연구들은 단일 모형들의 결과에 대한 상대적 가중치에 초점을 맞추고 있었다면, 본 연구의 제안 모형은 예측을 판단하는데 활용되는 결과치를 전역최적화 기법 중의 하나인 유전자 알고리즘을 이용해 최적화하도록 설계되어 있다는 점에서 차별화된다. 또한 지금까지의 연구들은 단일 기법에 의한 기업부실모형을 개별적으로 적용하거나, 단일 기법들을 결합하는 방식으로만 사용되었을 뿐이고, 이종의 인공지능 기법들을 동시에 적용하여, 결합하는 모형에 관한 연구는 많이 소개되지 않았으며, 결합 방식도 투표 혹은 단순 평균만을 통해 결합하는 것이 일반적이었다.

본 연구에서는 기업평가 실무에서 사용되는 기업들의 실제 재무 및 비재무 자료를 이용하여 부실예측을 위한 로지스틱 회귀분석, 의사결정나무, 인공신경망, 사례기반추론 모형의 결과들을 전역최적화 기법을 통해 가중치를 조정한 후 통합하고, 그 유용성을 검증하고자 한다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 본 논문을 위한 이론적 배경으로서 모형 결합에 관한 선행 연구를 기존 문헌을 통해서 살펴보고 이어서 유전자 알고리즘의 기본 작동원리에 대해서 알아본다. 제 3장에서는 본 연구의 제안 모형인 유전자 알고리즘 기반의 새로운 이종 인공지능 기법 간 통합 모형을 소개한다. 제 4장에서는 앞서 제시한 모형의 유용성을 검증하기 위한 실험 데이터 및 설계 내용을 설명하고, 제5장에서는 실험 결과를 종합적으로 정리해 제시하도록 한다. 로지스틱 회귀분석, 의사결정나무, 인공신경망 모형 및 사례기반추론인 K-Nearest Neighbor의 실험 결과를 정리하였고, 또한, 이러한 모형들을 통합한 모형의 실험 결과를 단순 평균모형과 가중 평균모형 및 유전자 알고리즘을 이용한 모형의 결과를 제시한다. 그리고 이러한 연구의 결과를 검증할 수 있는 방법에 대해서 설명을 한 후, 실험의 결과를 종합적으로 정리한다. 끝으로 마지막 장에서는 결언과 함께 본 연구와 관련된 향후 연구방향이 제시된다.

2. 이론적 배경

서론에서 언급한 바와 같이 본 연구에서는 기업부실예측을 위한 새로운 결합모형을 제안하고자 하며, 이에 앞서 결합모형의 기반이 되는 단일모형의 구축이 필요하다. 본 연구에서는 단일모형을 구현하기 위하여 이분류 회귀분석기법인 LOGIT(Logistic Regression, 로지스틱 회귀분석), 인공지능 기법인 ANN(Artificial Neural Networks, 인공신경망), CBR(Case Based Reasoning, 사례기반추론), DT(Decision Tree, 의사결정나무) 모형을 사용한다. 이들 기법에 대해서는 이미 많은 선행연구에서 충분히 설명되었으므로 이에 대한 상세한 설명은 생략하도록 한다. 본 장에서는 결합모형의 선행연구들과 본 연구에서 제안하는 통합모형의 기반 알고리즘인 유전자 알고리즘의 기본적인 작동원리에 대해서 설명한다.

2.1 모형 결합에 관한 연구

모형 간 결합에 관한 연구의 특징은 여러 가지 서로 다른 예측 모형의 결과들을 결합하거나 통합하여 적용하면 한 가지 모형을 사용할 때 보다 통합된 모델의 사용이 성능을 더 높인다는 것을 제시하는 통합 방법론의 중요성을 강조하고 있다.

이와 관련된 연구로 홍승현과 신경식(2003)은 기업부도예측을 위해서 인공신경망 모형을 여러 개 결합하는 결합모형을 제시 하였다. Kim(2004)은 기업의 재무부실이나 부도의 예측에 있어서 인공신경망이 높은 예측력을 보여 주고 있으나, 잘못된 자료로 인해 예측의 불일치가 나타나는 단점을 해결하기 위한 자료 축소의 방법을 유전자알고리즘과 인공신경망을 결합한 방법을 제안하였고, 실험결과는 유전자알고리즘과 인공신경망의 결합이 신뢰성 있는 방법임을 보여 주었다.

추휘석 등(2004)은 다수의 인공신경망 모형을 통합한 부실예측 모형을 제시하였고, 다수의 신경망 모형의 결과에 따른 데이터를 학습시켜 보다 나은 데이터의 패턴을 신경망에 적용하였다. 배재권과 김진화(2006)는 기업부도예측을 위해서 통계적 방법인 다변량 판별분석, 로지스틱 회귀분석과, 인공지능적인 방법인 인공신경망, 규칙유도기법, 베이지안 망의 5가지 방법론을 통합한 인공신경망 통합 모형을 제시하였고, 이에 대한 실험결과 인공신경망 통합모형이 기존의 모형들에 비해 우수한 예측 정확성을 나타냄을 보여 주었다. Kim et al.(2006)은 서로 다른 예측기법들인 인공신경망 모형, 전문가 판단, 사용자 판단의 결과들을 동시에 결합하는 결합모형을 제시한 바 있다.

홍태호와 신택수(2007)는 기업신용등급 산출모형 개발에 있어 재무지표를 이용하여 로짓모형과 인공신경망을 결합한 재무모형을 만들고 전문가의 주관적 판단을 합리적으로 측정해주는 AHP (Analytic Hierarchy Process)를 이용해 비재무 모형을 만든 후 이것을 하나의 모형으로 통합하여 기업신용등급을 산출하는 모형을 제시하였다. 또한, 최영수와 장욱(2007)은 데이터 마이닝 기법간의 결합이 아닌, 변수간의 결합을 시도하였다. 재무변수와 시장변수를 결합하여 은행에 대한 부도예측 모형을 제시하였고, 이에 대한 적합성 검증에서 결합모형의 경우 우수한 변별력을 나타냈다. 최근 연구로서, 이형용(2008)은 주가지수 등락 예측을 위해 로지스틱 회귀분석, SVM(Support Vector Machine), 인공신경망 등의 예측결과를 결합하여 유전자 알고리즘을 적용해서 보다 정확한 예측 모형을 산출하는 결합 모형을 제시하였다.

위에서 설명한 모형결합방법들은 대체로 단일 모형의 결과를 평균하거나, 다른 모형의 입력변수로 활용하는 방법을 취하였다. 이러한 방법은 다중모형의 결합에는 유용하고 사용이 용이하다는 장점이 있으나, 여러 모형의 결과를 유기적으로 결합하지는 못할 수 있다는 한계점이 있다. 한편, 이형용(2008)의 연구처럼, 유전자 알고리즘과 같은 전역 최적화 기법을 활용하여 여러 모형의 결합가중치를 결정하는 방법은 다중모형의 결과를 유기적으로 최적화할 수 있다는 장점이 있다. 그러나 이형용(2008)의 연구에서 제시한 결합방법 중 상당수의 모형은 보다 단순화한 방법을 통하여 유사한 결합결과를 얻을 수 있으므로 이에 대한 보완이 된다면 보다 활용이 용이하면서도 유기적인 결합이 가능한 다중모형결합방법을 제시할 수 있을 것이다. 이상에서 논의된 모형 결합에 관한 연구들을 정리하면 <표 1>과 같다.

2.2 유전자 알고리즘의 기본 작동원리

유전자 알고리즘은 모든 생물이 주어진 다양한 환경 속에 적응함으로써 살아 남는다는 Darwin의 적

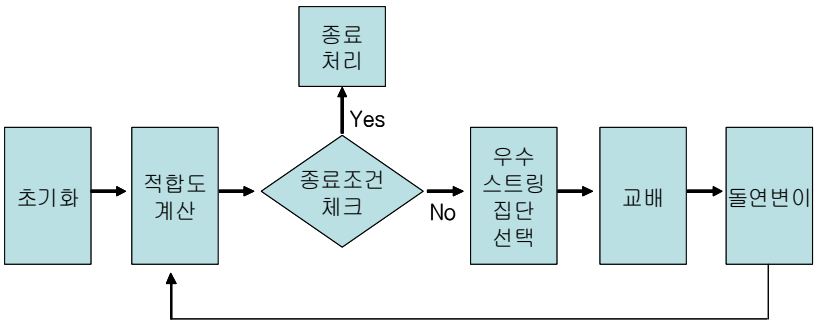
<표 1> 모형 결합에 관한 국내외 주요 연구

연 도	연구자	연구 내용
2003	홍승현, 신경식	유전자 알고리즘을 이용하여 기업부도 예측을 위한 입력 변수를 도출하였고, 인공신경망 모형을 여러 개 결합하는 결합모형을 제시
2004	Kim, Kyoung-jae	실험을 위한 방대한 자료의 축소 방법을 위해서 유전자 알고리즘과 인공신경망을 결합한 방법을 제안
2004	추휘석, 민지경, 이인호	다수의 인공신경망 모형을 통합한 부실예측 모형을 제시했고, 다수의 신경망 모형의 결과에 따라 데이터를 학습시켜 보다 나은 데이터의 패턴을 신경망에 적용
2006	배재권, 김진화	판별분석, 로지스틱 회귀분석, 인공신경망, 규칙유도기법, 베이지안 망의 5가지 방법론을 통합한 인공신경망 통합 모형을 제시
2006	Kim, M. J., Min, S. H., Han, I.	서로 다른 예측기법들인 인공신경망 모형, 전문가 판단, 사용자 판단의 결과들을 동시에 결합하는 결합모형을 제시
2007	홍태호, 신택수	로짓모형과 인공신경망을 결합한 재무모형과 AHP를 이용한 비재무 모형을 통합하는 모형을 제시
2007	최영수, 장욱	재무변수와 시장변수를 결합하여 은행에 대한 부도예측 모형을 제시하였고, 결합모형이 우수함을 제시
2008	이형용	로지스틱 회귀분석, SVM, 인공신경망 등 예측 결과를 결합하여 유전자 알고리즘을 적용해서 보다 정확한 예측 모형을 산출하는 결합 모형을 제시

자생존의 이론과 Mendel의 유전의 법칙을 응용하여 큰 공간을 탐색할 수 있는 확률적 탐색 기법이다(홍승현과 신경식, 2003). 또한, 유전자 알고리즘은 점에 의한 탐색이 아니라 개체들이 모여 이룬 개체군에 의한 병렬적인 탐색이라는 점에서 기존의 최적화 알고리즘과 다르며, 탐색의 방향이나 영역이 초기 값에 의해서 결정되지 않고 세대마다 확률적으로 결정되므로 지역 최소점에 빠질 가능성이 적어 전역 최적화가 가능한 알고리즘이다(http://icat.snu.ac.kr : 3000/Toolboxes/genetic_algorithm/).

유전자 알고리즘의 작동원리는 초기화, 적합도 계산, 종료조건 체크, 선택, 교배와 돌연변이로 구분할 수 있으며 이를 도식화하면 <그림 1>과 같다(Pal and Wang, 1996 수정).

초기화 단계에서 유전자 알고리즘으로 해결하기 위한 문제를 적당한 형태로 표현해야 하는데, 2진 형태인 비트(Bit)열로 표기하며, 이를 염색체(Chromosome)라고 한다. 이후 임의로 염색체를



<그림 1> 유전자 알고리즘 작동원리

선정하고, 염색체의 성과를 수치적으로 표현한 목적함수에 적합한지 선정된 염색체의 적합도를 계산한다. 목적함수에 대한 적합도에 따라서 성과가 우수한 유전자는 보존하고 성과가 나쁜 유전자는 탈락시키는 우수 스트링 선택 작업을 수행한다. 즉, 초기화 단계 후에는 각각의 염색체는 적합함수에 의해 평가되며 적합 함수의 값에 따라 잘 적응된 개체들이 적응하지 못한 개체들보다 더 많이 재생산 된다.

유전자 알고리즘은 반복적으로 사용되는 연산자로 작동된다, 교배연산자는 하나의 자식을 조합하기위해 다른 부모로부터 염색체의 일부를 교환받아 새로운 염색체를 생성하며, 교배과정을 통해서 부모 염색체가 탐색하던 방향과 다른 방향에서 해를 생성하게 되고, 부모세대와는 다른 새로운 영역에서 최적 해를 탐색하도록 해준다(홍승현과 신경식, 2003; Kim, 2004).

돌연변이 연산자는 선택된 염색체의 구성 요소인 비트를 무작위로 변경한다. 교배과정을 통해서도 새로운 염색체가 만들어지지만, 계속적인 새로운 정보가 만들어지지 못한다는 한계점이 있다. 그러나 돌연변이 과정을 통해서 새로운 염색체가 만들어짐으로써 이러한 한계점을 극복할 수 있다(홍승현과 신경식, 2003; Kim, 2004). 상기와 같은 과정을 반복하면서 새로 생성된 염색체는 적합도를 다시 계산한 후에 종료조건을 체크한다. 이때, 종료조건이 만족되면 유전자 알고리즘은 작동을 멈추게 되고, 만족하지 않으면 <그림 1>의 과정을 반복하면서 보다 적응도가 높은 우수 염색체를 만들어 간다.

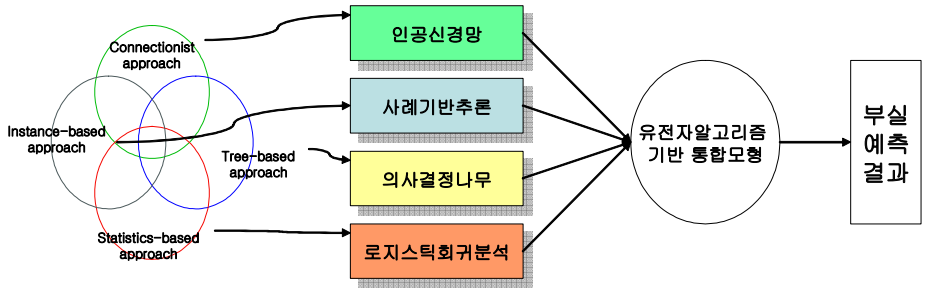
3. 유전자 알고리즘 기반의 통합모형

기업의 부실 예측에 적용되는 인공지능 기법은 일반적으로 <표 2>에서 보는 것처럼 Connectionist Approach, Instance-based Approach, Tree-based Approach 그리고 Statistics-based Approach의 4가지로 분류할 수 있으며, Connectionist 방법에서 많이 사용되는 기법은 인공신경망이고, Instance-based 방법은 사례기반추론, Tree-based 방법은 의사결정나무, Statistics-based 방법은 로지스틱 회귀분석이다.

기존의 부실예측 연구에서는 여러 가지 단일모형의 한계점이 지적되고 있는데, 이러한 한계점으로는 통계학적 모형이 비선형 예측에 한계를 보일 수 있다는 점, Tree 기반모형은 연속형 자료의 처리가 곤란한 단점이 있으며, Connectionist approach에서는 예측 성능은 좋으나 결과에 대한 설

<표 2> 기업의 부실 예측에 적용되는 인공지능 기법

인공지능 기법	특 징	대표적인 예
Connectionist Approach	학습과정을 통해 분석데이터의 전체적인 평균 오류에 기반하여 예측결과를 산출하는 연결선 방식의 Data Mining 방법으로 선형, 비선형 예측이 가능	ANN
Instance-based Approach	예측하고자 하는 사례와 유사한 사례와의 유사성에 기반하여 예측결과를 산출하는 방법	CBR, CF
Tree-based Approach	나무형식의 검색 공간 분할 방식의 Data Mining 방법	CART, CHAID, ID3
Statistics-based Approach	통계학적 분석방법을 근간으로 하는 Data Mining 방법으로 선형 예측이 가능	LOGIT, DA



<그림 2> 유전자알고리즘 기반 인공지능 기법 결합모형

<표 3> 각 모형 별 결과 값의 형태

모 형	로지스틱회귀분석 (LOGIT)	의사결정나무 (CART)	인공신경망 (ANN)	사례기반추론 (K-NN)
결과값	0과 1사이의 확률	0과 1사이의 확률	0과 1사이의 값	0 또는 1

명력이 부족하며, Instance based 방법은 사용이용이하나 connectionist approach 등에 비해 예측력이 떨어질 수 있다는 점이 한계점으로 지적되고 있다.

따라서 본 연구에서는 상기의 한계점을 보완하기 위해서 4가지의 주요 기법들인 인공신경망, 사례기반추론, 의사결정나무 그리고 로지스틱 회귀분석의 기법을 사용하고, 그들의 예측 결과를 유기적으로 결합하여 보다 정확성을 높일 수 있는 기업 부실예측의 통합 모형을 설계하고자 하며, 다양한 모형의 유기적 통합을 위하여 최적화 알고리즘인 유전자 알고리즘 (GA, Genetic Algorithms)를 활용한다. 본 연구에서 제안하는 연구의 프로세스와 모형은 <그림 2>와 같다.

3.1 통합모형용 실험 데이터의 특성

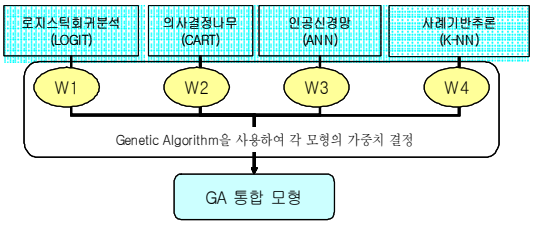
유전자 알고리즘 기반의 통합모형 구축을 위한 실험에서는 의사결정나무 모형에서는 CART(Classification And Regression Trees)를 사용하며, 사례기반추론 모형에서는 K-NN(K-nearest Neighbor)을 사용한다. 실험 데이터는 각 단일모형의 결과값을 사용하게 된다. 각 단일모형의 특성 상 결과값은 상이한 형태로 출력되는데, 로지스틱 회귀분석 및 의사결정나무의 결과 값은 0과 1사이의 확률 값으로 주어지고, 인공신경망의 경우에도 0과 1사이의 값으로 주어진다. 이에 반해 K-NN의 결과값은 최종 분류 카테고리 값인 0(정상)과 1(부도)로서 주어진다. 각 단일모형의 결과 값의 형태를 정리하면 <표 3>과 같다.

또한 K-NN을 제외한 다른 모형들에서는 모든 데이터 셋에 대한 결과 값이 주어지는 반면에 K-NN은 검증용 데이터 셋에 대해서만 결과 값이 주어지기 때문에 유전자 알고리즘 기반의 통합모형 구축을 위해서는 K-NN에서 학습용 데이터 셋의 결과 값을 All-but-one 방법[1]을 통하여 새롭게 결정한다.

1) 검증용 데이터 셋을 제외한 학습용 데이터 셋 내에서 한 기업의 부도 및 비부도 여부를 그 기업을 제외한 다른 모든 기업에 비추어 결정하는 방법이다. 즉, 학습용 데이터 셋에서 하나의 기업을 선택하여 새로운 검증용 데이터 셋으로 가정한 후, 선택된 기업을 제외한 다른 모든 기업을 학습용 데이터 셋으로 간주하여 K-NN을 적용하였을 때의 결과 값을 출력하는 방법을 의미한다.

3.2 유전자 알고리즘 기반 통합모형의 개념

유전자 알고리즘 기반의 통합모형은 유전자 알고리즘을 사용하여 각각의 단일모형에 배당될 최적의 가중치를 찾고자 하는 모형을 일컫는다. 이때 각 단일모형의 가중치의 합은 1이 되어, 유전자 알고리즘 통합모형을 통한 최종 결과 값 역시 0과 1사의 값이 출력되게 된다. 유전자 알고리즘 기반의 통합모형을 그림으로 묘사하면 <그림 3>과 같다.



<그림 3> 유전자 알고리즘 기반 통합모형의 개념도

3.3 유전자 알고리즘 기반 통합모형의 절차

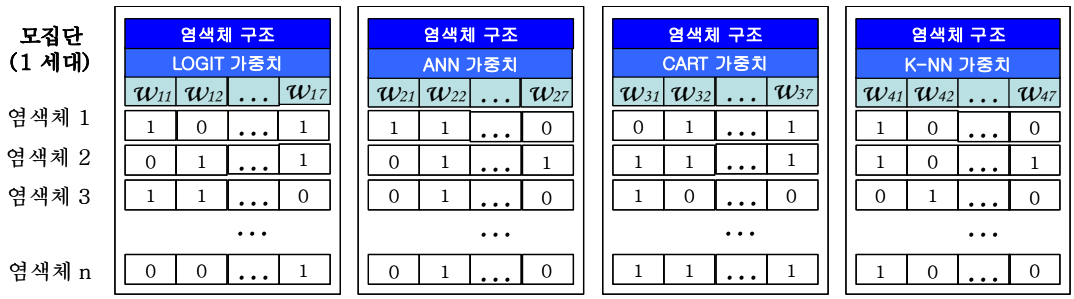
제안 모형은 5단계의 과정을 거쳐 진행되는데, 각 단계별로 어떤 처리가 이루어지는지에 대해 살펴보기로 한다.

3.3.1 염색체 구조 설계

본 연구에서는 유전자 알고리즘에 의해서 ANN, LOGIT, CART, K-NN의 4가지 기법에 대한 상대적 가중치가 탐색된다. 탐색되는 각 가중치들은 GA가 탐색할 4개의 변수가 된다. GA를 적용하기 위해서는 변수들을 '2진 문자열의 형태'로 구성된 염색체의 구조를 표현하여야 한다. 여기서 각 기법별 가중치의 경우 상대적 가중치를 도출해야 하므로, 각각 1에서 128사이의 값을 갖도록 설계하였다. 128의 표현은 2진수로 2^7이므로, 각 기법의 가중치 변수를 위한 염색체에는 각각 7비트씩 할당하였다. <그림 4>는 본 연구에서 제안하고 있는 모형에서 적용하고자 하는 유전자 알고리즘의 염색체 구조를 나타내고 있다.

3.3.2 초기 모집단 생성 및 적합도 함수 정의

탐색에 들어가지 전에, 초기 모집단인 각 기법의 최적 가중치를 찾기 위한 염색체들의 집합을 초기화해야 하는데, 이때 초기화는 각 염색체의 비트값들을 이진 무작위 값으로 부여하는 방법을 통해 이루어진다. 그런 다음, 이 모집단을 진화시키기 위해서는, 모집단을 구성하는 각 염색체들을 평가할 어떤 기준이 필요하게 된다. 이러한 기준을 '적합도 함수(fitness function)'이라고 부른다. 본 연구에서는 'ANN, LOGIT, CART, K-NN의 4가지 기법들을 어떤 가중치로 결합할 때 기업부실에



<그림 4> 제안모형의 염색체 구조

대한 가장 정확한 예측을 할 수 있는 최적의 가중치를 결정하는 것'이 목적이므로, 아래의 식 (1)을 이용하여 4가지 기법에 대한 최적 가중치를 계산하며 이것이 적합도 함수가 된다. 식 (1)은 'Real Weight = 개별 weight/weight의 총합'을 나타내며, 최종 Weight는 학습용 데이터 셋의 예측률을 최대화하는 weight 중에서 검증용 데이터 셋의 예측률을 최대화하는 Weight로 결정됨을 의미한다.

$$W_i = \frac{w_i}{\sum_i w_i} \quad (1)$$

3.3.3 유전자 알고리즘의 탐색 및 새로운 세대의 생성

이 단계에서는 앞의 단계에서 도출된 각 개별 염색체의 적합도를 기준으로 하여 결과 값이 최대화하는 방향으로 유전자 알고리즘의 진화가 이루어지도록 탐색이 수행된다. GA는 2단계에서 설정된 염색체를 대상으로 교배, 돌연변이 등 다양한 과정을 적용하여, 모집단의 새로운 세대(generation)를 생성하게 되는데, 이로 인해서 계속 새로운 각 기법별 가중치 후보들이 만들어지며, 이것을 다시 학습용 데이터에 적용해 보게 된다. 위와 같은 과정은 중지조건이 만족될 때까지 상기 작업을 계속 반복하게 된다. 이러한 과정을 통해 학습용 데이터에 가장 우수한 성과를 보이는 4개의 변수 값을 탐색하게 된다.

3.3.4 종료 조건 체크

유전자 알고리즘은 최적 또는 유사한 최적 해를 찾을 때까지 진화를 반복하도록 설계되어있다.

본 연구의 모형에서도 앞의 3단계를 통해서 새로운 모집단을 생성하고 그 모집단을 다시 반복 수행하게 된다. 이 과정은 사전에 정해진 종료조건에 도달할 때까지 계속해서 반복 수행되는데, 종료조건은 일반적으로 '최대 진화 세대수(the number of generation)'로 설정된다. 종료조건에 따라 진화과정이 마무리되면, 최적 혹은 최적에 근접한 4개의 변수 값이 도출되게 된다.

3.3.5 최종 검증 및 확인

최종 검증 및 확인과정에서는 앞의 단계에서 찾아낸 가중치가 모형구축에 사용되지 않은 데이터(unknown data)들에 대해서도 유효한 예측 성과를 보이는지 최종적으로 확인하는 단계가 된다. 이를 위해 본 단계에서는 최종적으로 선택된 각 기법별 가중치를 모형 구축에 사용하지 않은 검증용 데이터(Hold-out Data)에 적용하여, 그 결과를 살펴보게 된다.

본 과정이 필요한 이유는 GA가 학습용 데이터에 대해 예측 적중률을 최대화 하는 방향으로 가중치를 최적화 하려고 하는데, 이 경우 최적화된 가중치가 일반성을 지니는 값이 아니라서, 학습용 데이터에 대해서는 예측을 잘 하지만 새로운 데이터에 대해서는 예측을 잘 하지 못하는 문제가 발생할 수도 있기 때문이다. 그러므로 모형 구축에 사용되지 않은 검증용 데이터에 적용해 봄으로써, 이른바 '과잉학습' 문제가 발생한 상황인지 아닌지를 최종적으로 확인해 볼 필요가 있다.

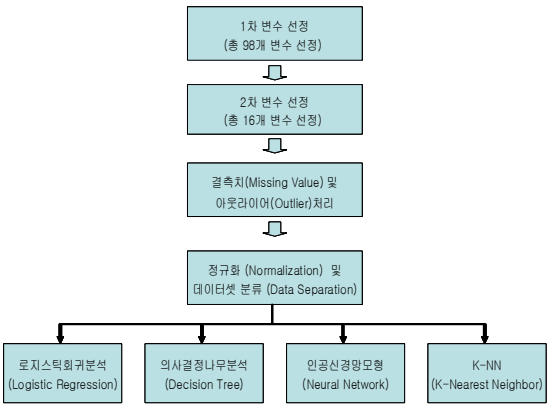
4. 실험 설계

4.1 실험 데이터

제안된 연구 모형의 유용성을 검증하기 위해서 국내의 중공업 외감 기업의 자료를 이용하여 모형

을 구축한다. 모형 구축에 사용된 표본은 도산 기업의 경우, 국내 외감 중공업으로 1994년부터 최근인 2008년까지의 부도나 폐업을 한 기업의 데이터를 사용했으며, 정상 기업의 경우는 표준산업 코드 기준으로 도산 기업과 같은 업종으로 구분되는 기업들 중에서 도산 기업과 같은 수를 추출하여 실험 데이터를 준비하였다.

실험을 위해 준비된 데이터는 정상 기업이 1,570개, 도산 기업이 1,570개로 모두 3,140개의 기업 데이터가 사용되었고, 각 기업의 데이터는 각 기업의 재무비율 및 규모와 관련된 지표들로 총 148개의 변수로 이루어져 있다. 데이터를 실제 실험에 사용하기에 앞서 변수 선정을 포함한 데이터 정제 과정을 <그림 5>와 같이 실시하였다.



<그림 5> 실험 설계 도표

<그림 5>에서 제시한 변수 선정 및 데이터 정제 과정을 아래에 간략히 설명하고자 한다.

4.1.1 1차 변수 선정

총 148개의 변수 중, 전체 데이터 수의 95% 이상(2,983개 이상)을 지닌과 동시에 단일표본 t검정(Independent-samples t-test)을 통한 p-value값이 0.05이하의 값을 가지는 변수만을 우선적으로 선정하였다. 1차 변수 선정 과정의 목적은 독립변수와 종속변수 간의 기본적인 관련성을 검정하기 위한 것이다. 이 결과, 91개의 변수가 선정되었다. 1차 변수선정과정에서 유의성을 나타내지는 못하였으나 선행연구에서 많이 사용되었고 신용평가 전문가들이 추천한 7개의 변수를 추가하여 총 98개의 변수를 1차 선정과정을 통하여 추출하였다.

4.1.2 2차 변수 선정

1차 변수 선정 과정을 통하여 선택된 변수들에 한하여, 단계선택 로지스틱 회귀분석(Logistic Regression with Stepwise Variable Selection)을 실시하였다. 2차 변수 선정 과정의 목적은 1차 변수 선정을 통해 유의한 것으로 확인된 독립변수들에 대해 다중공선성을 제거하기 위하여 실행되었다. 이

<표 4> 2차 변수 선정 과정을 통해 선정된 독립변수

변 수	변 수 명
B9	재료비/매출액
B13	순금융비용부담율
B15	감가상각율
B18	금융비용대총부채비율
B19	금융비용대총비용비율
S3	EBITDA/총차입금
S4	차입금의존도
S14	장단기대요금/총자산
S23	법인세부담율
S24	순운전자본대총자본비율
S26	적립금총자산비율
S28	총현금흐름대부채비율
S30	유보액대납입자본비율
D7	매입채무회전기간
D15	자본금회전율(회)
K1	업력(설립년수)

과정에서 사용된 로지스틱 회귀분석은 다변량 분석의 일환으로 진행되었다. SPSS 15.0에서 Binary Logistic Regression중 Conditional Forward Step-wise 방법을 사용한 결과 최종적으로 16개의 변수가 선정되었다. 이 과정을 통해 선정된 독립변수는 <표 4>와 같다.

4.1.3 결측치 및 극단치 처리

각 변수 데이터를 기준으로 결측치에 해당하는 기업 데이터 전체를 삭제하는 방법을 통하여 결측치를 처리하였다. 결측치 처리 후 총 2,852개의 기업 데이터 표본이 선정 되었다.

본 실험에서는 극단치(Outlier)를 정의함에 있어서, 데이터를 일렬로 정렬 하였을 때 앞 뒤 각각 0.5%에 위치한 데이터를 극단치로 정의하였다. 본 실험에서는 극단치 제거방법으로 각 변수 별로 극단치를 삭제하고 계산한 각 변수 별 평균값으로 극단치를 대체하는 방식을 이용하였다.

4.1.4 정규화 및 데이터 셋 분류

'1차 변수 선정'에서 부터 3번째 단계인 '결측치 및 극단치 처리' 까지의 과정을 거친 데이터를 사용하여 각 변수 별로 정규화(normalization) 과정을 진행하였다. 정규화 과정 이후 각 실험에 적합한 데이터셋으로 분류하였는데, 수집된 전체 데이터 중 80%에 해당되는 2,282건의 데이터를 학습용으로 활용하였고, 나머지 20%를 검증용으로 사용하였다. 한편, 인공신경망 모형과 같은 인공지능기법에 사용되는 데이터셋의 경우는 학습용 데이터셋(Training Set), 평가용 데이터셋(Test Set), 검증용 데이터셋(Validation Set)의 비중을 각각 6:2:2로 정하여 사용하였다. 각각의 데이터셋에 포함되는 데이터의 추출은 난수발생에 의한 임의추출방식(Random Sampling)에 의하여 진행되었다. 각 데이터셋의 크기와 상대비중은 다음의<표 5>와 같다.

<표 5> 데이터셋 크기 및 상대 비중

데이터 구분	데이터 개수	전체 데이터에 대한 비율
학습용	1712	60%
평가용	570	20%
검증용	570	20%
전 체	2852	100%

5. 실험 결과

5.1 단일모형 실험결과

5.1.1 로지스틱 회귀분석 모형

본 연구에서 로지스틱 회귀분석은 SPSS 15.0을 이용하였다. 이 과정에서는 데이터 정제 과정을 거쳐 선정된 16개의 변수가 모두 사용되었다. 실험결과는 <표 6>과 같이 학습용 데이터셋과 검증용 데이터셋에서 모두 80%에 도달하는 높은 예측율을 보여 주었다. 또한 학습용 데이터셋과 검증용 데이터셋의 예측율 차이는 1% 정도로, 두 개의 데이터셋에서 거의 비슷한 예측율을 보여 줌을 확인할 수 있었다.

5.1.2 의사결정 나무 모형

SPSS 15.0을 사용하여 의사결정 나무를 사용하여 예측한 결과는 <표 7>과 같았다. Tree Growing Method로는 CART를 선정하였고, 이때 가지치기(pruning)을 위하여 최대나무깊이(Maximum Tree Depth)는 5로 설정하였다. 또한 불순도 측정지표로는 Gini 방법을 사용하였으며, pruning을 하는

<표 6> 로지스틱 회귀분석 결과표

예측 / 관찰	학습용 데이터셋			검증용 데이터셋		
	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	906	206	81.5%	233	45	83.8%
1(부실)	240	930	79.5%	72	220	75.3%
전체비율			80.5%			79.5%

주) Cut Value는 0.5로 회귀식의 결과 값이 0.5보다 작으면 0으로 0.5보다 크거나 같으면 1로 분류하였다.

<표 7> 의사결정 나무 결과표

예측 / 관찰	학습용 데이터셋			검증용 데이터셋		
	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	829	283	74.6%	199	79	71.6%
1(부실)	157	1013	86.6%	50	242	82.9%
전체비율			80.7%			77.4%

데 있어 overfitting을 피하도록 설정하였다.

표의 결과를 살펴 보면, 학습용 데이터셋의 경우 로지스틱 회귀분석의 결과(80.5%)와 의사결정나무에 의한 결과(80.7%)가 거의 동일하게 나왔으나, 검증용 데이터셋의 경우 77.4%로 학습용 데이터셋의 결과와 비교하였을 때 약 3.3% 정도 낮게 나옴을 확인할 수 있었다.

5.1.3 인공신경망 모형

인공신경망 모형은 설정하는 모형에 따라 여러 가지 형태를 지닐 수 있다. 이번 연구에서 사용한 모형은 그 중 선행연구에서 가장 많이 사용된 3층 형 오류역전파 인공신경망을 사용하였다. 인공신경망의 실험을 위하여 본 연구에서는 Neuroshell2를 사용하였으며, 은닉층 처리요소의 숫자에 의한 실험결과의 변화를 관찰하기 위하여, 은닉층의 노드 수를 입력층 노드 수의 n/2개, n개, 2n개로 각각 변화시켜 가며 실험해 보았다. 따라서, 본 연구에서는 최종 선정된 입력변수가 16개 이므로, 은닉층에는 8개, 16개, 32개의 노드가 각각 사용되었다. 또한 각 층의 처리요소의 전이함수로는 비선형적인 특성을 살릴 수 있도록 로지스틱 함수를 사용하였다. 은닉층의 노드 수를 변화시켜가며 얻은 결과는 다음 <표 8>~<표 10>와 같다.

<표 8> 인공신경망 모형 결과표(n = 8)

예측 / 관찰	학습용 데이터셋			평가용 데이터셋			검증용 데이터셋		
	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	706	128	84.7%	228	50	82.0%	239	39	86.0%
1(부실)	201	677	77.1%	57	235	80.5%	78	214	73.3%
전체			80.8%			81.2%			79.5%

<표 9> 인공신경망 모형 결과표(n = 16)

예측 / 관찰	학습용 데이터셋			평가용 데이터셋			검증용 데이터셋		
	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	700	134	83.9%	228	50	82.0%	235	43	84.5%
1(부실)	199	679	77.3%	57	235	80.5%	77	215	73.6%
전체			80.5%			81.2%			78.9%

<표 10> 인공신경망 모형 결과표(n = 32)

예측 / 관찰	학습용 데이터셋			평가용 데이터셋			검증용 데이터셋		
	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	712	122	85.4%	232	46	83.5%	240	38	86.3%
1(부실)	212	666	75.9%	62	230	78.8%	81	211	72.3%
전체			80.5%			81.1%			79.1%

위 <표 8~10>에서 확인할 수 있듯이 은닉층의 노드 수가 8개(n = 8)일 때, 검증용 데이터셋의 예측율이 79.5%로 가장 높았다. 하지만 이러한 수치는 다른 케이스에서의 예측율(n = 16일 때 78.9%, n = 32일 때 79.1%)과 비교해 보았을 때 거의 차이가 나지 않음을 알 수 있다. 즉 위 모형에서 은닉층의 노드 수 변화는 예측율 변화에 그리 큰 영향을 주지 못함을 확인할 수 있었다.

5.1.4 사례기반 추론 모형

본 논문에선 사례기반추론의 일환으로 K-NN(K-Nearest Neighbor) 방법을 사용하였다. 데이터를 학습용 데이터 셋과 검증용 데이터 셋으로 나눈 뒤, Matlab 7.0의 K-NN toolbox를 적용한 결과는 <표 11>과 같다.

<표 11>과 <그림 7>을 살펴 보면, K-Parameter값이 7이 될 때까지는 예측율이 조금씩 증가함을 확인할 수 있다. 하지만 K-Parameter가 그 이상의 값을 가질 때에는 더 이상의 예측율 개선 없이 일정한 수준의 예측율을 유지하고 있는 것을 확인 할 수 있다. 따라서 본 연구에서는 K가 7일 때의 예측결과를 이용하였다.

<표 11> K-NN을 사용하였을 때의 예측 결과

K-Parameter	정확하게 분류된 데이터	잘못 분류된 데이터	예측율
1	403	167	70.70%
2	411	159	72.11%
3	424	146	74.39%
4	428	142	75.09%
5	427	143	74.91%
6	426	144	74.74%
7	435	135	76.32%
8	434	136	76.14%
9	432	138	75.79%
10	431	139	75.61%
11	430	140	75.44%
12	434	136	76.14%
13	432	138	75.79%
14	431	139	75.61%
15	431	139	75.61%

K-NN을 사용하여 얻은 예측율 중 가장 높은 값은 K-Parameter값이 7일 때의 예측율 76.3% 이

지만, 이러한 예측율은 앞서 사용한 여러 모형들에 비하여 상대적으로 낮은 예측성과라고 할 수 있다.

<그림 7> K-Parameter값 변화에 따른 예측율의 변화

5.1.5 단일모형 실험결과 정리

지금까지 설명한 실험 결과를 종합해 보면 <표 12>와 같다. 학습용 데이터셋에 대해서는 로지스틱 회귀분석이 80.5% 정도의 예측율을 보여 가장 저조한 것으로 나타났고, 인공신경망의 예측율이 81% 이상으로 가장 우수한 결과를 나타내고 있다. 반면, 검증용 데이터셋으로 검증한 결과는 인공신경망과 로지스틱 회귀분석이 가장 우수한 예측력을 보이고 있고, 사례기반추론 기법인 K-NN이 76.3%로 가장 저조한 예측력을 나타내고 있다.

5.2 GA 통합모형 실험결과

본 연구에서 제안하는 유전자 알고리즘 기반 통합모형의 실험 결과를 도출하기 위하여 유전자 알고리즘은 Palisade사의 Evolver 프로그램을 사용하였다. 이 때 모집단(Population)은 100으로 설정하였고 5000번의 iteration을 통하여 50세대 진화 후, 형성된 population 중에서 최적의 가중치를 선정하였다. 유전자 알고리즘을 사용할 때 개별 염색체에 주어지는 값은 최대 7bit 정수 값(0~127사이의 정수 값)으로 제한하였다. 한편 실제 결과 값 도출에 사용되는 가중치(weight)는 그 총합이 1이 되어야 하므로 앞서 '3.3 유전자 알고리즘 기반 통합모형의 절차'에서 언급한 식 (1)에서 표현한 것처럼 *'Real Weight = 개별weight/weight의 총합'*으로 계산하였다. 최종 Weight는 학습용 데이터셋의 적중율을 최대화하는 가중치 중 검증용 데이터셋의 적중률을 최대화하는 가중치로 선정하였다. 이러한 과정을 통해 도출된 가중치는 다음의 <표 13>과 같다.

다음의 <표 13>의 결과를 살펴보면, K-NN을 제외한 각 모형에 비교적 동등한 가중치가 설정되어 있음을 확인할 수 있다. 이는 GA 통합 모형이 어느 한 모형의 결과 값에 치중되어 있는 것이 아니라, 각 모형의 결과 값을 골고루 반영하고 있음

<표 12> 실험 결과 종합표

Data Mining 방법	학습용 데이터 셋	검증용 데이터 셋
로지스틱 회귀분석	80.5%	79.5%
의사결정나무(CART)	80.7%	77.4%
인공신경망(n = 8)	학습(80.8%)/평가(81.2%)	79.5%
인공신경망(n = 16)	학습(80.5%)/평가(81.2%)	78.9%
인공신경망(n = 32)	학습(80.5%)/평가(81.1%)	79.1%
사례기반추론(K-NN)		76.3%(k = 7)

<표 13> GA 통합 가중치

구 분	로지스틱회귀분석 (LOGIT)	의사결정나무 (CART)	인공신경망 (ANN)	사례기반추론 (K-NN)
Weight	125	82	117	20
Real Wgt.	0.363	0.238	0.340	0.057

주) 7bit(0～127 사이의 값) Real Weight의 총합은 1.

<표 14> GA 통합 모형 결과표

예측 / 관찰	예측 결과					
	학습용 데이터셋			검증용 데이터셋		
	0 (정상)	1(부실)	예측율	0(정상)	1(부실)	예측율
0(정상)	948	164	85.3%	245	33	88.1%
1(부실)	253	917	78.4%	73	219	75.0%
전체비율			81.7%			81.4%

을 나타낸다. 위와 같이 선정된 가중치를 사용하여 GA 통합 모형의 예측율을 살펴보면 <표 14>와 같이, GA 통합 모형의 결과가 학습용 데이터셋과 검증용 데이터셋 모두에서 월등한 예측율을 보여주고 있음을 확인할 수 있다. <표 12>와 <그림 14>의 검증용 데이터셋의 결과만 놓고 비교하였을 때에도, GA 통합 모형은 단일 모형에 비해 예측율이 약 2～5% 이상 상승하였다. 또한 학습용 데이터셋과 검증용 데이터셋의 예측율의 차이도 거의 없는 것으로 나타나 매우 안정적인 모형 구축이 가능한 것으로 판단된다.

5.3 GA 통합모형과 단순 통합 모형들 간의 비교

본 연구에서 제안하는 유전자 알고리즘 기반 통합모형의 유용성을 검증하기 위해 여러 가지 다양한 단순 통합모형과 유전자 알고리즘 기반의 통합모형을 <표 15>와 같이 비교 모형들을 제시 하고자 한다.

<표 15>에서 제시한 결합 모형에 대해서 간략히 설명하고자 한다.

5.3.1 VOTNG 모형

VOTNG 모형은 단일모형의 결과 값을 voting하여 결과 값을 통합하는 모형이다. 즉, 문자 그대로 가장 많이 선택된 결과를 선택하는 모형이다. 한편, 실험에 사용된 단일모형은 총 4개이기 때문에 4개의 모형에서 도출된 결과 값이 0(정상)이 2개, 1(부도)이 2개와 같이 동률을 이룰 경우에는 0 또는 1의 결과값을 제시한 모형들의 평균값[2] 중 0 또는 1에 가장 가까운 값을 갖는 모형의 결과 값을

2) 여기서의 평균값은 각 모형들이 보여 주는 원 결과값, 즉 0~1사이의 값(K-NN의 결과값은 제외)을 의미한다. 예를 들어 로지스틱 회귀분석 = 0.8(1 선택), 의사결정나무 = 0.9(1 선택), 신경망 = 0.4(0 선택), K-NN = 0 (0 선택)과 같이 주어졌을 경우, 1을 선택한 모형과 0을 선택한 모형이 동률이므로 이 때에는 (0.8+0.9)/2 = 0.85와 (0.4+0)/2 = 0.2 중 0 또는 1에 가장 가까운 값(이 예에서는 0.85)을 선택한다. 즉 이와 같은 예시에서 VOTNG모형의 결과값은 1이 된다.

<표 15> 비교 모형 및 제안 모형

모 형	모형 결합	가중치 최적화	비 고
로지스틱회귀분석(LOGIT)	없 음	해당 없음	비교 모형
의사결정 나무(CART)	〃	〃	〃
인공신경망(ANN)	〃	〃	〃
사례기반추론(K-NN)	〃	〃	〃
VOTNG	결 합	다수결 투표	〃
SELCT	〃	최적모형선택	〃
SAVRG	〃	단순 평균	〃
WAVRG	〃	가중 평균	〃
GENAG	〃	GA 탐색	제안 모형

선정한다.

5.3.2 SELCT 모형

SELCT 모형에서는 VOTNG 모형에서처럼 0과 1의 결과 값을 사용하는 것이 아니라, 각 단일모형을 통하여 1차적으로 출력되는 0에서 1사이의 원래 결과 값을 사용한다. (K-NN은 0과 1의 결과 값만을 가지므로 제외) 각 모형의 결과 값 중, 원래 결과 값이 0 또는 1에 가장 가까운 값을 가지는 모형을 승자로 선택하여 그 결과 값을 SELCT 모형의 결과 값으로 사용한다. 이때, 나머지 세 모형(LOGIT, CART, ANN) 중에서 승자가 나오지 않았을 경우에는 K-NN의 결과 값을 SELCT 모형의 결과 값으로 사용한다.

5.3.3 SAVRG 모형

SAVRG 모형은 단순 평균모형으로써, 각 모형의 결과 값에 동일한 가중치(본 연구에서는 네 개의 모형밖에 없으므로 각 0.25의 가중치를 가짐)부여한 뒤 가중합계를 산출함으로써 통합 모형의 결과 값을 도출한다.

5.3.4 WAVRG 모형

WAVRG 모형은 가중 평균모형으로써, 학습용 데이터에 대한 각 기법별 예측정확도를 가중치로 활용하여 통합모형의 결과 값을 도출하는 방법이다.

5.3.5 GENAG 모형

GENAG 모형은 GA를 이용하여 각 모형의 최적 가중치를 산출한 후 이에 따른 가중 평균값을 통합모형의 결과 값으로 이용하는 방법으로 본 연구에서 제안하는 방법이다.

본 연구의 제안 모형인 GENAG 모형을 검증하기 위해 단일 모형 4개(로지스틱회귀분석, 의사결정나무, 인공신경망, 사례기반추론)와 통합 모형 5개(VOTNG, SELCT, SAVRG, WAVRG와 GA 탐색을 활용한 GENAG 모형)의 총 9개 모형을 설정하여 실험한 결과를 검증하였으며, 그 검증 결과를 전체 모형의 예측력 평가 결과로 요약하여 <표 16>에 나타내었다.

SAVRG 모형과 WAVRG 모형을 제외한 통합 모형들은 모두 단일 모형의 예측율 보다 더 나은

예측율을 보여주고 있다. 하지만 이러한 통합 모형 중에서도, 검증용 데이터셋에서 가장 높은 예측율을 보여주고 있는 모형은 본 연구에서 제안하는 GA 통합 모형인 GENAG임을 <표 16>을 통하여 확인할 수 있다.

5.4 예측성과에 대한 검증

5.4.1 단일 모형의 McNemar Test

위와 같은 연구 결과의 차이가 통계적으로 유의한 것인지를 파악하기 위하여, McNemar 검정을 수행하고자 한다. 이 검정은 비모수 통계 분석 기법 중 하나로써 관련 있는 두 집단의 분류값의 분포에 대한 차이를 검정하는 방법이다. McNemar Test는 두 집단씩 이루어지므로 Test는 총 10번($_5C_2$)행해졌다. McNemar Test를 통한 유의성 검정 결과는 <표 17>과 같다.

결과 값을 살펴보면 단일 모형의 비교에선 LOGIT 및 K-NN 그리고 ANN 및 K-NN의 패턴이 유의확률 10% 수준에서 유의한 것으로 확인되었다. 즉, ANN과 LOGIT은 예측성과 면에서 K- NN의 결과와 비교하여 우수하며 이는 통계적으로 유의한다는 것을 의미한다. 한편, GA 통합 모형인 GENAG와 단일모형의 성과비교에서는 GENAG의 예측성과

<표 16> 각 모형의 예측율 비교

구 분	모 형	예측율	
		학습용 데이터셋	검증용 데이터셋
단일 모형	로지스틱회귀분석(LOGIT)	80.46%	79.47%
	의사결정 나무(CART)	80.72%	77.37%
	인공신경망(ANN)	80.89%	79.47%
	사례기반추론(K-NN)	66.96%	76.32%
통합 모형	VOTNG	82.56%	80.18%
	SELCT	81.77%	79.82%
	SAVRG	82.03%	79.30%
	WAVRG	78.35%	78.07%
	GENAG(GA 통합모형)	81.73%	81.40%

<표 17> McNemar 검정 결과표

	LOGIT 및 CART	LOGIT 및 ANN	LOGIT 및 K-NN	LOGIT 및 GENAG	CART 및 ANN	CART 및 K-NN	CART 및 GENAG	ANN 및 K-NN	ANN 및 GENAG	K-NN 및 GENAG
N	570	570	570	570	570	570	570	570	570	570
카이제곱[a]	1.061		3.284	3.226	1.120	0.223	5.438	3.440		10.740
근사 유의확률	0.303		0.07	0.072	0.290	0.637	0.02	0.064		0.001
정확한 유의확률 (양측)		1.000[b]							0.27b	

주) a. 연속 수정 b. 이항 분포를 사용함.

가 K-NN의 예측성과에 비해 우수하며, 이는 유의확률 1% 수준에서 유의하였다. 또한, GENAG는 CART와 ANN에 비해서도 예측성과가 우수하며 그 차이는 유의확률 5% 수준에서 유의하였으며, LOGIT에 대해서는 유의확률 10% 수준에서 유의한 차이를 보였다. 따라서 본 연구에서 제안하는 GA 통합 모형이 단일 모형보다 예측성과 면에서 우수하며, 그 차이는 통계적 유의성을 가지는 것을 나타났다.

5.4.2 단일 모형과 단순통합모형의 McNemar Test

기존의 단순통합 모형과 각 개별 모형들을 대상으로 McNemar Test를 실시한 결과, 다음과 같은 <표 18>의 결과를 얻을 수 있었다.

<표 18>을 살펴보면, <표 17>의 결과와 유사하게 K-NN과 단순통합모형을 비교하였을 때에만 근사 유의 확률이 높은 것을 확인할 수 있다. 이를 제외하고는 CART와 VOTNG만이 유의수준 10% 수준의 약한 유의성을 보여주고 있었다.

5.4.3 GA 통합모형의 McNemar Test

GA 통합 모형을 포함한 모든 통합 모형들 간의 McNemar Test를 실시하여 다음과 같은 <표 19>의 결과를 얻을 수 있었다.

<표 19>에서 나타난 바와 같이 GA 통합모형인 GENAG 및 WAVRG 그리고 WAVRG 및 VOTNG의 관계만이 통계적으로 유의할 뿐, 이 둘을 제외한 서로 각기 다른 통합 모형들의 비교에서 유의한 모형의 짝은 없다. 이는 통합모형 간의 예측력 비교에서 GA 통합모형의 예측력이 가장 뛰어 나지만 그 차이가 WAVRG 모형에 대해서만 통계적

<표 18> 단일모형과 단순통합모형의 McNemar 검정 결과표

구 분	모형 비교	N	카이제곱[a]	근사 유의확률
McNemar Test for VOTNG	LOGIT 및 VOTNG	570	0.205	0.651
	CART 및 VOTNG	570	2.885	0.089
	ANN 및 VOTNG	570	0.225	0.635
	K-NN 및 VOTNG	570	10.023	0.002
McNemar Test for SELCT	LOGIT 및 SELCT	570	0.020	0.888
	CART 및 SELCT	570	2.641	0.104
	ANN 및 SELCT	570	0.023	0.880
	K-NN 및 SELCT	570	4.011	0.045
McNemar Test for SAVRG	LOGIT 및 SAVRG	570	0.000	1.000
	CART 및 SAVRG	570	1.053	0.305
	ANN 및 SAVRG	570	0.000	1.000
	K-NN 및 SAVRG	570	9.481	0.002
McNemar Test for WAVRG	LOGIT 및 WAVRG	570	0.907	0.341
	CART 및 WAVRG	570	0.115	0.734
	ANN 및 WAVRG	570	0.845	0.358
	K-NN 및 WAVRG	570	1.397	0.237

주) a. 연속 수정.

<표 19> 통합 모형 간 McNemar 검정 결과표

	GENAG 및 SAVRG	GENAG 및 WAVRG	GENAG 및 SELCT	GENAG 및 VOTNG	SAVRG 및 WAVRG	SAVRG 및 SELCT	SAVRG 및 VOTNG	WAVRG 및 SELCT	WAVRG 및 VOTNG	SELCT 및 VOTNG
N	570	570	570	570	570	570	570	570	570	570
카이제곱[a]	80.942	0.832	101.227	73.758	166.006	7.557		188.046	158.006	19.22
근사 유의확률	0.105	0.017		0.281	0.418	0.801		0.221	0.090	0.883
정확한 유의확률 (양측)			0.108b				0.332b			

주) a. 연속 수정 b. 이항 분포를 사용함.

유의성을 가질 뿐, 나머지 모형과는 유의적인 차이가 없음을 나타낸다. 이는 GA 통합모형이 예측성과 면에서 가장 우수하였지만 검증에 사용된 데이터의 표본 수가 많지 않아서 발생한 결과라고 생각되며, 향후 연구에서 데이터의 표본 수가 보완된다면 유의한 성과의 차이를 나타낼 것으로 판단된다.

6. 결언

서론에서 언급한 바와 같이 기업의 부실 예측은 학계와 업계 모두에서 매우 관심 있게 연구되어 온 분야이다. 기업 부실에 대한 정확한 예측을 하기 위해서는 2가지 측면의 영역이 있는데, 그 중 한 측면은, 예측에 사용되는 변수들을 어떻게 선정하여 미래 부실의 징후를 잘 대변하는 변수를 선택하는가와, 또 다른 측면은, 어떤 분석 기법을 사용하여 보다 예측력 높은 결과치를 도출하는가로 크게 나누어 볼 수 있다.

본 논문에서는 우수한 예측력을 보이고 있는 인공지능 기법들을 이용하고, 유전자 알고리즘을 활용하여 통합한, 유전자 알고리즘 기반의 통합모형을 제안하였다. 본 연구의 제안모형은 각 기법별 결과의 가중치를 최적화하여 기업의 부실여부를 판단하는 결과를 최적화함으로써 보다 정밀한 예측이 가능하도록 설계하였다. 또한 가중치의 결정과정이 여러 모형의 예측치와 이를 통한 통합모형의 예측오류를 최소화하기 위해 전역 최적화 기법을 활용하여 유기적인 최적 통합 가중치를 도출할 수 있었다.

본 연구에서는 연구모형의 유용성을 확인하기 위하여, 수년간 수집된 국내의 외감 중공업 자료를 사용하여 부실예측의 제안모형을 적용해 본 결과, 단일 모형이나 기존에 제시된 단순 통합모형을 적용하는 것에 비해 제안모형인 GA 기반의 통합 모형의 결과가 학습용 데이터셋과 검증용 데이터셋 모두에서 우수한 예측율을 보여 주고 있음을 확인할 수 있었다. 검증용 데이터셋의 결과만 놓고 비교하였을 때에도, GA 통합 모형은 다른 모형에 비해 예측율이 약 2~5% 이상 상승하는 더 높은 예측 성과를 제공하였다. 연구의 결과를 살펴 보면, 단일모형에 비교하여 GA 통합모형의 성과는 매우 우수하고, 그 차이도 통계적 유의성을 가지는 것으로 나타났으며, 모든 단순 통합모형에 비해서도 그 예측성과가 가장 우수하지만, 통계적 유의성은 일부 모형에서만 나타났다. 이는 검증용으로 사용된 데이터의 표본 수가 부족하여 발생한 것으로 생각되며, 향후 연구에서는 충분한 데이터 표본의 확보를 통해 이런 문제점을 보완할 수 있을 것이라 생

각된다.

마지막으로 현재 본 연구는 제안모형을 기업부실 예측분야에 적용하였지만, 모형 자체는 다른 경영분야의 의사결정 문제에도 얼마든지 적용될 수 있다. 그러므로, 향후 연구에서는 제안된 GA 기반의 통합모형을 여러 산업 분야에 적용할 수 있을 것으로 생각된다.

참고문헌

김경재, 안현철, "개인화된 추천시스템을 위한 사용자-상품 매트릭스 축약기법", *Jounal of Information Technology Applications and Management*, 16권 1호(2009), 97~113.

김경재, 한인구, "퍼지 신경망을 이용한 기업부도예측", *한국지능정보시스템학회논문지*, 7권 1호(2001), 135~147.

배재권, 김진화, "기업부도예측을 위한 통합알고리즘", *한국지능정보시스템학회 춘계학술대회 논문집,* (2006), 195~202.

안현철, "데이터 마이닝을 활용한 인터넷 쇼핑몰의 상품추천 시스템 개발", *석사학위논문*, 한국과학기술원, 2002.

안현철, 김경재, 한인구, "Support Vector Machine을 이용한 고객구매 예측모형", *한국지능정보시스템학회 논문지*, 11권 3호(2005), 9~82.

이건창, 김명종, 김혁, "기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법 : MDA, 귀납적 학습방법, 인공신경망 모형과 성과비교", *경영학연구*, 23권 3호(1994), 109~149.

이형용, "한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합 모형", *Entrue Journal of Information Technology*, 7권 2호(2008), 33~43.

최영수, 장욱, "재무변수와 시장변수를 결합한 은행에 대한 부도예측모형", *한국경영학회 통합학술대회 논문집*, (2007), 1~52.

추휘석, 민지경, 이인호, "다수의 인공신경망 모형을 통한 기업데이터의 분류 및 부도예측에 관한 연구", *연세경영연구*, 41권 1호(2004), 514~539.

홍승현, 신경식, "유전자알고리즘을 활용한 인공신경망모형 최적입력변수의 선정-부도예측모형을 중심", *한국지능정보시스템학회논문지*, 9권 1호(2003), 227~247.

홍태호, 신택수, "부도확률맵과 AHP를 이용한 기업신용등급 산출모형의개발", *정보시스템연구*, 16권 3호(2007), 1~20.

Barniv, R., A. Agarwal, and R. Leach, "Predicting the Outcome Following Bankruptcy Filing : A Three-state Classification Using Neural Networks", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.6, No.3(1997), 177~194.

Han I., H. Jo, and K. Shin, "The Hybrid Systems for Credit Rating", *Journal of the Korea OR/MS Society*, Vol.22, No.3(1997), 163~173.

Jo, H., I. Han, and H. Lee, "Bankruptcy prediction using case based reasoning, neural networks, and discriminant analysis", *Expert Systems With Application*, Vol.13, No.2(1997), 97~108.

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

Kim, E., W. Kim, and Y. Lee, "Combination of Multiple Classifiers for the Customerd's Purchase Behavior Prediction", *Decision Support Systems*, Vol.34, No.2(2003), 167~175.

Kim, K. J. and W. B. Lee, "Stock Market Prediction

using Artificial Neural Networks with Optimal Feature Transformation", *Neural Computing and Applications*, Vol.13, No.3(2004), 255~260.

Kim, M. J., S. H. Min, and I. Han, "An Evolutionary approach to the Combination of Multiple Classifiers to Predict a Stock Price Index", *Expert Systems with Applications*, Vol.31, No.2(2006), 241~247.

Pal, S. K. and P. P. Wang, *Genetic Algorithms for Pattern Recognition*, CRC Press, 1996.

Shin, K., "A GA-based Rule Extraction for Bankruptcy Prediction Modeling", *Journal of Korea Intelligent Information Systems Society*, Vol.7, No.2(2001), 83~93.

Shin, K. S. and K. J. Lee, "Bankruptcy Prediction Modeling Using Multiple Neural Network Models", *Lecture Notes in Computer Sciences, Vol.3214(2004), 668~674.*

Abstract

Integrated Corporate Bankruptcy Prediction Model Using Genetic Algorithms

Joong-kyung Ok* · Kyoung-jae Kim*

Recently, there have been many studies that predict corporate bankruptcy using data mining techniques. Although various data mining techniques have been investigated, some researchers have tried to combine the results of each data mining technique in order to improve classification performance. In this study, we classify 4 types of data mining techniques via their characteristics and select representative techniques of each type then combine them using a genetic algorithm. The genetic algorithm may find optimal or near-optimal solution because it is a global optimization technique. This study compares the results of single models, typical combination models, and the proposed integration model using the genetic algorithm.

Key Words : Integrated Model, Corporate Bankruptcy Prediction, Genetic Algorothms, Intelligent Credit Rating System, Data Mining

* Dept. of MIS, Dongguk Univ-Seoul

저 자 소 개

옥중경

현재 한국기업데이터㈜에서 상무로 재직 중이다. 연세대학교에서 경영학석사를, 동국대에서 경영학박사과정을 수료하였다. 주요경력은 삼성SDS에서 정보시스템 실장 및 금융부문 프로젝트매니저를 역임하였다. 관심분야는 데이터마이닝, 기업신용평가시스템, 프로젝트매니지먼트 등이다.

김경재

현재 동국대학교 경영대학 경영정보학과 부교수로 재직 중이다. 한국과학기술원에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, Annals of Operations Research, Applied Intelligence, Applied Soft Computing, Computers in Human Behavior, Expert Systems, Expert Systems with Applications, Intelligent Data Analysis, Intelligent Systems in Accounting, Finance and Management, Neural Computing and Applications, Neurocomputing 등의 학술지에 논문을 게재하였다. 연구 관심분야는 데이터마이닝, 지능형 신용평가시스템, 고객관계관리 등이다.