인공신경망 학습단계에서의 Genetic Algorithm을 이용한 입력변수 선정

Input Variables Selection using Genetic Algorithm in Training an Artificial Neural Network

이 재 식†, 차 봉 근‡
†아주대학교 경영대학 부교수
‡아주대학교 대학원 경영정보학과 석사과정
경기도 수원시 팔달구 원천동 산5번지
Tel: 0331-219-2719, Fax: 0331-219-2190
E-mail: leejsk@madang.ajou.ac.kr

Abstract

Determination of input variables for artificial neural network (ANN) depends entirely on the judgement of a modeler. As the number of input variables increases, the training time for the resulting ANN increases exponentially. Moreover, larger number of input variables does not guarantee better performance. In this research, we employ Genetic Algorithm for selecting proper input variables that yield the best performance in training the resulting ANN.

1. 서 론

인공신경망은 규칙기반 전문가시스템의 한계점인 지식획득 문제를 완화시켜줄 수 있다는 장점때문에 문제해결의 규칙을 명시적으로 추출할 수 없는 문제에 많이 적용되어 왔다[11]. 경영문제의 경우에는 투자 의사결정[9], 기업 도산예측[2,3,5,12], 재무 위험 관리[4], 원가시스템[10] 등에 적용되었다. 인공신경망은 입력변수들의 값과 목표출력값을 주고 학습을 통해서 이들간의 관계를 추출한다. 그러므로 입력변수의 수가 많으면 학습시간이 오래 걸리게 되며, 더욱이 입력변수들 사이에 잡음(noise)이 섞여 있는 경우에는 적정한 입력변수들을 선정하는 것이 예측율과 학습 효율을 높이는데 중요한 요소가 된다. 본 연구에서는 유전자 알고리즘(Genetic Algorithm : GA)을 이용하여 인공신경망의 학습 에러율을 최소로 하는 입력변수의 구성을 구하고자 한다. 실험 자료는 기업 도산예측을 위한 재무정보 22개를 이용하였다 [1].

2. 인공신경망의 입력변수 선정을 위한 유전자 알고리즘

2.1 유전자 알고리즘 (Genetic Algorithm : GA)

John Holland에 의해서 개발된 GA는 자연의 법칙인 "적자생존의 원리"에 근거한 문제해결 기법으로서, 우리가 풀고자 하는 문제의 최적해에 근접한 해들의 특성을 다음 세대에 유전시키므로써, 새대를 거듭할수록 더욱 근접한 해를 찾아가는 탐색기법이다[6,7]. GA는 스케줄링[13], 기업도산예측[5] 등의 문제에 성공적으로 적용되었다. GA의 특징은 첫째, 문제에 속한 파라메터의 숫자 자체를 사용하지 않고 그것을 문자, 숫자 또는 부호의 열(string)로 변환시킨 파라메터 집합을 사용한다. 둘째, 미분함수와 같은 부수적인 함수를 사용하지 않고 최적화하고자 하는 함수 하나만을 사용한다. 셋째, 확정적 전이규칙(deterministic transition rule)보다는 확률적 전이규칙(probabilistic transition rule)을 이용한다. 또한 해를 찾는 과정에 있어서 지역적(local)이기 보다는 전역적(global)으로 접근하며, 병렬 탐색을 하여 보다 빠르게 최적해에 도달할 수 있다. 문제의 파라메터를 string으로 변환시키고 이를 평가할 적응함수(fitness function)를 만드는 일이 문제의 해를 찾는데 있어 주요

작업이 된다.

2.2 파라메터의 변환 (Encoding)

문제의 가능해를 나타내는 string은 생물학에서 빌려온 용어로 chromosome이라 한다. 보통 chromosome은 n개의 string으로 구성되고 각 문자는 gene이라고 한다. Chromosome m개가 모여서 한 세대(generation)를 구성한다. 본 연구에서는 22개의 재무변수를 인공신경망에의 입력 가능변수로 선정하였는데, 이 변수들의 선정 유·무의 결정을 gene으로 표현하였다. 즉, gene의 값은 '0' 혹은 '1'을 가지며, 한 chromosome의 길이는 22가 된다. 여기서 '0'과 '1'은 숫자라기보다는 부호의 의미를 가진다.

2.3 유전자 조작자 (GA Operators)

2.3.1 재생산 (Reproduction)

재생산 조작자는 각 chromosome이 갖는 적응함수값을 참고하여 전 세대에서 다음 세대로 chromosome을 복사하는 과정이다. 적응함수값을 참고한다는 것은 적응함수값이 클수록 즉, 최적해에 근접해 있을수록 다음 세대로 유전될 확률이 커진다는 것을 의미한다.

2.3.2 교배 (Crossover)

교배는 임의의 두 chromosome의 임의의 부분을 서로 바꾸는 프로세스이다. 교배가 일어날 위치 k는 난수에 의해 선정하는데, 교배가 일어날 두 chromosome에서 이 위치는 같다. 여기서 k는 1부터 chromosome의 총길이(22)보다 작은 21 사이의 값을 갖는다. 위치가 선정되었으면 k+1에서 22사이에 있는 부호열을 서로 바꾼다. 교배는 매번 일어나는 것은 아니고 자연의 법칙대로 확률을 갖게 된다. 교배가 일어날 확률은 0.65로 하였다.

2.3.3 돌연변이 (Mutation)

교배 후에는 돌연변이가 일어날 수 있다. 돌연변이는 각 chromosome의 gene의 값을 `0'에서 `1'로, 혹은 그 반대로 바꾼다. 돌연변이가 필요한 이유는 재생산이나 교배에서 놓칠 수 있는 잠재적인 해를 놓치지 않기 위함이다. 돌연변이가 일어날 확률은 0.08로 하였다.

2.4 적응함수 (Fitness Function)

적응함수는 가능해 즉, chromosome의 해 근접도를 나타낸다. 문제의 성격에 따라 적응함수의 식은 달라질 수 있다. 본 연구에서는 선정된 입력변수 개수, 인공신경망의 학습 에러율(%), GA 세대 번호 등을 고려하였다. 본 실험에서는 다음과 같은 식을 적응함수로 사용하였는데[8], 선정된 입력변수의 개수가 적을수록, 인공신경망의 학습 에러율이 작을수록 적응함수의 값은 커지게 되며, 세대 번호가 클수록 chromosome간의 적응함수간의 격차가 커지게 된다.

$$f = \left(1.0 - e^{-(x-1)^{0.15(z+1)}}\right) \times e^{-0.01y^{0.7(z+1)^{1/3}}}$$

x=선택전 입력변수의 개수 / 선택후 입력변수의 개수

y=인공신경망 학습 에러율(%)

z=세대 번호

2.5 기업 도산예측 인공신경망

기업 도산예측을 위한 인공신경망은 입력층, 은닉층, 출력층으로 구성되어 있는데, 입력층의 처리요소 개수는 GA에 의해 선정된 입력변수 개수에 따라 변하며, 은닉층의 처리요소 개수는 5개, 출력층의 처리요소 개수는 1개로 하였다. 학습은 Backpropagation algorithm에 의하여 수행하였으며, 입력층에서의 전이함수는 선형이고, 은닉층과 출력층에서의 전이함수는 Sigmoid 함수이다.

2.6 입력변수 선정과정

초기의 입력변수는 난수에 의하여 선정한 후, 특정 세대 수만큼 아래의 단계를 반복하여 입력변수를 선정하였다.

① 선정된 변수들을 이용하여 인공신경망을 구성한다.
② 인공신경망을 학습시키고, 학습 에러율(%)을 계산한다.
③ 적응함수값을 구한다.
④ GA 조작자를 이용하여 다음 세대의 변수들을 선정한다.

3. 실험결과 및 결론

입력 가능변수는 다음 <표 1>과 같이 기업 도산예측을 위해 고려할 수 있는 재무변수 22개를 사용하였다.

<표 1> 입력 가능 재무변수

안정성 비율	(1)유동성 비율, (2)당좌 비율, (3)고정 비율, (4)고정장기적합율, (5) 부채비율, (6)자기자본비율, (7)매출채권대매입채무비율
수익성 비율	(8)총자본수익율, (9)총자본경상이익율, (10)매출액순이익율, (11)매출액경상이익율
활동성 비율	(12)총자본회전율, (13)자기자본회전율, (14)경영자본회전율, (15)매출채권회전율, (16)매입채무회전율
생산성 비율	(17)부가가치율, (18)노동생산성, (19)총자본투자효율
성장성 비율	(20)총자본증가율, (21)매출액증가율, (22)유형고정자산증가율

GA는 MS Visual Basic을 이용하여 구축하였고 인공신경망은 NeuralWorks (NeuralWare, Inc.), 그리고 자료 교환과 정리를 위하여 MS Excel을 이용하였다. 한 세대의 population의 크기 즉, chromosome의 수는 10개로 하였고 세대 수는 20까지 하였다. 인공신경망의 입력처리요소는 GA에 의해서 생성된 chromosome을 바탕으로 구성하였으며, 학습 회수는 10만 번으로 고정하였다. 학습 데이타 개수는 60개로, 도산 기업 30개와 비도산 기업 30개로 구성하였는데 이 자료는 한국신용평가주식회사로 부터 얻은 1991년에서 1993년까지의 것이다.

본 연구의 실험 결과는 <표 2>에 제시되어 있는데, 10번째 세대에서 학습 에러율이 가장 작은 것을 알 수 있다. 그리고 같은 수의 입력변수를 선정하였다 하더라도 그 구성에 따라 학습 에러율에서 차이가 많이 남을 알 수 있는데, 예를 들면 4번째 세대와 10번째 세대를 비교해 보면 선정된 입력변수의 수는 9개이지만 학습 에러율은 20.00%와 8.33%로 다르다. 참고로 22개의 입력변수를 모두 사용하여 10만번 학습한 경우의 학습 에러율은 18.33%이었다. 즉, 인공신경망의 입력변수가 많다고 학습 결과가 좋다고 말할 수는 없으며, 같은 수의 입력변수라 할지라도 어떠한 변수들이 선정되었는가에 따라 학습 결과가 달라진다는 것을 알 수 있다. 학습 효율적 측면에서도 입력변수가 적은 경우가 시간이 적게 걸림은 당연하다.

<표 2> 선정된 입력변수들과 학습 에러율

Generation	Best Solutions with generations(1 to 20)	# of input	Training Error(%)
1	1 4 8 12 14 17 21	7	10.00
2	10 14 17 18 20 21	6	16.67
3	1 7 10 14 15 17 18 19	8	16.67
4	1 2 3 4 6 15 19 20 21	9	20.00
5	4 5 10 15 19 20 21	7	18.33
6	5 6 8 15 19 20 21	7	20.00
7	1 2 4 5 6 7 8 10 12 22	10	13.33
8	2 10 14 16 19 21	6	16.67
9	10 14 16 19 21	5	21.67
10	4 7 8 14 16 17 19 21 22	9	8.33
11	5 10 12 18 21	5	20.00
12	5 8 9 15 17 18 21	7	21.67
13	11 12 14 21	4	26.67
14	5 7 8 14 16 20 21	7	18.33
15	7 8 21	3	33.33
16	2 7 8 16 21	5	18.33
17	2 7 8 10 12 16 21	7	23.33
18	7 9 18 20	4	35.00
19	4 6 9 14 15 17 18 21 22	9	18.33
20	8 11 17 18 20 21	6	10.00

4. 향후 연구과제

GA의 효율을 높이기 위하여 본 실험에서 이용한 기본적인 방법 외에 진보적인 기술, 예를 들어 안정단계의 재생산(steady-state reproduction)[6], 문제 특성을 고려한 유전자 조작자 등을 적용할 필요가 있다. 본 연구에서는 인공신경망의 학습 횟수를 고정시켰지만, 인공신경망의 학습에 소요되는 시간의 장기화를 감수할 수 있다면 각 인공신경망마다 최소의 학습 에러율을 얻을 때까지 학습을 계속하는 것이 더 바람직할 것이다. 본 연구는 인공신경망의 학습에 그쳤지만, 향후에는 학습된 인공신경망의 테스트에까지 연구를 확장할 예정이다.

참고문헌

1. 기업경영분석, 한국은행, 1991.
2. 이재식, 한재홍, "인공신경망을 이용한 중소기업도산예측에 있어서의 비재무정보의 유용성 검증," 한국전문가시스템학회지, Vol. 1, No. 1, 1995, pp. 123~134.

3. 한인구, 권영식, 이건창, 주성도, "지능형 기업 신용평가 시스템의 개발:NICE-AI," 한국경영정보학회 추계학술대회, 1994. pp. 229~252.
4. Bansal A., R. J. Kauffman, R. M. Mark and E. Peters, "Financial Risk and Financial Risk Management Technology," Information & Management, 24, 1993, pp. 267~281.
5. Barbro B., Teija L., and Kaisa S., "Neural Networks and Genetic Algorithms for Bakruptcy Predictions," *Proc. of the Third World Congress on Expert Systems*, Seoul, Feb., Vol. I, 1996, pp. 123~130.
6. Davis L., "Handbook of Genetic Algorithms," Van Nostrand Reinhold, 1991.
7. Goldberg D. E. , "Genetic Algorithms in Search, Optimization & Machine Learning," Addison Wesley, 1989.
8. Guo Z. and Uhrig R. E., "Use of Genetic Algorithms to Select inputs for Neural Networks," *Proc. of the Special Workshop on Combination of Genetic Algorithms and Neural Networks, Int'l Joint Conference on Neural Networks*, Baltimore, MD, June 6-11, 1992.
9. Kamijo, K. and T. Tanigawa, "Stock Price Pattern Recognition: A Recurrent Neural Network Approach," *Proc. of Int'l Joint Conf. on Neural Networks*, Vol. I, 1990, pp. 215-222.
10. Lee, J. S. and T. S. Ahn, "An Artificial Neural Network Approach for Activity-Based Cost Allocation," *Proc. of Korea/Japan Joint Conf. on Expert Systems*, Feb. 1993, pp. 380-393.
11. Nelson M. M. and Illingworth W. T., "A Practical Guide to Neural Nets," Addison Wesley, 1991.
12. Odom, M. D. and R. Sharda, "A Neural Network Model for Bankruptcy Prediction," *Proc. of Int'l Joint Conf. on Neural Networks*, Vol. II, 1990, pp. 163-168.
13. Schultz A. C., J. J. Grefenstette, and K. A. De Jong, "Test and Evaluation by Genetic Algorithms," IEEE Expert, Oct., 1993, pp. 9~14.